THE AMERICAN STATISTICIAN
2023, VOL. 77, NO. 2, 151–159: General
https://doi.org/10.1080/00031305.2022.2128421

Optimal and Fast Confidence Intervals for Hypergeometric Successes

Jay Bartroff[a], Gary Lorden[b], and Lijia Wang[c]

[a]Department of Statistics and Data Sciences, University of Texas at Austin, Austin, TX; [b]Department of Mathematics, Caltech, Pasadena, CA; [c]Department of Mathematics, University of Southern California, Los Angeles, CA

ABSTRACT

We present an efficient method of calculating exact confidence intervals for the hypergeometric parameter representing the number of "successes," or "special items," in the population. The method inverts minimum-width acceptance intervals after shifting them to make their endpoints nondecreasing while preserving their level. The resulting set of confidence intervals achieves minimum possible average size, and even in comparison with confidence sets not required to be intervals it attains the minimum possible cardinality most of the time, and always within 1. The method compares favorably with existing methods not only in the size of the intervals but also in the time required to compute them. The available R package `hyperMCI` implements the proposed method.

ARTICLE HISTORY
Received May 2022
Accepted September 2022

KEYWORDS
Binary data analysis; Estimation; Exact methods; Inference; Quality control; Sampling

1. Introduction

1.1. Summary of Our Approach

This article concerns exact confidence intervals for the number M of "successes" in the hypergeometric distribution. Given integers $0 < n \leq N$ and $0 \leq M \leq N$, a random variable X has the *hypergeometric distribution* $Hyper(M, n, N)$ if

$$P_M(X = x) = \binom{M}{x}\binom{N-M}{n-x} \Big/ \binom{N}{n} \tag{1}$$

for all integer values of x such that the quotient (1) is defined, with $P_M(X = x) = 0$ otherwise.

Our approach to constructing $(1 - \alpha)$-confidence intervals for M based on X is by inverting tests of the hypotheses $H : M = M_0$, which we denote as $H(M_0)$, for $M_0 = 0, 1, \ldots, N$. For testing $H(M)$, we use acceptance intervals $[a_M, b_M]$ that *maximize* the acceptance probability $P_M(X \in [a_M, b_M])$ among all shortest possible level-α intervals, a property we call α *max optimal* which is discussed in Section 2, along with a novel method of shifting a set of α max optimal intervals so their endpoints a_M, b_M form nondecreasing sequences. This guarantees that the confidence sets that result from inversion are intervals, which is our goal here. After obtaining and shifting a set of α max optimal intervals, in Section 3 we discuss how to further modify them to make them symmetrical, and discuss the case $M = N/2$ when N is even, which needs separate handling. Our proposed confidence intervals are the inversion of these symmetrical and monotonic acceptance intervals, and Example 4.1 illustrates the process of starting with α max optimal intervals, modifying them to make them symmetrical and monotonic, and inverting them to yield confidence intervals. In Section 4 we prove the size-optimality results for the confidence intervals that result from inversion. In Section 5 we present some numerical examples and compare with two existing methods, including the notable, recent method of Wang (2015). There we also apply our method to an air quality dataset of particulate matter concentration drawn at multiple sites in China.

1.2. Background

1.2.1. The Hypergeometric Distribution

The most common setting in which the hypergeometric distribution arises is when X counts the number of items with a certain binary "special" property, sometimes called a "success," in a simple random sample (i.e., sampled uniformly without replacement) of size n from a population of size N containing M special items. But the hypergeometric arises in many other ways not involving a simple random sample, such as the analysis of a 2×2 contingency table using Fisher's Exact Test, and in other sampling schemes. Readers interested in other aspects of the hypergeometric distribution are referred to Hald (1990) for its history and naming, Keilson and Gerber (1971) for log-concavity and other properties, and Chvátal (1979) and Skala (2013) for exponential tail bounds, to name a few.

1.2.2. Exact Confidence Intervals

For exact confidence sets, there is much more literature on the related problem of the Binomial success probability than for the hypergeometric, beginning with Clopper and Pearson (1934) who applied the method of pivoting the CDF to the Binomial problem. Sterne's (1954) method for the Binomial inverts hypothesis tests with the p-value as the test statistic,

CONTACT Jay Bartroff bartroff@austin.utexas.edu Department of Statistics and Data Sciences, University of Texas at Austin, Austin, TX.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/TAS.

and he observed that the resulting intervals are "sometimes narrower" (Sterne 1954, p. 278) than the Clopper-Pearson intervals. Sterne's method can alternatively be described as inverting acceptance intervals with maximal acceptance probability, which is similar to the method we apply here to the hypergeometric. Crow (1956) showed that Sterne's (1954) method yields intervals with minimal total (or average) width, but also pointed out some "irregularities" in the method, such as occcasionally producing nonintervals, or giving longer intervals for *lower* confidence levels based on the same data. Crow (1956) proposed a modification of Sterne's method eliminating these irregularities while maintaining minimal total width. Blyth and Still (1983) proposed a further modification of Sterne's method giving intervals with more regular monotonic endpoint sequences than Sterne's and Crow's, while also achieving minimal total width. Blaker (2000, 2001) proposed an improvement of the Clopper-Pearson method giving shorter intervals, nested by confidence level, by choosing a more efficient partition of the error probabilities than the "equal tails" approach of the earlier method. The recent method of Schilling and Doi (2014) produces length-minimizing, exact intervals for the Binomial problem by shifting acceptance intervals to achieve monotonicity of endpoints before inverting; this is similar to our approach to the hypergeometric.

For the hypergeometric, pivoting the CDF was proposed by Konijn (1973) and Buonaccorsi (1987), but length-optimality was not addressed until Wang (2015), who proposed a computationally intensive method for both 1- and 2-sided intervals, and proved that the 1- sided intervals were length-minimizing. See Section 5 for a more detailed description and comparison of these methods.

Casella and Berger (2002, p. 463) give a summary of work on confidence sets for some other discrete distributions. One notable example is Crow and Gardner's (1959) for the Poisson mean, a method similar to Crow's (1956) for the binomial.

1.3. Additional Notation

Throughout the article we treat the positive integers n and N, and the desired confidence level $1-\alpha \in (0,1)$, as fixed quantities, known to the statistician, and inference centers on the unknown value of M. Since the parameter M of interest is an integer, the *intervals* we consider are actually sets of consecutive integers, which we denote by $[a,b]$ but actually mean $\{a, a+1, \ldots, b\}$. For an arbitrary set A we let $P_M(A)$ denote $P_M(X \in A)$ where $X \sim \text{Hyper}(M,n,N)$, which X will denote throughout unless otherwise specified. For a scalar x we let $P_M(x)$ denote $P_M(X = x)$. It is not hard to see from (1) that $P_M(X = x)$ is nonzero if and only if

$$x_{\min} := \max\{0, M+n-N\} \le x \le \min\{M,n\} =: x_{\max}. \quad (2)$$

We let $\lfloor y \rfloor$ denote the largest integer $\le y$ and $\lceil y \rceil$ the smallest integer $\ge y$. For sets A, B let $A \setminus B = \{a \in A : a \notin B\}$ denote the set difference and $|A|$ denote set cardinality, for example, $|[a,b]| = b-a+1$ for integers $a \le b$. For a nonnegative integer j we let $[j] = \{0, 1, \ldots, j\}$.

2. α Max Optimal Acceptance Sets and Modifying Intervals for Monotonicity

In this section we establish properties of acceptance intervals that will guarantee that they still enjoy size optimality when they are appropriately shifted to make their endpoints monotonic. The next definition makes this precise, and we call the property *α max optimal*. Theorem 2.1 shows how to modify any set of α max optimal acceptance intervals to produce intervals whose endpoints a_M, b_M are nondecreasing in M, thus, producing confidence *intervals* upon inversion rather than non-interval confidence *sets*; see also Section 4. It is not difficult to construct α max optimal acceptance intervals, and a simple and straightforward algorithm to do so is given in Algorithm S.1 in the supplementary materials, where we prove that it produces α max optimal intervals in Lemma S.2.1.

For the next definition we consider more general acceptance *sets* (not necessarily intervals): A *level-α acceptance set* for $H(M)$ is any subset $S_M \subseteq [n]$ such that

$$P_M(S_M) \ge 1-\alpha.$$

Definition 2.1. Fix n, N, and $\alpha \in (0,1)$.

1. Given $M \in [N]$, a subset $S \subseteq [n]$ is *α optimal for M* if $P_M(S) \ge 1-\alpha$ and $P_M(S^*) < 1-\alpha$ whenever $S^* \subseteq [n]$ with $|S^*| < |S|$. A collection $\{S_M : M \in \mathcal{M}\}$, $\mathcal{M} \subseteq [N]$, is *$\alpha$ optimal (for $\mathcal{M}$)* if, for all $M \in \mathcal{M}$, S_M is α optimal for M.
2. Given $M \in [N]$, a subset $S \subseteq [n]$ is *P_M-maximizing* if all elements of S have positive P_M-probability and $P_M(S) \ge P_M(S^*)$ whenever $|S^*| = |S|$. A collection $\{S_M : M \in \mathcal{M}\}$, $\mathcal{M} \subseteq [N]$, is *$P_{\mathcal{M}}$-maximizing* if, for all $M \in \mathcal{M}$, S_M is P_M-maximizing.
3. A collection $\{S_M : M \in \mathcal{M}\}$, $\mathcal{M} \subseteq [N]$, is *$\alpha$ max optimal (for $\mathcal{M}$)* if it is α optimal and $P_{\mathcal{M}}$-maximizing.

The link between the more general probability-maximizing *sets* in the definition, and intervals, is the remarkable fact that, for the hypergeometric distribution, probability-maximizing sets are *always* intervals. This result is recorded and proved as Proposition S.1.1 in the supplementary materials. That fact underlies our main result concerning α max optimal acceptance intervals in Theorem 2.1, that they can always be modified in order to make both sequences of endpoints nondecreasing in M while still being α optimal.

Theorem 2.1. Fix $n, N, \alpha \in (0,1)$. Let $\mathcal{M} \subseteq [N]$ be an arbitrary set of consecutive integers, and $\{[a_M, b_M] : M \in \mathcal{M}\}$ a set of α max optimal acceptance intervals. For $M \in \mathcal{M}$ define

$$\bar{a}_M = \max_{M' \le M} a_{M'} \quad \text{and} \quad \underline{b}_M = \min_{M' \ge M} b_{M'}. \quad (3)$$

Finally, define

$$\mathcal{M}_a = \{M \in \mathcal{M} : a_M < \bar{a}_M\} \quad \text{and}$$
$$\mathcal{M}_b = \{M \in \mathcal{M} : b_M > \underline{b}_M\}. \quad (4)$$

Then the following hold.

1. The sets $\mathcal{M}_a$ and $\mathcal{M}_b$ are disjoint.

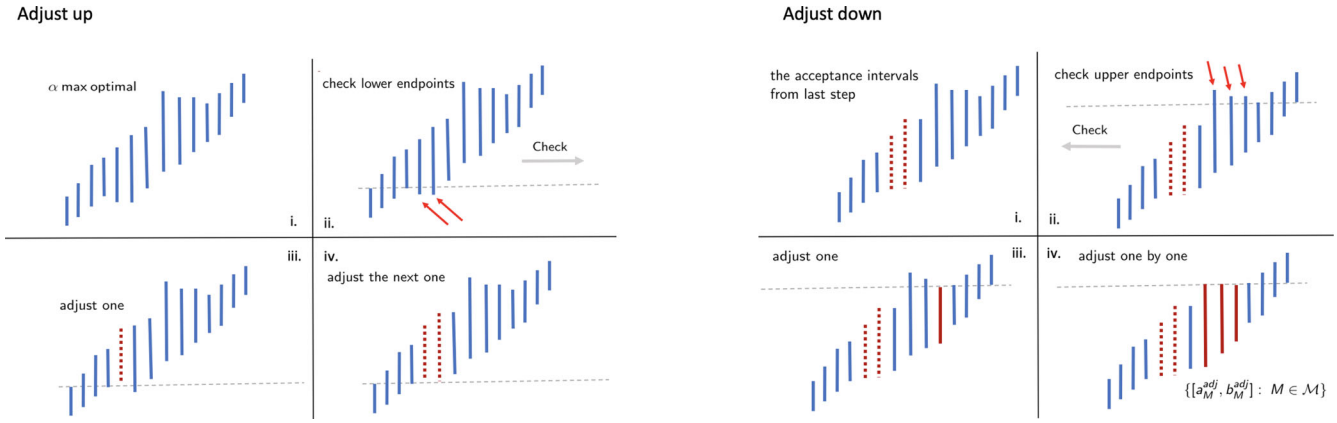

Figure 1. An illustration of the shifting process in Theorem 2.1. In the left panel the lower endpoints of two intervals with $M \in \mathcal{M}_a$ are adjusted up; in the right panel three intervals with $M \in \mathcal{M}_b$ are adjusted down.

2. The adjusted intervals

$$[a_M^{\text{adj}}, b_M^{\text{adj}}] := \begin{cases} [\overline{a}_M, b_M + (\overline{a}_M - a_M)], & M \in \mathcal{M}_a \\ [a_M - (b_M - \underline{b}_M), \underline{b}_M], & M \in \mathcal{M}_b \\ [a_M, b_M], & \text{all other } M \in \mathcal{M}, \end{cases} \tag{5}$$

are α optimal and have nondecreasing endpoint sequences.

The proof of the theorem and auxiliary results are given in Section S.3 of the supplementary materials. The process in the theorem of beginning with α max optimal intervals and shifting them to produce the monotonic, adjusted intervals (5) is illustrated in Figure 1.

3. α Optimal, Symmetrical, Nondecreasing Acceptance Intervals

3.1. Initial Modification of α Max Optimal Intervals

A set of acceptance intervals $\{[a_M, b_M] : M \in [N]\}$ is *symmetrical* if the intervals are equivariant with respect to the reflections $M \mapsto N - M$, $[a_M, b_M] \mapsto [n - b_M, n - a_M]$. That is, if

$$[a_{N-M}, b_{N-M}] = [n - b_M, n - a_M] \quad \text{for all } M \in [N]. \tag{6}$$

This can equivalently be stated as the intervals $[a_M, b_{N-M}]$, $M \in [N]$, all having midpoint $n/2$, or having endpoints summing to n. We seek symmetrical acceptance intervals because they will result in symmetrical confidence intervals (defined below analogously to (6)) upon inversion in Section 4.

Let

$$\{[a_M, b_M] : M \in [\lfloor N/2 \rfloor]\} \tag{7}$$

be α max optimal acceptance intervals, such as those produced by Algorithm S.1, and

$$\{[a_M^{\text{adj}}, b_M^{\text{adj}}] : M \in [\lfloor N/2 \rfloor]\} \tag{8}$$

the result of applying Theorem 2.1 to these intervals. One way to expand (8) to a full set $M \in [N]$ of symmetric intervals is to define $[a_M^{\text{adj}}, b_M^{\text{adj}}]$ for $M > \lfloor N/2 \rfloor$ as the reflection of the intervals (8) across $n/2$. This is guaranteed to achieve symmetry everywhere except possibly at $M = N/2$ when N is even and $[a_{N/2}^{\text{adj}}, b_{N/2}^{\text{adj}}]$ is not symmetric about $n/2$. This is the strategy taken in Theorem 3.1, with the $M = N/2$ interval taken to be (9), and the resulting intervals are α optimal, symmetrical, and have nondecreasing endpoint sequences. Algorithm S.2 in the supplementary materials produces the result of applying Theorem 3.1 to α max optimal intervals, which could be those produced by Algorithm S.1 or any other α max optimal intervals (7).

3.2. Reflection and Modification at $N/2$

Starting with a set of α max optimal intervals $\{[a_M, b_M] : M \in [\lfloor N/2 \rfloor]\}$ we will now define a new set of intervals $\{[a_M^*, b_M^*] : M \in [N]\}$ by (i) applying the adjustments in Theorem 2.1, (ii) reflecting across $n/2$ to obtain symmetrical intervals for $M > \lfloor N/2 \rfloor$, and (iii) if N is even setting $[a_{N/2}^*, b_{N/2}^*]$ to be the interval

$$[h_{\alpha/2}, n - h_{\alpha/2}], \quad \text{where} \quad h_{\alpha/2} = \max\left\{x \in [n] : P_{N/2}(X < x) \le \alpha/2\right\}. \tag{9}$$

The next theorem establishes that the resulting intervals are α optimal, symmetrical, and have nondecreasing endpoint sequences.

Theorem 3.1. Given a set of α max optimal intervals $\{[a_M, b_M] : M \in [\lfloor N/2 \rfloor]\}$, let $\{[a_M^{\text{adj}}, b_M^{\text{adj}}] : M \in [\lfloor N/2 \rfloor]\}$ denote the result of applying Theorem 2.1, and

$$[a_M^*, b_M^*] = \begin{cases} [a_M^{\text{adj}}, b_M^{\text{adj}}], & \text{for } M = 0, 1, \ldots, \lceil N/2 \rceil - 1; \\ [n - b_{N-M}^{\text{adj}}, n - a_{N-M}^{\text{adj}}], & \text{for } M = \lfloor N/2 \rfloor + 1, \ldots, N; \\ [h_{\alpha/2}, n - h_{\alpha/2}], & \text{for } M = N/2 \text{ if } N \text{ is even.} \end{cases} \tag{10}$$

Then $\mathcal{A}^* := \{[a_M^*, b_M^*] : M \in [N]\}$ are level-α, symmetrical, have nondecreasing endpoint sequences, and are size-optimal except possibly for $M = N/2$ when N is even; in this case, $[a_{N/2}^*, b_{N/2}^*]$ is size-optimal unless $[a_{N/2}^*, b_{N/2}^* - 1]$ has probability $1 - \alpha$ or greater, in which case $[a_{N/2}^*, b_{N/2}^* - 1]$ is α optimal and $\mathcal{A}^*$ is larger by one in total size than an α optimal collection. In any case, $\mathcal{A}^*$ is α optimal among symmetrical collections.

The proof of the theorem is in Section S.4 of the supplementary materials.

Note that if N is odd then the first two cases of (10) cover all $M \in [N]$. When N is even, the $M = N/2$ interval (9) is clearly the smallest symmetrical level-α acceptance interval for $H(N/2)$, and is α optimal. Also, we have $h_{\alpha/2} \leq n/2$ since

$$P_{N/2}(X < \lfloor n/2 \rfloor + 1) = P_{N/2}(X \leq \lfloor n/2 \rfloor) \geq 1/2 > \alpha/2. \quad (11)$$

Finally, it is not necessary to use special calculations to get $h_{\alpha/2}$ since it is easily obtained from an α max optimal interval $[a_{N/2}, b_{N/2}]$ by $h_{\alpha/2} = \min\{a_{N/2}, n - b_{N/2}\}$. Note that if $[a_{N/2}, b_{N/2}]$ is already symmetrical, then (9) is the same interval.

4. Optimal Symmetrical Confidence Intervals

4.1. Confidence and Acceptance Sets

For a set $\mathcal{S}$ let $2^{\mathcal{S}}$ denote the power set of $\mathcal{S}$, that is, the set of all subsets of $\mathcal{S}$. A *confidence set with confidence level* $1 - \alpha$ is a function $\mathcal{C} : [n] \to 2^{[N]}$ such that the coverage probability satisfies $P_M(M \in \mathcal{C}(X)) \geq 1 - \alpha$ for all $M \in [N]$. For short, we refer to such a $\mathcal{C}$ as a $(1 - \alpha)$-confidence set. If a confidence set $\mathcal{C}$ is interval-valued (i.e., for all $x \in [n]$, $\mathcal{C}(x)$ is an interval) we call it a *confidence interval*. A confidence set $\mathcal{C}$ is *symmetrical* if

$$\mathcal{C}(x) = N - \mathcal{C}(n - x) \quad \text{for all } x \in [n]. \quad (12)$$

Here, for a set $\mathcal{S}$, the notation $N - \mathcal{S}$ means $\{N - s : s \in \mathcal{S}\}$. Symmetry (12) is an equivariance condition requiring that the confidence set is reflected about $N/2$ when the data is reflected about $n/2$. See also Section 5 for how this definition compares with the regularity conditions of Wang (2015).

Similarly, we shall denote a level-α acceptance set by a function $\mathcal{A} : [N] \to 2^{[n]}$ such that $P_M(X \in \mathcal{A}(M)) \geq 1 - \alpha$ for all $M \in [N]$, and call an interval-valued (i.e., for all $M \in [N]$, $\mathcal{A}(M)$ is an interval) acceptance set an acceptance interval and write $\mathcal{A}(M) = [a_M, b_M]$, or similar. Note that whereas above we referred to an expression like (7) as a set of acceptance intervals, we will now call it *an* acceptance interval (singular). This is to coincide with our terminology for a confidence set, as well as avoid cumbersome phrases like "a set of acceptance sets."

We also need to generalize the concept of symmetry from (6) to handle general sets, so we say that an acceptance set $\mathcal{A}$ is *symmetrical* if $\mathcal{A}(M) = n - \mathcal{A}(N - M)$ for all $M \in [N]$. This says that the set is equivariant with respect to reflections $M \mapsto N - M$, and specializes to (6) for intervals.

4.2. Inverted Confidence Sets

We will construct confidence sets that are inversions of acceptance sets, and vice-versa. If $\mathcal{A}$ is a level-α acceptance set, then

$$\mathcal{C}_{\mathcal{A}}(x) = \{M \in [N] : x \in \mathcal{A}(M)\} \quad (13)$$

is a $(1 - \alpha)$-confidence set. Conversely, given a $(1 - \alpha)$-confidence set $\mathcal{C}$, $\mathcal{A}_{\mathcal{C}}(M) = \{x \in [n] : M \in \mathcal{C}(x)\}$ is a level-α acceptance set; see, for example, (Rice 2007, chap. 9.3). Moreover, $\mathcal{C}_{\mathcal{A}_{\mathcal{C}}} = \mathcal{C}$ and $\mathcal{A}_{\mathcal{C}_{\mathcal{A}}} = \mathcal{A}$, which are immediate from the definitions. However, neither $\mathcal{A}$ nor $\mathcal{C}$ being interval-valued guarantees that its inversion is.

We will evaluate confidence and acceptance sets by their *total size*, which we define as the sum of the cardinalities of each set: Recalling that $|\cdot|$ denotes set cardinality, define the *total size* of acceptance and confidence sets to be

$$|\mathcal{A}| = \sum_{M=0}^{N} |\mathcal{A}(M)| \quad \text{and} \quad |\mathcal{C}| = \sum_{x=0}^{n} |\mathcal{C}(x)|.$$

If $\mathcal{A}(M) = [a_M, b_M]$ is an acceptance interval, then

$$|\mathcal{A}| = \sum_{M=0}^{N} |[a_M, b_M]| = \sum_{M=0}^{N} (b_M - a_M + 1),$$

and similarly for a confidence interval $\mathcal{C}$.

Lemma 4.1 records some basic facts about inverted confidence sets, and is proved in Section S.6 of the supplementary materials.

Lemma 4.1. Let $\mathcal{A}$ be an acceptance set. Then the following hold.

1. $$|\mathcal{C}_{\mathcal{A}}| = |\mathcal{A}|. \quad (14)$$
2. $\mathcal{C}_{\mathcal{A}}$ is symmetrical if and only if $\mathcal{A}$ is symmetrical.
3. If, in addition, $\mathcal{A}(M) = [a_M, b_M]$ is interval-valued and the endpoint sequences $\{a_M\}$ and $\{b_M\}$ are nondecreasing, then $\mathcal{C}_{\mathcal{A}}$ is interval-valued.

4.3. Size Optimality

We say that a confidence set $\mathcal{C}$ is *size-optimal* among a collection of confidence sets if it achieves the minimum total size in that collection. The results in this section establish size-optimality of $\mathcal{C}^* = \mathcal{C}_{\mathcal{A}^*}$, where $\mathcal{A}^* = \{[a^*_M, b^*_M] : M \in [N]\}$ denotes the result of applying Theorem 3.1 to any α max optimal acceptance intervals $\{[a_M, b_M] : M \in [N]\}$. Thus, $\mathcal{A}^*$ could be the intervals given by Algorithm S.2, or the result of starting with any other α max optimal intervals. Whatever the choice of $\mathcal{A}^*$, note that $\mathcal{C}^*$ is a symmetrical, $(1 - \alpha)$-confidence interval by Lemma 4.1.

Before discussing the optimality of $\mathcal{C}^*$ in Theorems 4.1 and 4.2, we give a short example of its construction, starting with α max optimal acceptance intervals, their modification, and inversion to produce $\mathcal{C}^*$.

Example 4.1. Take $N = 23$, $n = 5$, and $\alpha = 0.54$, chosen only for the sake of example. The first few α max optimal acceptance intervals produced by Algorithm S.1 are

$$[a_M, b_M] = \begin{cases} \{0\}, & M = 0, 1, 2, 3 \\ \{1\}, & M = 4 \\ [0, 1], & M = 5 \\ [1, 2], & M = 6, 7, 8, 9, \end{cases} \quad (15)$$

and satisfy $a_M \geq 3$ for $M \geq 10$. Note the violation of monotonicity in the lower endpoints at $a_4 = 1 > 0 = a_5$. Applying Theorem 3.1 to these intervals therefore shifts up the $M = 5$ interval yielding $\mathcal{A}^*(5) = [1, 2]$, with $\mathcal{A}^*(M)$ unchanged for the other M in (15). Then the first three confidence intervals resulting from the inversion of $\mathcal{A}^*$ are $\mathcal{C}^*(0) = [0, 3]$, $\mathcal{C}^*(1) = [4, 9]$, and $\mathcal{C}^*(2) = [5, 9]$.

Theorems 4.1 and 4.2, which follow, are the main results of the article. Theorem 4.1 is the more powerful of the two in that it gives wide conditions under which $\mathcal{C}^*$ is size-optimal among symmetrical confidence *sets* (not just intervals) and shows that, even in the worst case, the total size $|\mathcal{C}^*|$ is at most 1 point larger than the optimal set. Theorem 4.2 specializes to intervals and gives conditions for optimality there. In particular, it shows that $\mathcal{C}^*$ is size-optimal among all symmetrical non-empty (i.e., $\mathcal{C}(x) \neq \emptyset$ for all x) intervals, which are usually preferred in practice.

Theorem 4.1. Let $\mathcal{C}^*$ be as defined above and $\mathfrak{C}_S$ the class of all symmetrical, $(1-\alpha)$-confidence sets. Then $\mathcal{C}^*$ is size-optimal in $\mathfrak{C}_S$, that is,

$$|\mathcal{C}^*| = \min_{\mathcal{C}\in\mathfrak{C}_S} |\mathcal{C}|,$$

if either of the following holds:

1. n or N is odd;
2. n, N are even and there is no size-optimal $\mathcal{C} \in \mathfrak{C}_S$ such that $|\mathcal{A}_\mathcal{C}(N/2)|$ is even. If n, N, and $|\mathcal{A}_\mathcal{C}(N/2)|$ are all even for some size-optimal $\mathcal{C} \in \mathfrak{C}_S$, then

$$|\mathcal{C}^*| \leq \min_{\mathcal{C}\in\mathfrak{C}_S} |\mathcal{C}| + 1. \tag{16}$$

In addition, $\mathcal{C}^*$ is size-optimal among all $\mathcal{C} \in \mathfrak{C}_S$ such that $\mathcal{A}_\mathcal{C}$ are all intervals.

The proofs of Theorems 4.1 and 4.2 use some auxiliary lemmas, stated and proved in Section S.6 of the supplementary materials. See also Example 5.3 for an instance of $\mathcal{C}^*$ failing to be optimal as a confidence set, under conditions satisfying part 2.

Proof of Theorem 4.1. First suppose N is odd, and let $\mathcal{C} \in \mathfrak{C}_S$ be arbitrary; we will show that $|\mathcal{C}^*| \leq |\mathcal{C}|$. Since $N/2$ is not an integer, by Lemma S.6.1 we have $|\mathcal{A}^*(M)| \leq |\mathcal{A}_\mathcal{C}(M)|$ for all $M \in [N]$ thus, using Lemma 4.1,

$$|\mathcal{C}| = |\mathcal{A}_\mathcal{C}| = \sum_{M=0}^{N} |\mathcal{A}_\mathcal{C}(M)| \geq \sum_{M=0}^{N} |\mathcal{A}^*(M)| = |\mathcal{A}^*| = |\mathcal{C}^*|, \tag{17}$$

as claimed.

Now suppose N is even and let $\mathcal{C} \in \mathfrak{C}_S$ be size-optimal. By Lemma S.6.1 we have $|\mathcal{A}^*(M)| \leq |\mathcal{A}_\mathcal{C}(M)|$ for all $M \in [N]$ other than $M = N/2$. If n or $|\mathcal{A}_\mathcal{C}(N/2)|$ is odd, then by Lemma S.6.2 there is an interval $[a, n-a]$ such that $n - 2a + 1 = |\mathcal{A}_\mathcal{C}(N/2)|$ and $P_{N/2}([a, n-a]) \geq P_{N/2}(\mathcal{A}_\mathcal{C}(N/2))$. Since $\mathcal{A}^*(N/2) = [a^*_{N/2}, b^*_{N/2}] = [a^*_{N/2}, n - a^*_{N/2}]$ is the shortest symmetrical acceptance interval for $M = N/2$, we have

$$|\mathcal{A}^*(N/2)| = b^*_{N/2} - a^*_{N/2} + 1 \leq n - 2a + 1 = |\mathcal{A}_\mathcal{C}(N/2)|.$$

This, with the above inequality for the $M \neq N/2$ cases, establishes (17) in this case.

The remaining case—when N, n, and $|\mathcal{A}_\mathcal{C}(N/2)|$ are all even—is handled by Lemma S.6.3, recalling that $\mathcal{C}$ was size-optimal to establish (16).

For the final statement in the theorem, for any such $\mathcal{C}$, $\mathcal{A}_\mathcal{C}$ is symmetrical and thus has total size at least $|\mathcal{A}^*|$, so $|\mathcal{C}| = |\mathcal{A}_\mathcal{C}| \geq |\mathcal{A}^*| = |\mathcal{C}^*|$. □

Theorem 4.2. Let $\mathcal{C}^*$ be as defined above and $\mathfrak{C}_I$ the class of all symmetrical, $(1-\alpha)$-confidence intervals. Then $\mathcal{C}^*$ is size-optimal in $\mathfrak{C}_I$, that is,

$$|\mathcal{C}^*| = \min_{\mathcal{C}\in\mathfrak{C}_I} |\mathcal{C}|,$$

if either of the following holds:

1. n or N is odd;
2. n, N are even and there is no size-optimal $\mathcal{C} \in \mathfrak{C}_I$ such that

$$\mathcal{C}(n/2) = \varnothing. \tag{18}$$

A sufficient condition for $\mathcal{C}^*$ to be size-optimal in this case is that

$$\alpha < \binom{N/2}{n/2}^2 \Big/ \binom{N}{n}. \tag{19}$$

In particular, $\mathcal{C}^*$ is size-optimal among all nonempty $\mathcal{C} \in \mathfrak{C}_I$ regardless of the parity of n, N.

We comment that the scenario (18) seems to be particularly rare since $\mathcal{C}(n/2)$ is typically the widest confidence interval. Thus, even allowing empty intervals, Theorem 4.2 establishes size optimality of $\mathcal{C}^*$ among intervals for most intents and purposes, and (16) holds in any case. However, it may be possible to construct an adversarial example with that property.

Proof of Theorem 4.2. Part 1 is a consequence of Theorem 4.1 since $\mathfrak{C}_I \subseteq \mathfrak{C}_S$.

Assume N and n are even, and there is no size-optimal $\mathcal{C}$ satisfying (18). Let $\mathcal{C}$ be any size-optimal interval and since $\mathcal{C}(n/2) \neq \varnothing$, there is some $M \in \mathcal{C}(n/2)$. Because $\mathcal{C}(n/2)$ is symmetrical, $N - M \in \mathcal{C}(n/2)$, and because $\mathcal{C}(n/2)$ is an interval, $N/2 \in \mathcal{C}(n/2)$ since it lies between M and $N - M$. This implies that $n/2 \in \mathcal{A}_\mathcal{C}(N/2)$, which is symmetrical about $n/2$. Using these facts,

$$\begin{aligned} |\mathcal{A}_\mathcal{C}(N/2)| &= 2|\{x \in \mathcal{A}_\mathcal{C}(N/2) \mid x < n/2\}| + |\{n/2\}| \\ &= 2|\{x \in \mathcal{A}_\mathcal{C}(N/2) \mid x < n/2\}| + 1, \end{aligned}$$

an odd number. We then have $|\mathcal{C}^*| \leq |\mathcal{C}|$ by Lemma S.6.3.

To see that (19) is sufficient, suppose there is a $\mathcal{C}$ with $\mathcal{C}(n/2) = \varnothing$. Then for $M = N/2$, we have

$$\alpha \geq P_M(M \notin \mathcal{C}(X)) \geq P_M(X = n/2) = \binom{N/2}{n/2}^2 \Big/ \binom{N}{n}.$$

□

5. Examples and Comparisons

In this section we show examples of our proposed method $\mathcal{C}^*$ using Algorithm S.2 in the supplementary materials as the acceptance interval $\mathcal{A}^*$, and give some comparisons with other methods. All calculations of our method were performed using the R package `hyperMCI`, available at *github.com/bartroff792/hyper*.

For comparisons we focus on *exact* methods with guaranteed coverage probability. A standard method for producing a $(1-\alpha)$-confidence interval for M is the so-called *method of pivoting the CDF*, giving $\mathcal{C}_{\text{Piv}}(x) = [L_{\text{Piv}}(x), U_{\text{Piv}}(x)]$ where, for fixed nonnegative $\alpha_1 + \alpha_2 = \alpha$,

$$L_{\text{Piv}}(x) = \min\{M \in [N] : P_M(X \geq x) > \alpha_1\},$$
$$U_{\text{Piv}}(x) = \max\{M \in [N] : P_M(X \leq x) > \alpha_2\}. \tag{20}$$

Taking $\alpha_1 = \alpha_2 = \alpha/2$ is a common choice, and all our calculations of $\mathcal{C}_{\text{Piv}}$ below use this. See Buonaccorsi (1987), Casella and Berger (2002, chap. 9), or Konijn (1973). This method is alternatively called the *quantile method* and the *method of extreme tails*.

Wang (2015) proposed a method producing a $(1-\alpha)$-confidence interval for M, which we denote by $\mathcal{C}_W$, that cycles through the intervals $\mathcal{C}_{\text{Piv}}(x)$, shrinking the intervals where possible while checking that coverage probability is maintained. The algorithm can require multiple passes through the intervals, calculating the coverage probability for all $M \in [N]$ multiple times, and is therefore computationally intensive. We compare the computational times of $\mathcal{C}_W$ and $\mathcal{C}^*$ in Examples 5.1 and 5.2. All calculations of $\mathcal{C}_W(x)$ were performed using that author's R code.

Although W. Wang proves that a 1-sided version of his algorithm produces size-optimal intervals (among 1-sided intervals), it is not claimed that $\mathcal{C}_W$ is size-optimal. Since $\mathcal{C}_W$ produces non-empty intervals we know that $|\mathcal{C}^*| \leq |\mathcal{C}_W|$ by Theorem 4.2. In the following example we compare $\mathcal{C}^*$ with $\mathcal{C}_W$ in terms of both size and computational time, and indeed exhibit a setting where $|\mathcal{C}^*| < |\mathcal{C}_W|$. We also note that the regularity conditions assumed in W. Wang's results are slightly more restrictive than our symmetry condition (12), which W. Wang calls a "natural restriction," by including two additional requirements that both sequences of endpoints of $\mathcal{C}_W(x)$ be nondecreasing in x, and any sub-interval of $\mathcal{C}_W(x)$ must have confidence level strictly less than $1-\alpha$. Our $\mathcal{C}^*$ satisfies these additional properties, and see also Figure S.1 for an example of monotonicity of $\mathcal{C}^*$. However, we do not require them of the confidence sets considered so that our optimality results apply to a broader class.

Example 5.1. We compare $\mathcal{C}^*$, $\mathcal{C}_{\text{Piv}}$, and $\mathcal{C}_W$ in the setting $\alpha = 0.05$, $N = 500$, and $n = 10, 20, 30, \ldots, 490$. The $\mathcal{C}^*$ intervals are much shorter than the $\mathcal{C}_{\text{Piv}}$ intervals in this setting, and Figure 2 shows the differences in size $|\mathcal{C}_{\text{Piv}}| - |\mathcal{C}^*|$ for $n = 10, 20, \ldots, 490$ which are substantial; all the $\mathcal{C}^*$ intervals are at least 200 points shorter than their corresponding $\mathcal{C}_{\text{Piv}}$ intervals, and some are as many as 260 points shorter. These differences are also sizable fractions of the largest possible range $[0, N] = [0, 500]$.

The $\mathcal{C}_W$ intervals are very similar to $\mathcal{C}^*$ and so are not shown in Figure 2. In fact, the sizes $|\mathcal{C}_W| = |\mathcal{C}^*|$ are exactly equal for all values of n considered, except $n = 100$. The confidence intervals for this case are given explicitly in Tables S.1–S.2 in the supplementary materials. These tables show very similar, but slightly different intervals, with neither method dominating the other. For example, $|\mathcal{C}^*(0)| = |[0, 14]| < |[0, 16]| = |\mathcal{C}_W(0)|$ and

$$|\mathcal{C}^*(13)| = |[40, 102]| > |[40, 101]| = |\mathcal{C}_W(13)|. \tag{21}$$

Totaling the sizes gives $|\mathcal{C}^*| = 7129 < 7131 = |\mathcal{C}_W|$, showing that the $\mathcal{C}_W$ intervals are indeed nonoptimal. One property of our method is that it does not necessarily producing intervals that are sub-intervals of $\mathcal{C}_{\text{Piv}}$, which $\mathcal{C}_W$ always does since it begins with these intervals before iteratively shrinking them. For example, in this setting $\mathcal{C}_{\text{Piv}}(13) = [39, 101]$ which, by (21), contains $\mathcal{C}_W(13)$ but not $\mathcal{C}^*(13)$.

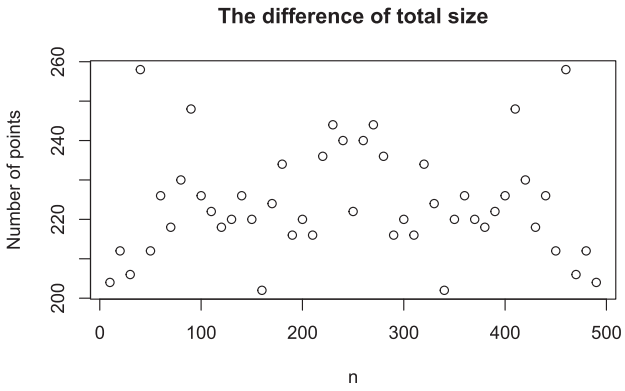

Figure 2. The differences in total size $|\mathcal{C}_{\text{Piv}}| - |\mathcal{C}^*|$, for $N = 500$, $\alpha = 0.05$, and $n = 10, 20, \ldots, 490$.

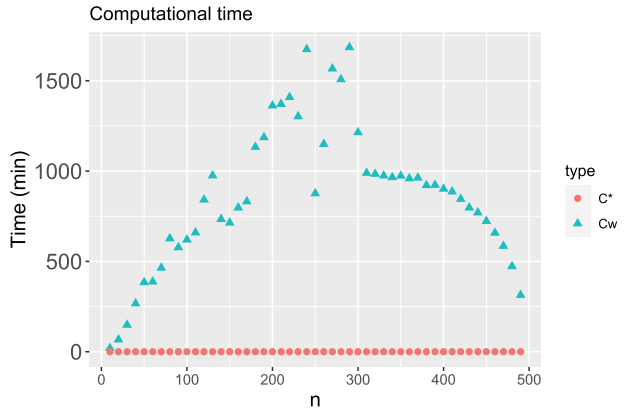

Figure 3. The computational time of the confidence intervals $\mathcal{C}_W$ and $\mathcal{C}^*$ for $N = 500$, $\alpha = 0.05$, and $n = 10, 20, \ldots, 490$.

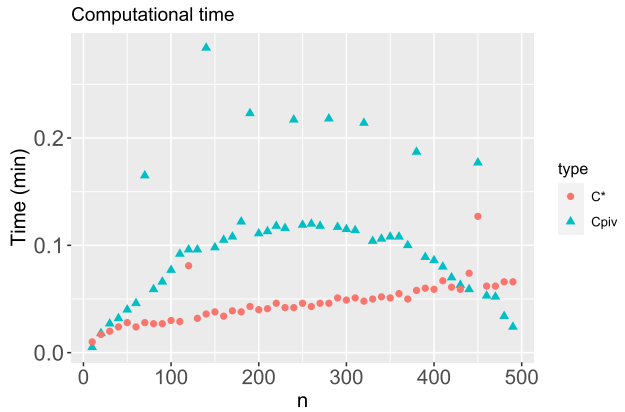

Figure 4. The computational time of the confidence intervals $\mathcal{C}_{\text{Piv}}$ and $\mathcal{C}^*$ for $N = 500$, $\alpha = 0.05$, and $n = 10, 20, \ldots, 490$.

In addition to the total sizes, Tables S.1–S.2 also show the computational times used by both methods, at the bottom of each table. All times were computed using R's `proc.time()` function. Whereas $\mathcal{C}^*$ took roughly 1/10th of a second (0.0019 min) to fill the table, $\mathcal{C}_W$ took more than 10 min. As mentioned above, this is due to the adjusting technique of $\mathcal{C}_W$ which requires repeated updating of intervals, whereas $\mathcal{C}^*$ just requires one pass through the acceptance intervals for adjustment. Figure 3 gives a more complete comparison of compu-

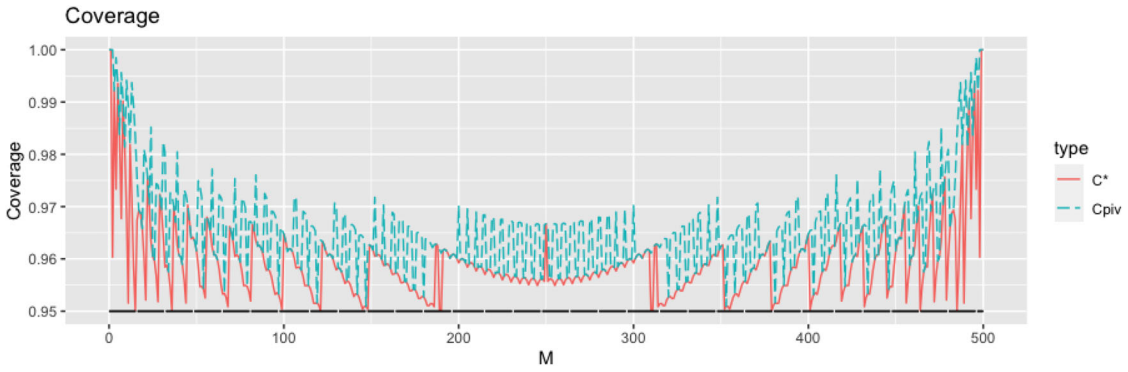

Figure 5. Coverage probability of $\mathcal{C}^*$ and $\mathcal{C}_{\text{Piv}}$ for $N = 500$, $n = 100$, and $\alpha = 0.05$.

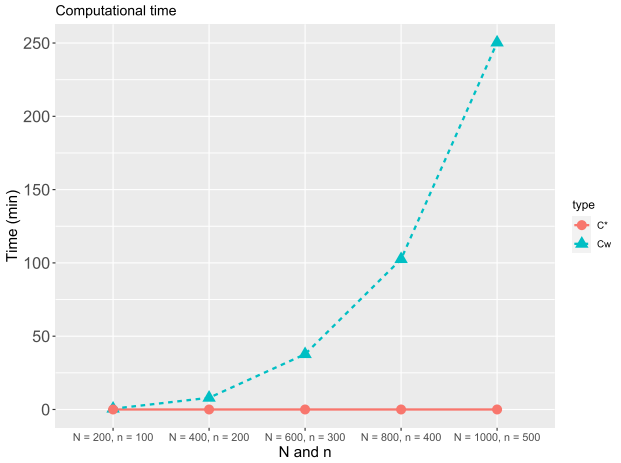

Figure 6. The computational time of the confidence intervals $\mathcal{C}_W$ and $\mathcal{C}^*$ for $N = 200, 400, \ldots, 1000$, and $n = N/2$.

tational times in this setting. The additional time required by $\mathcal{C}_W$ is sizable, even exceeding 25 min for values of n near the middle of the range. A comparison of computational times of $\mathcal{C}_{\text{Piv}}$ and $\mathcal{C}^*$ is shown in Figure 4, which shows that the times are much faster overall compared to $\mathcal{C}_W$ (the longest times being less than 1/3 of a second), and comparable between the two methods.

Figure 5 shows the coverage probability of $\mathcal{C}^*$ and $\mathcal{C}_{\text{Piv}}$ for the $n = 100$ case as a function of $M = 0, 1, \ldots 500$. The coverage probability of $\mathcal{C}_{\text{Piv}}$ is overall higher than that of $\mathcal{C}^*$, an *undesirable* property once it exceeds $1 - \alpha$. The coverage probability of $\mathcal{C}_W$ is very similar to that of $\mathcal{C}^*$ in this setting, and is shown in Figure S.2 of the supplementary materials.

Figure S.1 in the supplementary materials is a plot of the $\mathcal{C}^*$ intervals for the $n = 100$ case.

Example 5.2. We compare the computational time of $\mathcal{C}^*$ and $\mathcal{C}_W$ in the setting $\alpha = 0.05$, $N = 200, 400, \ldots, 1000$, and $n = N/2$. The computational times of $\mathcal{C}^*$ and $\mathcal{C}_W$ are shown in Figure 6. When $N = 1000$, $n = 500$ and $\alpha = 0.05$, the computational time for $\mathcal{C}_W$ is 250 min while that of $\mathcal{C}^*$ is 0.0111 min.

Example 5.3. In this example we show the necessity of part 2 of Theorem 4.1. That is, we exhibit a setting with n, N, and $\mathcal{A}(N/2)$ all even for a certain acceptance set $\mathcal{A}$ whose inversion $\mathcal{C}$ is size-optimal with $|\mathcal{C}| = |\mathcal{C}^*| - 1$. Set $N = 20$, $n = 6$, and $\alpha = 0.6$. For $M \neq N/2 = 10$ define $\mathcal{A}(M) = [a_M^*, b_M^*]$ to be the same intervals given by Theorem 3.1 and inverted to create $\mathcal{C}^*$, and define $\mathcal{A}(10) = \{2, 4\}$. For all $M \neq 10$, $\mathcal{A}(M)$ is a level-α interval, and $\mathcal{A}(10)$ is as well since

$$P_{M=10}(2) = P_{M=10}(4) = 0.244$$

to 3 decimal places, thus $P_{M=10}(\{2, 4\}) > 0.4 = 1 - \alpha$. It can be shown that $\mathcal{A}^*(10) = [2, 4]$, thus, the (noninterval) set $\mathcal{A}$ has 1 fewer point than $\mathcal{A}^*$, so by (14) we have that $|\mathcal{C}| = |\mathcal{C}^*| - 1$.

Example 5.4 (Air quality data). In this example we apply our confidence interval $\mathcal{C}^*$ to data collected by China's Ministry of Environmental Protection (MEP) and discussed by Liang et al. (2016). The MEP collects data on particulate matter ($PM_{2.5}$) concentration, measured in $\mu g/m^3$, of fine inhalable particles with diameters less than 2.5 micrometers. The U.S. Environmental Protection Agency (2012) classifies the air quality of a given day as "hazardous" if the day's 24-hour average $PM_{2.5}$ measurement exceeds the set threshold 250.5. Liang et al. (2016) analyzed the 2013–2015 MEP data and concluded that it was consistent with measurements taken at nearby U.S. diplomatic posts, the U.S. Embassy in Beijing and four U.S. Consulates in other cities. However, a persistent problem with the MEP data is a high degree of missing days. For a given year, if the missing days are assumed to be missing at random with each day of the year equally likely, then the number X of remaining "hazardous" days, conditioned on the number n of remaining days, follows a hypergeometric distribution with $N = 365$ and unknown actual number M of annual hazardous days, to be estimated as an indication of annual air quality. We focus on the 2015 data from 3 MEP sites in Beijing: Dongsi, Dongsihuan, and Nongzhanguan. For each of these sites, Table 1 shows the number n of days with complete measurements, the observed number x of days with complete measurements classified as hazardous, the point estimate Nx / n (with $N = 365$) of the number M of annual hazardous days, and the 90% confidence interval $\mathcal{C}^*(x)$ for M, which are also plotted in Figure 7. The point estimates from the MEP sites are similar to and surround the estimate at the U.S. Embassy data, similar to the conclusions drawn by Liang et al. (2016). But the confidence intervals also show that the MEP estimates are more variable, largely in the direction of indicating worse air quality, with two out of three upper confidence limits being much larger for the MEP sites than for the U.S. Embassy. For comparison, we note that the

Table 1. For the Beijing air quality data (Liang et al. 2016), the number n of days with complete measurements, the number x of days with complete measurements classified as hazardous, the point estimate Nx/n (to 1 decimal place) of the number M of annual hazardous days, and the 90% confidence interval $\mathcal{C}^*(x)$ for M.

Site	n	x	Nx/n	90% CI for M
Dongsi	292	16	20.0	[17, 24]
Dongsihuan	166	7	15.4	[10, 24]
Nongzhanguan	290	11	13.8	[11, 17]
U.S. Embassy	332	15	16.5	[15, 18]

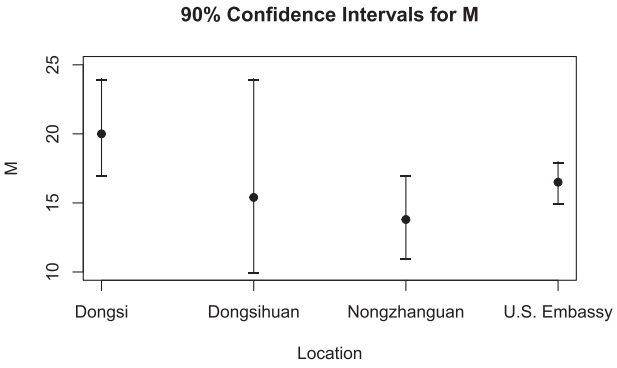

Figure 7. The 90% confidence interval $\mathcal{C}^*(x)$ for the number M of annual hazardous days at different locations in Beijing.

intervals $\mathcal{C}_W$ in this setting coincide with the intervals $\mathcal{C}^*$ in Table 1 and Figure 7, except for the Nongzhanguan data where $\mathcal{C}_W$ produces the shorter interval [12, 17].

6. Discussion

We have presented an efficient method of computing exact hypergeometric confidence intervals. Compared to the standard pivotal method, our method requires similar computational time but produces much shorter intervals. Our method produces intervals with total size no larger than, and strictly smaller than in some cases, the existing nearly optimal method of Wang (2015), which is computationally much more costly than our and the pivotal method. Therefore, we hope our method can provide something near the "best of both worlds" for this problem in terms of computational time and interval size, at least for two-sided intervals.

In practice there are many applications, such as quality control, where one-sided confidence intervals are desired. For these the method of Wang (2015) provides optimal intervals. On the other hand, there may be situations where two-sided intervals are appropriate but the statistician prefers the error probability on each side to be bounded by $\alpha/2$. For example, if the statistician wants the option to, post-hoc, use one or both endpoints of the interval as a "one-sided" confidence bound. Our proposed method does not satisfy this property because it would prevent size-optimality, which is our focus here, but in this case the statistician can use the pivotal method (20) with $\alpha_1 = \alpha_2 = \alpha/2$.

The key to our method is the novel shifting of acceptance intervals before inversion, developed in Sections 2 and 3. We have observed in the numerical examples included in Section 5, as well as extensive further computations not included in this article, that the needed shifts in Theorem 2.1 seem to never exceed a single point. This is not needed in our theory but we close by mentioning it as a tantalizing conjecture.

A similar approach to the one here of shifting optimal acceptance regions before inverting can be used to produce optimal confidence intervals for the hypergeometric population size N when it is unknown, such as in capture-recapture problems (Bailey 1951; Wittes 1972; Pollock et al. 1990). A forthcoming work will cover this problem.

Supplementary Materials

Supplementary materials are available online and include proofs, auxiliary results, and additional tables and figures.

Disclosure Statement

The authors report there are no competing interests to declare.

References

Bailey, N. T. (1951), "On Estimating the Size of Mobile Populations from Recapture Data," *Biometrika*, 38, 293–306. [158]

Blaker, H. (2000), "Confidence Curves and Improved Exact Confidence Intervals for Discrete Distributions," *Canadian Journal of Statistics*, 28, 783–798. [152]

——— (2001), "Corrigenda: Confidence Curves and Improved Exact Confidence Intervals for Discrete Distribution," *Canadian Journal of Statistics*, 29, 681–681. [152]

Blyth, C. R., and Still, H. A. (1983), "Binomial Confidence Intervals," *Journal of the American Statistical Association*, 78, 108–116. [152]

Buonaccorsi, J. P. (1987), "A Note on Confidence Intervals for Proportions in Finite Populations," *The American Statistician*, 41, 215–218. [152,156]

Casella, G., and Berger, R. L. (2002), *Statistical Inference*, Belmont, CA: Duxbury Press. [152,156]

Chvátal, V. (1979), "The Tail of the Hypergeometric Distribution," *Discrete Mathematics*, 25, 285–287. [151]

Clopper, C. J., and Pearson, E. S. (1934), "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, 26, 404–413. [151]

Crow, E. L. (1956), "Confidence Intervals for a Proportion," *Biometrika*, 43, 423–435. [152]

Crow, E. L., and Gardner, R. S. (1959), "Confidence Intervals for the Expectation of a Poisson Variable," *Biometrika*, 46, 441–453. [152]

Hald, A. (1990), *A History of Probability and Statistics and Their Applications Before 1750*, New York: Wiley. [151]

Keilson, J., and Gerber, H. (1971), "Some Results for Discrete Unimodality," *Journal of the American Statistical Association*, 66, 386–389. [151]

Konijn, H. S. (1973), *Statistical Theory of Sample Survey Design and Analysis*, Amsterdam: North-Holland Publishing Company. [152,156]

Liang, X., Li, S., Zhang, S., Huang, H., and Chen, S. X. (2016), "$PM_{2.5}$ Data Reliability, Consistency, and Air Quality Assessment in Five Chinese Cities," *Journal of Geophysical Research: Atmospheres*, 121, 10–220. Datasets available at *https://archive.ics.uci.edu/ml/datasets/PM2.5+Data+of+Five+Chinese+Cities*. [157,158]

Pollock, K. H., Nichols, J. D., Brownie, C., and Hines, J. E. (1990), "Statistical Inference for Capture-Recapture Experiments," *Wildlife Society Monographs*, 107, 3–97. [158]

Rice, J. A. (2007), *Mathematical Statistics and Data Analysis* (3rd ed.), Belmont, CA: Duxbury Press. [154]

Schilling, M. F., and Doi, J. A. (2014), "A Coverage Probability Approach to Finding an Optimal Binomial Confidence Procedure," *The American Statistician*, 68, 133–145. [152]

Skala, M. (2013), "Hypergeometric Tail Inequalities: Ending the Insanity," arXiv preprint arXiv:1311.5939. [151]

Sterne, T. E. (1954), "Some Remarks on Confidence or Fiducial Limits," *Biometrika*, 41, 275–278. [151,152]

U.S. Environmental Protection Agency (2012). "The National Ambient Air Quality Standards for Particle Pollution Revised Air Quality Standards For Particle Pollution And Updates To The Air Quality Index (AQI)," Available at *https://www.epa.gov/sites/production/files/2016-04/documents/2012_aqi_factsheet.pdf*. [157]

Wang, W. (2015), "Exact Optimal Confidence Intervals for Hypergeometric Parameters," *Journal of the American Statistical Association*, 110, 1491–1499. [151,152,154,156,158]

Wittes, J. T. (1972), "Note: On the Bias and Estimated Variance of Chapman's Two-Sample Capture-Recapture Population Estimate," *Biometrics*, 28, 592–597. [158]