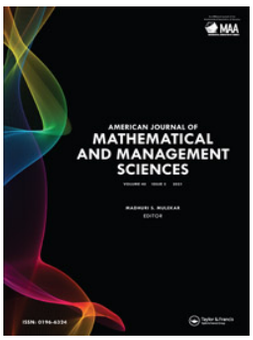

American Journal of Mathematical and Management Sciences

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/umms20

Confidence Intervals for a Population Size Based on Capture-Recapture Data

Bao-Anh Dang, K. Krishnamoorthy & Shanshan Lv

To cite this article: Bao-Anh Dang, K. Krishnamoorthy & Shanshan Lv (2021) Confidence Intervals for a Population Size Based on Capture-Recapture Data, American Journal of Mathematical and Management Sciences, 40:3, 212-224, DOI: 10.1080/01966324.2020.1835591

To link to this article: https://doi.org/10.1080/01966324.2020.1835591

AMERICAN JOURNAL OF MATHEMATICAL AND MANAGEMENT SCIENCES
2021, VOL. 40, NO. 3, 212–224
https://doi.org/10.1080/01966324.2020.1835591

Confidence Intervals for a Population Size Based on Capture-Recapture Data

Bao-Anh Dang[a], K. Krishnamoorthy[a], and Shanshan Lv[b]

[a]Department of Mathematics, University of Louisiana at Lafayette, Lafayette, Louisiana, USA;
[b]Department of Statistics, Truman State University, Kirksville, Missouri, USA

ABSTRACT
Capture-recapture is a popular sampling method to estimate the total number of individuals in a population. This method is also used to estimate the size of a target population based on several incomplete records/databases of individuals. In this context, a simple approximate confidence interval (CI) based on the hypergeometric distribution is proposed. The proposed CI is compared with a popular approximate CI, likelihood CI and an exact admissible CI in terms of coverage probability and precision. Our numerical study indicates that the proposed CI is very satisfactory in terms of coverage probability, better than the popular approximate CI, and much shorter than the admissible CI. The interval estimation method is illustrated using a few examples with epidemiological data.

KEYWORDS AND PHRASES:
Admissible confidence interval; coverage probability; multiple lists; precision; prevalence; score method

1. Introduction

Capture-recapture is a popular sampling design that is often used to estimate the size of animal populations in a wilderness. The procedure is that at one time a sample of M animals is captured, tagged and released; this is called the capture stage. At a later time, another sample of n animals will be caught (recapture) and the number of marked animals X is recorded. The number of animals in each sample, and the number of animals common to both, are used to estimate the animal population size. This simple method was first applied to the epidemiological problem of estimating prevalence by Sekar and Deming (1949). Since the work of Wittes, Colton and Sidel (1974), this method has been widely used in many epidemiological problems. For example, to estimate incidence of diseases (such as cancer, stroke, mental illness), homelessness and people who inject drugs (PWID). The capture-recapture models are also used to estimate the size of a target population based on several incomplete records/databases of individuals; for example, see Tilling (2001) and Chao et al. (2001). Some examples of epidemiological applications of capture-recapture are given in the example section.

A simple approach of merging data from different sources and eliminating duplicate cases are likely to underestimate the target population size. One of the methods for estimating the population size is based on the hypergeometric model. Specifically, the

CONTACT K. Krishnamoorthy krishna@louisiana.edu Department of Mathematics, University of Louisiana at Lafayette, Lafayette, LA 70504, USA.

number of marked animals M in the capture phase is regarded as the number of animals in the population of size N with an attribute. If the sample of size n at the recapture phase is drawn without replacement, then the number of marked animals X in the recapture sample has the hypergeometric distribution with the probability mass function (pmf),

$$P(X = x|n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \max\{0, n + M - N\} \le x \le \min\{n, M\}. \quad (1)$$

In this setup, both N and M could be unknown parameters. As noted earlier, in the capture-recapture sampling, the values of n and M are known, and the problem is to estimate the population size N.

There are a few approximate confidence intervals (CIs) available in the literature. Among them, the CI based on the Chapman's (1951) estimate is popular and commonly used. Wang (2015) proposed an admissible CI which is shorter than the exact CI. Wang's coverage studies indicated that the approximate CIs are too liberal, having coverage probabilities much smaller than the nominal level. In this sense, the approximate CI is inaccurate even for large samples. In general, exact CIs for discrete models are usually too conservative and unnecessarily wide. For example, see Agresti and Coull (1998) and Brown, Cai and DasGupta (2001). The admissible CI proposed by Wang (2015) could be narrower than the exact CI, but our numerical study indicates that the admissible CI is also too wide. Furthermore, calculation of admissible CI is quite involved and it is extremely time consuming for some cases where n is large. Even though the problem of estimating the population size based on the capture-recapture data has been in the literature since the 1940s, no simple approximate closed-form CI that is satisfactory in terms of coverage probability and precision is available.

In this article, we develop a simple approximate closed-form CI for the unknown population size by inverting the score statistic for the proportion in a finite population. In the following section, we describe the available approximate CI based on Chapman's estimate, the likelihood CI, the exact CI and the admissible CI due to Wang (2015). We then propose a score CI, which is formed by two roots (greater than M) of a cubic polynomial function. We also propose a simple closed-form approximate score CI from which the score CI can be deduced. In Section 3, we evaluate and compare all the CIs in terms of exact coverage probability and precision. In Section 4, we illustrate the interval estimation methods using two real examples. Some concluding remarks are given in Section 5.

2. Confidence Intervals

2.1. An Approximate CI

Petersen (1896) proposed a natural point estimate for N as $\hat{N} = M/(X/n)$. Chapman (1951) modified the Petersen estimator as,

$$\hat{N} = \frac{(M+1)}{(X+1)/(n+1)} - 1. \tag{2}$$

The above estimator is unbiased if $n + M \geq N$ and it is approximately unbiased if $X \geq 7$. (Robson and Regier, 1964).

There are a few closed-form approximate CIs available in the literature. The most popular approximate CI is based on the above Chapman estimate and is given by,

$$\left(\frac{(n+1)(M+1)}{X+1} - 1\right) \pm z_{\alpha/2} \left(\frac{(n+1)(n-X)(M+1)(M-X)}{(X+1)^2(X+2)}\right)^{1/2}, \tag{3}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution.

2.2. Likelihood Confidence Interval

The log-likelihood function is given by,

$$\ln f(N) = \ln f(x|n, M, N),$$

where $f(x|n, M, N)$ is the pmf given in (1). Noting that the maximum likelihood estimator $\hat{N}$ of N is the largest integer less than or equal to (nM/x), a $100(1-\alpha)$ likelihood CI for N is given by,

$$\left\{N : -2\left[\ln f(N) - \ln f(\hat{N})\right] \leq z^2_{\alpha/2}\right\}. \tag{4}$$

For a given (x, n, M), the likelihood CI (4) can be obtained numerically. For example, starting from $\hat{N}$, a forward search can be used to find the value of N, say N^*, for which $-2\left[\ln f(N^*) - \ln f(\hat{N})\right] > z^2_{\alpha/2}$. Then, $N^* - 1$ is the right endpoint of the CI. Similarly, the left endpoint of the CI can be obtained using backward search starting from $\hat{N}$.

2.3. Confidence Intervals Based on the Score Statistic

The score statistic for testing the proportion in a finite population is given by,

$$\frac{\hat{p} - p}{\sqrt{\text{Var}(\hat{p})}} = \frac{\hat{p} - p}{\sqrt{R_n p(1-p)/n}}, \tag{5}$$

where $p = M/N, \hat{p} = X/n$ and $R_n = (N-n)/(N-1)$, which is the finite population correction. Wald's (1943) result shows that the above quantity has an approximate standard normal distribution for large n. For a given (X, n, N), the roots of the equation $|\hat{p} - p|/\sqrt{R_n p(1-p)/n} = z_{\alpha/2}$ form a CI for p, which is referred to as the score CI for p.

We now use the same score statistic to obtain an approximate CI for N as follows. Let us write the score statistic (5) in terms of M and N as,

$$Z(N) = \frac{X/n - M/N}{\sqrt{\text{Var}(X/n)}} = \frac{X/n - M/N}{\sqrt{R_n \frac{M}{N}\left(1 - \frac{M}{N}\right)/n}}. \tag{6}$$

For a given (X, n, M), an approximate $1-\alpha$ confidence set is given by

$$\left\{N : Z^2(N) \leq z^2_{\alpha/2} \text{and} N \geq M\right\}.$$

The confidence set is an interval determined by the roots of the equation,

$$Z^2(N) - z^2_{\alpha/2} = 0, \tag{7}$$

and we shall refer to the CI for N as the score CI. The above equation is a cubic polynomial in N, and two real roots that are greater than M form a CI for N.

In order to find exact roots of the equation (7), we first find approximate closed-form roots, which form an approximate CI, referred to as the approximate score CI.

2.3.1. An Approximate Score Confidence Interval

A closed-form approximate score CI can be found as follows. Noting that the proportion X/n of the tagged animals in the second sample is similar to the proportion M/N of the animals trapped in the first occasion, we see that $X/n \simeq M/N$ or $n/N \simeq X/M$. Writing,

$$R_n = \frac{N-n}{N-1} \simeq 1 - \frac{n}{N} \simeq 1 - \frac{X}{M} = R_x, \text{say},$$

and replacing the R_n in (7) with R_x, we see that the CI is determined by the values of N that satisfy the equation,

$$\frac{(\hat{p}-p)^2}{R_x p(1-p)/n} - z^2_{\alpha/2} = 0, \tag{8}$$

where $\hat{p} = X/n$ and $p = M/N$. The roots can be expressed as follows. Solving the above equation for p, we find the CI interval for p as,

$$(\hat{p}_l, \hat{p}_u) = \frac{\hat{p} + \frac{z^2_{\alpha/2}R_x}{2n} \mp \frac{z_{\alpha/2}\sqrt{R_x}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{\alpha/2}R_x}{4n^2}}}{1 + \frac{z^2_{\alpha/2}R_x}{n}}}{1 + \frac{z^2_{\alpha/2}R_x}{n}}. \tag{9}$$

If we replace R_x with 1, then the above CI simplifies to the popular Wilson's (1927) score CI for the binomial proportion; for more details, see Agresti and Coull (1998). The CI for N on the basis of the above CI for p is given by,

$$\left[\hat{N}^{(1)}_l, \hat{N}^{(1)}_u\right] = \left[\lceil M/\hat{p}_u \rceil, \lfloor M/\hat{p}_l \rfloor\right], \tag{10}$$

where $\lceil x \rceil$ is the ceiling function and $\lfloor x \rfloor$ is the floor function. It should be noted that the CI $(p_l, p_u) = (0, z^2_{\alpha/2}/(n + z^2_{\alpha/2}))$ when $X = 0$, and it is $(n/[n + z^2_{\alpha/2}(1 - n/M)], 1)$ when $X = n$. Thus, the CI (10) for N is defined for all values of $X \in \{0, 1, ..., n\}$.

Remark 1. The hypergeometric pmf remains the same when n and M are swapped. Specifically, for any given $(n, M) = (n_0, M_0)$,

$$P(X = x | n = n_0, M = M_0, N) = P(X = x | n = M_0, M = n_0, N).$$

Any CI for N should reflect this property of the hypergeometric model. In particular, the CI for N at $(X, n = n_0, M = M_0)$ and the CI for N at $(X, n = M_0, M = n_0)$ should

be the same. Notice that the approximate CI in (3) possesses this property, whereas the CI (10) does not possess this property. To overcome this problem, we compute the CI $[N_l^{(1)}, N_u^{(1)}]$ in (10) at $(X, n = n_0, M = M_0)$, and construct the same at $(X, n = M_0, M = n_0)$, say, $[N_l^{(2)}, N_u^{(2)}]$, and propose the CI,

$$[N_l, N_u] = \left[\frac{N_l^{(1)} + N_l^{(2)}}{2}, \frac{N_u^{(1)} + N_u^{(2)}}{2}\right]. \tag{11}$$

We refer to the above CI as the approximate score CI or simply A-score CI.

2.3.2. Calculation of the Score CI. The roots for the extreme cases can be expressed in closed-form as follows.

Case of $X = 0$: In this case, noting that $N - 1 \simeq N$, we see that (7) becomes quadratic, and the finite root is given by,

$$\hat{N}_l = M\left(\frac{n}{2z_{\alpha/2}^2} + \frac{1}{2}\right) + \frac{n}{2} + \frac{1}{2z_{\alpha/2}^2}\sqrt{n^2(M + z_{\alpha/2}^2)^2 + z_{\alpha/2}^2 M\left(z_{\alpha/2}^2 M + 2n(M - z_{\alpha/2}^2)\right)}, \tag{12}$$

and $\hat{N}_u = \infty$. Thus, for the case of $X = 0$, we consider the interval $[\hat{N}_l, \infty]$ as the $1 - \alpha$ CI for N.

Case of $X = n$: In this case also (7) becomes quadratic, and the root that is greater than M is given by,

$$\hat{N}_u = M\left(\frac{1}{2} + \frac{z_{\alpha/2}^2}{2n}\right) + \frac{1}{2} + \sqrt{\frac{1}{4n^2}\left(n + M(n + z_{\alpha/2}^2)\right)^2 - M\left(1 + z_{\alpha/2}^2\right)}. \tag{13}$$

Thus, for the case of $X = n$, we consider the interval $[M, \hat{N}_u]$ as the $1 - \alpha$ CI for N.

We shall now describe a method computing the score CIs for $X \in \{1, \ldots, n-1\}$ from the approximate approach in the preceding subsection. Let,

$$Z_x^2(N) = \frac{(\hat{p} - p)^2}{R_x p(1 - p)/n}$$

so that (8) can be written as $Z_x^2(N) - z_{\alpha/2}^2 = 0$, or equivalently,

$$f_x(N) = n(\hat{p} - p)^2 - z_{\alpha/2}^2 R_x p(1 - p) = 0. \tag{14}$$

Similarly, we can write (7) as,

$$f_n(N) = n(\hat{p} - p)^2 - z_{\alpha/2}^2 R_n p(1 - p) = 0, \tag{15}$$

where $p = M/N$.

It can be easily checked that the coefficient of N^3 in $f_n(N)$ is positive, and so the function has two turning points, and is concave in the first half and convex in the second half. Also, $f_n(N)$ is positive for N close to M, and for a fixed (X, n, M), it can be easily verified that

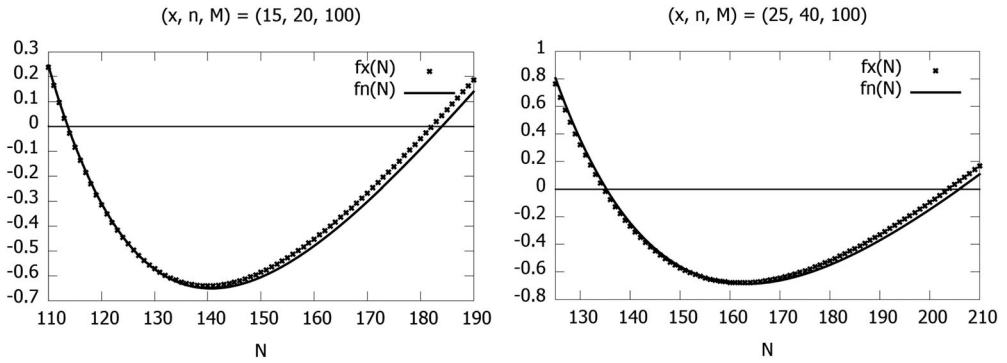

Figure 1. Plots of $f_x(N)$ in (14) and $f_n(N)$ in (15) as a function of N; $1-\alpha = 0.95$.

$$f_x(N) \le f_n(N) \text{ for } N \le \frac{(n-1)M}{X} + 1 \quad \text{and} \quad f_x(N) > f_n(N) \text{ for } N > \frac{(n-1)M}{X} + 1.$$

The above relations between $f_x(N)$ and $f_n(N)$ imply that both the smaller and the larger roots of the equation $f_n(N) = 0$ are larger than the corresponding smaller and larger roots of the equation $f_x(N) = 0$. To shed some light, we plotted these functions $f_x(N)$ and $f_n(N)$ for $(X, n, M, 1-\alpha) = (15, 20, 100, .95)$ and (25, 40, 100,.95) and presented them in Figure 1. As seen in the figure, the 95% CI formed by roots of $f_n(N) = 0$ falls on the right side of the 95% CI formed by the roots of $f_x(N) = 0$. On the basis of this relation, the score CI can be computed from the approximate score CI as follows.

Algorithm 1

For a given (x, n, M):

1. Compute the $1-\alpha$ CI $[\hat{N}_l^{(1)}, \hat{N}_u^{(1)}]$ in (10).
2. Starting from $\hat{N}_l^{(1)}$, search the value of N forward until $Z^2(N) - z_{\alpha/2}^2 \le 0$. The smallest value of N for which $Z^2(N) - z_{\alpha/2}^2 \le 0$ is the left endpoint of the score CI.
3. Starting from $\hat{N}_u^{(1)}$, search the value of N forward until $Z^2(N) - z_{\alpha/2}^2$ becomes positive, and denote the corresponding value of N by N^*. Then, $N^* - 1$ is the right endpoint of the score CI.

The score CI based on Algorithm 1 is invariant under the transformation $(X, n = n_0, M = M_0) \to (X, n = M_0, M = n_0)$. This is true because $Z(N)$ defined in (6) is invariant under the transformation $(X, n, M) \to (X, M, n)$.

2.4. Exact Confidence Intervals

For a given (x, n, M), the cumulative distribution function $F(x|n, M, N) = P(X \le x|n, M, N)$ is an increasing function of N. So using Theorem 9.2.14 of Casella and Berger (2002), we find the following exact one-sided confidence limits for N.

The $1-\alpha$ lower confidence limit L_o is determined by,

$$L_o = \begin{cases} \max\{n, M\} & \text{if } x = \min\{n, M\} \\ \min\{N : P(X \le x|n, M, N) \ge \alpha/2\}, & \text{for } x = 0, 1, \ldots, n-1. \end{cases} \tag{16}$$

The $1-\alpha$ upper confidence limit U_o is determined by,

$$U_o = \begin{cases} \infty & \text{if } x = 0 \\ \max\{N : P(X \geq x|n, M, N) \geq \alpha/2\}, & \text{for } x = 1, ..., n. \end{cases} \tag{17}$$

The exact confidence limits given in Lemmas 5 and 6 of Wang (2015) are defined in a different way, but it can be easily verified that they are the same as the above exact confidence limits. Furthermore, the exact CIs are invariant under the transformation $(X, n, M) \rightarrow (X, M, n)$, because the hypergeometric cdf is invariant under the transformation.

Wang (2015) has shown that for any $1-\alpha$ lower confidence limit L^* with nondecreasing in x, $L_o \geq L^*$. Similarly, for any upper confidence limit U^* with nondecreasing in x, $U_o \leq U^*$. Even though the one-sided confidence limits have some desirable properties, the $1-2\alpha$ two-sided confidence interval $[L_o, U_o]$, formed by these one-sided limits are unnecessarily wide. Wang (2015) has proposed an iterative algorithm to find an exact admissible two-sided CI $[L_e, U_e]$ which is a subset of $[L_o, U_o]$. However, calculation of the admissible two-sided CI is numerically quite involved. Wang (2015) has provided an R function which can be used to compute the admissible CIs for $x \in \{0, 1, ..., n\}$. Specifically, for a given (n, M) and the confidence level, the R function returns CIs corresponding to all $x \in \{0, 1, ..., n\}$.

2.5. Comparisons of Confidence Intervals

To judge the disparity among the CIs by different methods, we computed 95% likelihood CI, score CI, approximate score CI and Wang's (2015) exact admissible CI for $(n, M) = (30, 400)$ and (20, 1000) and is presented in Table 1. By examining the CIs in Table 1, we see that the admissible CIs are wider than the corresponding score and approximate score CIs. Indeed, the admissible CIs are much wider than the approximate score and score CIs when X is much smaller than the sample size n. The likelihood CIs are also wider than the score CIs. As noted earlier, the score CIs always fall on the right side of the corresponding approximate score (A-score) CIs. Furthermore, we note that the score CIs are slightly wider than the corresponding approximate score CIs.

3. Coverage and Precision Studies

For a given (X, n, M, α), let $(C_{X,n,M,\alpha/2}, C_{X,n,M,1-\alpha/2})$ be a $1-\alpha$ CI for N. Then, for an assumed value of N, the exact coverage probability of this CI can be computed using the expression

$$\sum_{x=L}^{U} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} I_{[C_{x,n,M,\alpha/2}, C_{x,n,M,1-\alpha/2}]}(N), \tag{18}$$

where $L = \max\{0, n+M-N\}, U = \min\{n, M\}$ and $I_A(x)$ is the indicator function. Similarly, the expected width of $(C_{X,n,M,\alpha/2}, C_{X,n,M,1-\alpha/2})$ can be computed using the expression

Table 1. 95% confidence intervals based on three methods.

	$n = 30, M = 400$					$n = 20, M = 1000$			
X	Admis.	Likelihood	A-Score	Score	X	Admis.	Likeihood	A-Score	Score
5	(1180, 5862)	(1240, 6300)	(1197, 5424)	(1201, 5440)	4	(2441, 14003)	(2472, 14941)	(2407, 12376)	(2410, 12389)
7	(979, 4010)	(1000, 3695)	(982, 3371)	(985, 3382)	5	(2149, 9598)	(2167, 10207)	(2137, 8921)	(2139, 8931)
8	(906, 3029)	(916, 3017)	(904, 2802)	(907, 2810)	6	(1935, 7157)	(1940, 7575)	(1930, 6858)	(1931, 6866)
10	(786, 2234)	(788, 2169)	(785, 2065)	(787, 2071)	7	(1772, 6487)	(1765, 5930)	(1766, 5505)	(1767, 5512)
11	(735, 1972)	(739, 1889)	(738, 1814)	(739, 1820)	8	(1607, 4774)	(1624, 4818)	(1633, 4558)	(1634, 4564)
15	(600, 1215)	(598, 1216)	(601, 1196)	(602, 1199)	9	(1473, 4329)	(1509, 4023)	(1522, 3862)	(1523, 3867)
16	(574, 1148)	(572, 1110)	(576, 1097)	(577, 1100)	10	(1418, 3406)	(1413, 3428)	(1429, 3331)	(1430, 3335)
19	(511, 905)	(510, 872)	(514, 870)	(515, 873)	11	(1329, 3165)	(1333, 2969)	(1350, 2914)	(1351, 2918)
20	(493, 841)	(493, 811)	(497, 812)	(498, 814)	12	(1268, 2767)	(1264, 2605)	(1282, 2578)	(1282, 2582)
21	(481, 785)	(477, 757)	(482, 760)	(482, 762)	13	(1201, 2440)	(1205, 2308)	(1223, 2303)	(1223, 2305)
22	(464, 734)	(463, 708)	(468, 713)	(468, 715)	14	(1164, 2148)	(1154, 2063)	(1172, 2072)	(1172, 2074)
23	(451, 688)	(450, 664)	(455, 670)	(455, 672)	15	(1118, 1934)	(1110, 1855)	(1127, 1876)	(1127, 1878)
24	(442, 646)	(439, 624)	(444, 632)	(444, 633)	16	(1078, 1771)	(1073, 1677)	(1089, 1706)	(1089, 1708)
25	(431, 599)	(429, 588)	(433, 596)	(433, 597)	17	(1045, 1606)	(1042, 1521)	(1056, 1558)	(1056, 1559)
26	(421, 573)	(420, 554)	(424, 563)	(424, 564)	18	(1019, 1438)	(1019, 1381)	(1029, 1426)	(1029, 1427)
27	(413, 541)	(412, 522)	(415, 533)	(415, 533)	19	(1003, 1328)	(1003, 1251)	(1010, 1304)	(1010, 1305)
30	(400, 450)	(400, 425)	(400, 447)	(400, 447)	20	(1000, 1200)	(1000, 1099)	(1000, 1189)	(1000, 1189)

$$\sum_{x=L}^{U} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \left(C_{x,n,M,1-\alpha/2} - C_{x,n,M,\alpha/2}\right). \tag{19}$$

Recall, that the right endpoints of all CIs at $X = 0$ are infinity, so the expected widths of all CIs are infinity. However, comparison of CIs in Table 1 indicates that the score CIs are much narrower than the admissible CIs for almost all $X \neq 0$. Thus, to compare these CIs in terms of precision, we calculated the right endpoint of the approximate score CI using $X = .5$ when the actual value of $X = 0$. This right endpoint of approximate score CI was used as the right endpoint of all other CIs when $X = 0$. This adjustment ensures that the expected widths of all CIs are finite and a meaningful comparison can be made. Furthermore, all the CIs are invariant under the transformation $(X, n, M) \rightarrow (X, M, n)$, and therefore, we can assume without loss of generality that $n \leq M$ in our comparison studies.

In Table 2, coverage probabilities and expected widths of the approximate CI (3), the likelihood CI (4), the approximate score CI (11) and the score CI based on Algorithm 1 are reported for some values of (n, M, N) and confidence coefficient.95. We observe from this table that the approximate CI is not satisfactory in terms of coverage probabilities. Even for a sample size of 80 (see $n = 80$, $M = 400$), the coverage probability of the approximate CI could be as low as.898. The approximate CIs are narrower than the other two CIs, because the coverage probabilities are much smaller than the nominal level.95. The likelihood CI is quite comparable with the approximate and the score CIs in terms of coverage probabilities, but it is wider than the score CIs in almost all cases. There are some cases where the score CIs have better coverage probabilities with smaller width than the likelihood CIs; see when $(n, M, N) = (30, 200, 425)$, (60,200,550) and (30, 400, 2000). The approximate score and the score CIs are very similar in terms of coverage probability. However, the approximate score CIs are shorter than the score CIs for all the cases considered in Table 2.

Table 2. Coverage probabilities and expected widths of 95% confidence intervals.

	$n = 20, M = 200$				$n = 30, M = 200$			
N	Approx.	Likelihood	A-Score	Score	Approx.	Likelihood	A-Score	Score
300	.857(170)	.952(219)	.952(212)	.952(214)	.927(139)	.935(163)	.935(160)	.936(162)
325	.851(207)	.943(273)	.943(261)	.944(264)	.885(169)	.952(201)	.952(196)	.953(198)
350	.926(245)	.939(333)	.939(313)	.940(317)	.904(200)	.945(243)	.967(234)	.945(237)
375	.905(285)	.936(398)	.936(370)	.936(373)	.910(233)	.965(287)	.965(275)	.966(278)
400	.875(325)	.963(470)	.963(431)	.964(435)	.908(266)	.964(334)	.964(318)	.965(321)
425	.922(367)	.935(551)	.935(497)	.936(501)	.900(301)	.941(384)	.964(363)	.964(366)
450	.885(410)	.964(641)	.964(568)	.964(572)	.885(336)	.943(438)	.943(411)	.944(415)
475	.832(454)	.938(743)	.938(645)	.938(650)	.863(373)	.945(495)	.945(462)	.945(466)
500	.877(499)	.966(859)	.966(729)	.967(734)	.910(410)	.942(556)	.967(515)	.967(519)
525	.910(546)	.942(992)	.942(820)	.942(826)	.882(449)	.948(622)	.948(571)	.949(576)
550	.854(594)	.943(1144)	.943(920)	.944(926)	.918(489)	.945(693)	.947(631)	.947(636)
575	.886(642)	.939(1316)	.971(1028)	.972(1034)	.885(529)	.953(769)	.953(694)	.953(699)
600	.813(692)	.948(1511)	.948(1145)	.949(1152)	.916(571)	.951(852)	.928(760)	.928(765)
	$n = 60, M = 200$				$n = 80, M = 200$			
300	.915(94)	.954(101)	.954(100)	.954(101)	.921(78)	.947(82)	.947(82)	.947(82)
325	.948(115)	.944(125)	.944(123)	.945(124)	.947(96)	.952(102)	.952(101)	.952(102)
350	.931(137)	.937(150)	.937(147)	.955(149)	.928(115)	.960(122)	.960(120)	.946(122)
375	.939(160)	.934(176)	.950(172)	.934(174)	.925(134)	.957(144)	.957(141)	.957(143)
400	.937(183)	.948(203)	.948(199)	.965(200)	.944(155)	.939(166)	.953(163)	.939(165)
425	.928(208)	.949(232)	.949(226)	.949(228)	.925(175)	.953(190)	.953(186)	.953(188)
450	.910(233)	.949(262)	.949(255)	.950(257)	.927(197)	.953(215)	.953(210)	.953(212)
475	.927(259)	.950(294)	.950(285)	.950(288)	.923(219)	.953(241)	.953(235)	.953(237)
500	.938(285)	.949(327)	.949(316)	.933(319)	.945(242)	.938(268)	.952(261)	.938(263)
525	.904(313)	.934(361)	.951(349)	.951(352)	.931(266)	.955(296)	.955(287)	.955(290)
550	.907(341)	.935(397)	.953(382)	.953(386)	.944(290)	.941(325)	.941(315)	.941(318)
575	.906(370)	.938(435)	.954(417)	.938(421)	.922(315)	.942(355)	.956(344)	.942(347)
600	.902(399)	.940(474)	.940(453)	.940(457)	.928(341)	.943(386)	.958(373)	.958(377)
	$n = 30, M = 400$				$n = 40, M = 400$			
1000	.906(834)	.964(1136)	.964(1052)	.964(1057)	.930(725)	.950(902)	.931(857)	.931(861)
1100	.915(992)	.941(1414)	.944(1287)	.968(1292)	.906(862)	.956(1104)	.956(1039)	.957(1043)
1200	.913(1157)	.948(1737)	.947(1549)	.925(1555)	.920(1006)	.960(1329)	.960(1237)	.961(1241)
1300	.902(1330)	.956(2119)	.956(1842)	.956(1848)	.925(1157)	.964(1577)	.944(1452)	.944(1457)
1400	.883(1509)	.933(2573)	.961(2170)	.961(2177)	.923(1314)	.947(1855)	.947(1686)	.948(1692)
1500	.851(1694)	.942(3116)	.942(2539)	.943(2547)	.915(1477)	.954(2167)	.954(1940)	.955(1947)
1600	.896(1886)	.943(3764)	.969(2953)	.969(2961)	.900(1646)	.933(2520)	.960(2217)	.960(2224)
1700	.855(2083)	.952(4530)	.952(3416)	.952(3425)	.876(1820)	.943(2921)	.943(2519)	.943(2526)
1800	.892(2285)	.956(5426)	.956(3932)	.956(3942)	.914(1999)	.942(3379)	.944(2848)	.945(2856)
1900	.839(2491)	.932(6459)	.957(4504)	.958(4515)	.884(2183)	.952(3904)	.952(3208)	.953(3217)
2000	.873(2701)	.932(7633)	.965(5133)	.965(5146)	.841(2372)	.954(4504)	.931(3600)	.931(3610)

The coverage probabilities and expected widths of the admissible CI (Admis.), the likelihood CI (4), the score CI and the approximate score CI (A-Score) are calculated as a function of n, and reported in Table 3, and they are calculated as a function of M and reported in Table 4. The approximate CI (3) is omitted for further evaluation, because its coverage probabilities (see Table 2) are appreciably smaller than the nominal level. From Tables 3 and 4, we first observe that the admissible CI, being exact, has coverage probabilities greater than or equal to the nominal level. However, it is too conservative yielding confidence intervals that are unnecessarily wider. In fact, there are many situations where the admissible CIs are twice wider than the corresponding score and approximate score CIs with similar coverage probabilities; see the values at $n = 11$, 13, 17 and $(M, N) = (20, 100)$ in Table 3; $n = 9, 11, 13, 15, 28, 30$ and $(M, N) = (30, 300)$. The likelihood CIs are much wider than the score CIs. Although the likelihood CIs have satisfactory coverage probabilities, they are too wide. The approximate score CI

Table 3. Coverage probabilities and (expected widths) of 95% CIs as a function of n.

$M = 20, N = 100$					$M = 20, N = 200$				
n	Admis.	Likelihood	A-Score	Score CI	n	Admis.	Likelihood	A-Score	Score
4	.975(1258)	.975(1101)	.975(764)	.975(769)	4	.996(1409)	.949(1302)	.949(1066)	.949(1068)
5	.994(1442)	.946(1263)	.946(819)	.946(827)	5	.992(1726)	.992(1588)	.920(1241)	.920(1246)
6	.986(1565)	.986(1373)	.907(841)	.986(851)	6	.985(2018)	.985(1849)	.981(1386)	.985(1392)
7	.971(1632)	.971(1435)	.971(841)	.971(852)	7	.976(2279)	.976(2083)	.976(1504)	.976(1511)
8	.951(1654)	.951(1457)	.951(822)	.951(834)	8	.965(2512)	.965(2290)	.965(1596)	.965(1606)
9	.985(1637)	.985(1446)	.924(791)	.985(804)	9	.951(2716)	.951(2470)	.951(1667)	.951(1678)
10	.974(1588)	.974(1408)	.974(753)	.974(766)	10	.989(2887)	.989(2623)	.934(1719)	.934(1732)
11	.959(1518)	.959(1351)	.959(709)	.959(724)	11	.984(3032)	.984(2751)	.916(1754)	.984(1769)
12	.986(1432)	.986(1281)	.940(663)	.940(679)	12	.978(3149)	.978(2855)	.978(1776)	.978(1792)
14	.959(1238)	.968(1117)	.968(573)	.933(588)	14	.962(3307)	.961(2995)	.961(1782)	.962(1801)
15	.963(1137)	.927(1030)	.953(530)	.927(545)	15	.951(3351)	.990(3033)	.951(1771)	.951(1791)
16	.963(1036)	.963(945)	.916(489)	.963(504)	16	.987(3373)	.987(3053)	.939(1752)	.939(1774)
17	.960(940)	.960(863)	.960(451)	.960(465)	17	.983(3375)	.983(3057)	.926(1726)	.983(1749)
18	.953(849)	.953(786)	.953(416)	.953(430)	18	.978(3360)	.978(3046)	.978(1696)	.978(1718)
19	.978(766)	.925(713)	.943(384)	.943(398)	19	.972(3331)	.972(3021)	.972(1660)	.972(1684)
20	.975(689)	.932(646)	.929(355)	.975(368)	20	.965(3288)	.965(2985)	.965(1621)	.965(1647)
$M = 120, N = 800$					$M = 80, N = 1000$				
n	Admis.	Likelihood	A-Score	Score CI	n	Admis.	Likelihood	A-Score	Score
10	.951(13460)	.990(12234)	.951(7241)	.951(7255)	10	.960(12697)	.960(11737)	.960(8257)	.960(8269)
15	.984(12780)	.984(11730)	.940(6364)	.940(6382)	12	.988(14262)	.988(13160)	.935(8834)	.935(8848)
20	.979(10350)	.979(9643)	.979(5108)	.979(5128)	15	.973(15942)	.973(14692)	.973(9295)	.973(9314)
25	.960(7774)	.960(7378)	.960(3992)	.960(4011)	17	.959(16657)	.959(15348)	.959(9394)	.959(9415)
30	.967(5675)	.929(5495)	.967(3137)	.967(3154)	20	.982(17210)	.982(15864)	.931(9315)	.931(9341)
35	.951(4156)	.951(4103)	.951(2520)	.951(2536)	25	.957(17030)	.988(15745)	.957(8787)	.957(8818)
40	.962(3122)	.928(3133)	.962(2085)	.962(2100)	30	.972(15936)	.972(14808)	.972(8006)	.972(8039)
45	.950(2440)	.950(2476)	.950(1777)	.950(1791)	35	.982(14363)	.982(13429)	.946(7144)	.946(7179)
50	.962(1994)	.932(2033)	.962(1556)	.962(1570)	40	.955(12609)	.965(11878)	.965(6302)	.965(6338)
60	.963(1496)	.938(1516)	.963(1270)	.963(1282)	45	.956(10874)	.956(10331)	.956(5530)	.956(5565)
70	.966(1245)	.945(1244)	.943(1095)	.966(1106)	53	.968(8401)	.916(8102)	.935(4487)	.968(4520)
80	.952(1094)	.952(1078)	.950(975)	.932(984)	60	.951(6650)	.945(6505)	.951(3769)	.951(3800)
95	.950(940)	.934(918)	.953(849)	.953(857)	68	.964(5120)	.964(5090)	.941(3139)	.941(3168)
100	.964(901)	.949(878)	.949(815)	.949(824)	70	.963(4808)	.963(4798)	.937(3008)	.937(3037)

and the score CI perform similar in terms of coverage probabilities. The expected widths of approximate score CIs are shorter than the score CIs for all the cases considered in Tables 2–4.

Overall, we see that the admissible CIs, even though they are exact, are much wider than the other three CIs. The likelihood CIs are also unnecessarily wide. The approximate score CI and the score CI perform very similar in terms of coverage probabilities, and the former is in closed-form, simple to compute and slightly narrower than the latter.

4. Examples

Example 1. This example is taken from Allen et al. (2019) The opioid epidemic has had a severe impact across the USA, and has fueled outbreaks of HIV and hepatitis C virus (HCV) infections among people who inject drugs (PWID). The outbreaks were linked to the injection of prescription opioids and syringe sharing. Allen et al. have conducted capture-recapture population estimation of PWID in Cabell County, West Virginia in June and July 2018. The capture phase occurred in June 2018 at the Cabell Huntington Harm Reduction Program (CHHRP), where the data were collected anonymously

Table 4. Coverage probabilities and (expected widths) of 95% CIs as functions of M.

$n = 20, N = 400$					$n = 40, N = 400$				
M	Admis.	Likelihood	A-Score	Score	M	Admis.	Likelihood	A-Score	Score
30	.988(6759)	.988(6204)	.946(3667)	.946(3693)	60	.967(1470)	.936(1469)	.967(986)	.967(1001)
35	.977(6829)	.977(6265)	.977(3563)	.977(3589)	70	.954(1053)	.954(1064)	.954(793)	.954(805)
40	.961(6694)	.961(6145)	.961(3393)	.961(3418)	80	.962(823)	.942(828)	.942(669)	.942(679)
45	.983(6409)	.983(5893)	.938(3183)	.938(3207)	90	.953(689)	.953(686)	.950(584)	.930(592)
50	.972(6022)	.972(5549)	.972(2953)	.972(2976)	100	.967(603)	.944(593)	.942(522)	.967(529)
55	.957(5570)	.988(5150)	.957(2718)	.957(2740)	110	.961(540)	.935(526)	.961(475)	.961(480)
60	.981(5088)	.980(4724)	.937(2488)	.980(2509)	120	.956(491)	.956(476)	.956(436)	.956(441)
65	.965(4601)	.944(4291)	.944(2269)	.944(2288)	130	.951(451)	.951(435)	.951(403)	.951(408)
70	.968(4126)	.968(3869)	.938(2065)	.938(2083)	140	.963(418)	.946(402)	.946(375)	.946(379)
75	.966(3677)	.966(3468)	.966(1878)	.966(1894)	150	.959(389)	.943(373)	.943(351)	.943(355)
80	.961(3262)	.907(3096)	.961(1708)	.961(1723)	160	.957(363)	.939(347)	.939(329)	.939(333)
85	.952(2885)	.937(2757)	.952(1556)	.952(1570)	170	.955(341)	.937(325)	.937(310)	.937(313)
90	.976(2548)	.943(2452)	.939(1421)	.976(1433)	180	.953(320)	.935(305)	.935(292)	.935(295)
95	.963(2250)	.944(2180)	.944(1301)	.944(1312)	190	.952(302)	.934(287)	.934(276)	.934(278)
100	.966(1989)	.941(1941)	.941(1195)	.941(1206)	200	.952(285)	.934(270)	.934(261)	.934(263)

Table 5. 95% Confidence intervals for the total number of PWID.

Method	CI
Approx. Method	(1139, 2440)
Likelihood	(1311, 2827)
A-Score	(1284, 2734)
Score	(1295, 2755)
Adm. CI	(1293, 2893)

through audio computer-assisted self-interview. The recapture phase occurred in community locations where PWID congregate and commenced two weeks after the completion of the capture phase. The reported data are as follows. The number of PWID in the capture phase is 194, in the recapture phase is 201 and in both capture and recapture the sample is 21. The problem is to estimate the total number of PWID in Cabell County during the study period.

To estimate the total number of PWID in Cabell County during the study period, we write the data in our model notation as $(X, n, M) = (21, 201, 194)$. Noting that all CIs are invariant under the transformation $(X, n, M) \to (X, M, n)$, we can also take $(X, n, M) = (21, 194, 201)$. The Chapman estimate is calculated as

$$\frac{(M+1)(n+1)}{X+1} - 1 = 1789.$$

We calculated 95% CIs based on different methods and presented in Table 5. We first observe from Table 5 that the approximate CI is the shortest and is quite different from other CIs. The approximate score CI is a little narrower than the score CI. The admissible CI is the widest among all five CIs. The left endpoints of the score and admissible CIs are practically the same, but the right endpoint of the admissible CI is appreciably larger than that of the score CI.

Example 2. Chartier et al. (2015) have analyzed the data on chronic kidney disease (CKD) collected from residents of Manitoba as of April 1, 2012, who met the definition for CKD at any point from April 1, 2004, to March 31, 2012. There are two sources of recorded data: the administrative data and Diagnostic Services Manitoba (DSM)

Table 6. 95% Confidence intervals for the adult population with CKD.

Method	CI
Approx. Method	(132155, 135380)
Likelihood	(132180, 135404)
A-Score	(132174, 135398)
Score	(132179, 135403)
Exact CI	(131949, 135647)

Laboratory Data. The administrative data covers virtually all Manitoba residents in the province and the DSM laboratory data covers Winnipeg residents. The total number of adult Manitoba residents (18 years or older) with CKD found in both administrative data and laboratory data was 69,905. Of these, 32,371 were identified using administrative data, 24,909 were found in the Diagnostic Services Manitoba (DSM) laboratory data, and 12,625 people were found in both data.

To estimate the prevalence of CKD during the study period, we note that $M = 32371 + 12625 = 44996, n = 24909 + 12625 = 37534$ and $X = 12625$. The Chapman estimate (2) is calculated as 133767. The 95% CIs for the prevalence of CKD in Manitoba during the study period are estimated using different methods and are given in Table 6. As the values of M and n are very large, the R code provided by Wang (2015) to compute the admissible CI does not response for hours and so we are unable to report the admissible CI for this data. Even for such large data, the CIs based on the exact method is appreciably wider than the other CIs. The likelihood CI and the score CI are practically the same, and they are not appreciably different from the approximate score CI. The approximate CI is narrower than all other CIs, because it is too liberal.

5. Conclusion

In this article, we have used the score statistic that is used to obtain approximate CIs for the proportion in finite populations to find a CI for population size. The proposed score CIs are much shorter than the exact admissible CI and they are relatively easier to compute than the admissible CI. Indeed, the approximate closed-form score CIs are straightforward to calculate, and they are far better than the existing approximate CI in terms of coverage probability. It is hoped that this present study will be beneficial to researchers and practitioners in other areas of science.

Finally, we note that we have addressed the estimation problem based on single capture-recapture data or data from two sources. The capture-recapture methodology can be applied to multiple data-sources, by the linkage of individuals across the multiple lists. Estimation based on such multiple lists is often referred to as Multiple Systems Estimation (MSE); see Bird and King (2018). We are currently investigating estimation based on a multiple system.

Acknowledgements

The authors are grateful to a reviewer for providing useful comments and suggestions.

References

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, *52*, 119–126.

Allen, S. T., O'Rourke, A., White, R. H., Schneider, K. E., Kilkenny, M. M., & Sherman, S. G. (2019). Estimating the number of people who inject drugs in a rural county in Appalachia. *American Journal of Public Health*, *109*, 445–450. https://doi.org/10.2105/AJPH.2018.304873

Bird, S. M., & King, R. (2018). Multiple systems estimation (or capture-recapture estimation) to inform public policy. *Annual Review of Statistics and Its Application*, *5*, 95–118. https://doi.org/10.1146/annurev-statistics-031017-100641

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, *16*(2), 101–133. https://doi.org/10.1214/ss/1009213286

Casella, G., & Berger, R. L. (2002). *Statistical Inference*. Duxbury.

Chao, A., Tsay, P. K., Lin, S.-H., Shau, W.-Y., & Chao, D.-Y. (2001). The applications of capture-recapture models to epidemiological data. *Statistics in Medicine*, *20*(20), 3123–3157. https://doi.org/10.1002/sim.996

Chapman, D. G. (1951). *UC Publications in Statistics*, Vol 1 #7. University of California Press.

Chartier, M., Dart, A., Tangri, N., Komenda, P., Walld, R., Bogdanovic, B., Burchill, C., Koseva, I., McGowan, K.-L., & Rajotte, L. (2015). *Care of manitobans living with chronic kidney disease*. MB: Manitoba Centre for Health Policy.

Petersen, C. G. J. (1896). The yearly immigration of young plaice into the Limfjord from the German Sea. *Report of the Danish Biological Station for 1895*, *6*, 5–48.

Robson, D. S., & Regier, H. A. (1964). An autopsy technique for zinc-caused fish mortality. *Transactions of the American Fisheries Society*, *93*, 174–182. https://doi.org/10.1577/1548-8659(1964)93[215:SSIPME]2.0.CO;2

Sekar, C. C., & Deming, W. E. (1949). On a method of estimating birth and death rates and the extent registration. *Journal of the American Statistical Association*, *44*(245), 101–115. https://doi.org/10.1080/01621459.1949.10483294

Tilling, K. (2001). Capture-recapture methods-useful or misleading? *International Journal of Epidemiology*, *30*(1), 12–14. https://doi.org/10.1093/ije/30.1.12

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*(3), 426–482. https://doi.org/10.1090/S0002-9947-1943-0012401-3

Wang, W. (2015). Exact optimal confidence intervals for hypergeometric parameters. *Journal of the American Statistical Association*, *110*(512), 1491–1499. https://doi.org/10.1080/01621459.2014.966191

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*(158), 209–212. https://doi.org/10.1080/01621459.1927.10502953

Wittes, J. T., Colton, T., & Sidel, V. W. (1974). Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information sources. *Journal of Chronic Diseases*, *27*(1-2), 25–36. https://doi.org/10.1016/0021-9681(74)90005-8