Statistics and Computing (2024) 34:177
https://doi.org/10.1007/s11222-024-10485-7

ORIGINAL PAPER

# Optimal confidence interval for the difference between proportions

**Almog Peer**[1] · **David Azriel**[1]

Received: 8 January 2024 / Accepted: 19 August 2024


**Abstract**
Estimating the probability of the binomial distribution is a basic problem, which appears in almost all introductory statistics courses and is performed frequently in various studies. In some cases, the parameter of interest is a difference between two probabilities, and the current work studies the construction of confidence intervals for this parameter when the sample size is small. Our goal is to find the shortest confidence intervals under the constraint of coverage probability being at least as large as a predetermined level. For the two-sample case, there is no known algorithm that achieves this goal, but different heuristics procedures have been suggested, and the present work aims at finding optimal confidence intervals. In the one-sample case, there is a known algorithm that finds optimal confidence intervals presented by Blyth and Still (J Am Stat Assoc 78(381):108–116, 1983). It is based on solving small and local optimization problems and then using an inversion step to find the global optimum solution. We show that this approach fails in the two-sample case and therefore, in order to find optimal confidence intervals, one needs to solve a global optimization problem, rather than small and local ones, which is computationally much harder. We present and discuss the suitable global optimization problem. Using the Gurobi package we find near-optimal solutions when the sample sizes are smaller than 15, and we compare these solutions to some existing methods, both approximate and exact. We find that the improvement in terms of lengths with respect to the best competitor varies between 1.5 and 5% for different parameters of the problem. Therefore, we recommend the use of the new confidence intervals when both sample sizes are smaller than 15. Tables of the confidence intervals are given in the Excel file in this link (https://technionmail-my.sharepoint.com/:f:/g/personal/ap_campus_technion_ac_il/El-213Kms51BhQxR8MmQJCYBDfIsvtrK9mQIey1sZnZWIQ?e=hxGunl).

**Keywords** Binomial distribution · Exact confidence intervals · Optimization

## 1 Introduction

The task of constructing confidence intervals for the proportion of the binomial distribution is a basic problem in statistics, which appears in almost all introductory statistics courses and is performed frequently in many studies. In some cases, the parameter of interest is the difference between two proportions, and the present work studies the construction of confidence intervals for this parameter. Specifically, if $p_1$ and $p_2$ are two proportions, the parameter of interest is $\Delta = p_1 - p_2$. Other functions, such as the ratio $p_1/p_2$ or the log odds ratio $\log\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right)$ will not be discussed here, but we believe that our methodology can be extended to these functions. For a discussion about comparisons among different functions of $p_1$, $p_2$ see Brumback and Berg (2008). Our aim in this work is to study confidence intervals with minimal length. Even when the sample size is small, this is not a trivial problem, and we show below how computational difficulties can be overcome and present (almost) optimal confidence intervals for the stated problem.

First, one needs to distinguish between an exact confidence interval (henceforth, CI) and an approximate CI. An exact CI has a guarantee that the coverage probability is above some predetermined level of $1-\alpha$ for all the parameter space, while an approximate CI achieves this level only asymptotically, and might have a smaller coverage probability for some values of the parameter. An exact CI has the advantage of guaranteeing the desired level for every sample size and for every value of the parameter. However, it might come at the cost of wider intervals. This work focuses on exact CI and small sample sizes.

✉ David Azriel
davidazr@technion.ac.il

[1] Technion - Israel Institute of Technology, Haifa, Israel

Published online: 02 September 2024

We now review some widely-used methods for the one-sample case. The most popular one is the Wald CI, which is based on the normal approximation of the binomial variable. Specifically, let $X \sim \text{Binomial}(n, p_1)$, and let $\hat{p}_1 = X/n$. Wald CI is $\hat{p}_1 \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n}}$, where $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution. The Wald CI is symmetric around the observed proportion $\hat{p}_1 = X/n$, and its width depends on the variance estimator and the level of confidence $1-\alpha$. Among the approximate CIs, the Wilson score (Wilson 1927) gained some popularity. Similar to the Wald CI, the Wilson CI is based on the normal approximation, but with a different variance estimator. Agresti and Coull (1998) showed that the performance of the Wald CI is much inferior to the Wilson CI in terms of coverage probability. Agresti and Coull also suggest another CI, which they call an adjusted Wald CI. The idea is to simply take $X^* = X + 2$, $n^* = n + 4$ and compute the Wald CI with $X^*$ and $n^*$.

Brown et al. (2001) provided a comprehensive review of different methods to construct CIs. They compared performance in terms of minimum coverage level, average coverage level, and average diversion from $1-\alpha$. Based on the above criteria, they recommended the Wilson score CI or the Jeffreys CI for $n < 40$. The Jeffreys CI is obtained by using a prior $BETA(\frac{1}{2}, \frac{1}{2})$, known as the Jeffreys prior, and taking the middle $1-\alpha$ area under the posterior distribution. For $n \geq 40$, Brown et al. suggested using either the Wilson or the Jeffreys CIs or the Agresti Coull method that was mentioned above.

The first exact CI for the one-sample case was suggested by Clopper and Pearson (1934), and it is the intersection of two one-sided CIs. The Clopper and Pearson CI is generally too conservative—the intervals are fairly wide. Correspondingly, the coverage probability is higher than the desired level, especially for small $n$. Sterne (1954) developed an exact CI that is shorter than the Clopper and Pearson CI, and is optimal in the sense of having a minimal length, i.e., the sum of $n+1$ confidence regions is minimal among all CIs with the correct coverage probability. However, Crow (1956) showed that the Sterne method might lead to confidence regions that are the union of intervals and not a single interval. Crow further modified the Sterne method to return only confidence regions consisting of one interval for any $x$, preserving the above optimality property for CIs. Blyth and Still (1983) proposed an algorithm that finds all optimal CIs that are intervals, including the Crow CI. In Sect. 3.1 the Blyth and Still algorithm is described in detail, as we wish to generalize it to the two-sample case. Blaker (2000) show that the Blyth and Still CI does not maintain nestedness. This means that for two given coefficients $1-\alpha > 1-\alpha'$, the corresponding CI of coefficient $1-\alpha'$ is not necessarily contained in the other CI. Therefore, Blaker proposed an algorithm for constructing a CI that is always an interval and maintains nestedness but is not necessarily optimal.

Now, we will review several CIs for the two-sample case, i.e., for $\Delta = p_1 - p_2$. The Wald CI can be easily generalized based on the normal approximation of the differences of the averages. Specifically, let $X \sim \text{Binomial}(n, p_1)$, $Y \sim \text{Binomial}(m, p_2)$, where $X$ and $Y$ are independent and let $\hat{p}_1 = X/n$, $\hat{p}_2 = Y/m$. The Wald CI for $\Delta$ is

$$\hat{p}_1 - \hat{p}_2 \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}.$$

Miettinen and Nurminen (1985) demonstrated the poor coverage of this CI in a few examples and suggested relying on more stabilized estimators of the variance, which are based on quantiles of the chi-square distribution, and result in an approximate CI. Recently, Martín Andrés et al. (2024) revisited the work of Miettinen and Nurminen (1985) and suggested a bias correction factor in the context of hypothesis testing.

Newcombe (1998) reviewed 11 methods for creating CIs for $\Delta$, including the methods that were mentioned above. Newcombe compared the methods by the average coverage, the minimal coverage, and the percentage of non-coverage. Furthermore, Newcombe suggested a methodof his own called 'hybrid score' and it performed well in the above criteria; see Sect. 5.1 for more details. Another recommended method for constructing a CI for $\Delta$ was proposed by Agresti and Caffo (2000). Generalizing Agresti and Coull CI, they proposed to add four pseudo observations, one to each group, i.e., define $X^* = X+1$, $Y^* = Y+1$, $n^* = n+2$, $m^* = m+2$ and then calculate the Wald CI for the difference.

A few exact CIs for $\Delta$ were also developed. Santner and Snell (1980) proposed three different methods to construct exact CIs. One of them, called the tail method, has gained popularity due to its simplicity and ease of calculation. This method can be thought of as a two-dimensional analog of the Clopper and Pearson CI for one proportion, where the CIs are an intersection of two one-sided intervals. This method typically leads to too conservative intervals, as shown by Chan and Zhang (1999). The latter paper suggests a different method for constructing exact CIs. Agresti and Min (2001) studied exact CIs for the two-sample case. They reviewed the Chan and Zhang CI and suggested a modification that results in significant improvement in performance. The method is described in detail in Sect. 5.1. Fagerland et al. (2015) compared several methods including the ones mentioned above and also others both approximate and exact. Their main criterion for comparison was the closeness to the nominal level $1-\alpha$. They recommended using the Agresti and Min CI. Fay and Hunsberger (2021) reviewed different methods and compared them both in terms of testing and confidence intervals. They found that no one method can meet all the

desirable properties and provide recommendations based on which properties are given more importance. A related topic is exact tests in $2 \times 2$ contingency tables; see Keer (2023) and references therein. This work uses optimization methods to maximize the number of outcomes in a rejection region similarily to what is done in the current paper.

To sum up, for the one-sample case there exists an algorithm that minimizes the sum of interval's length of the CI under the constraint of obtaining a certain coverage level. For the two-sample case, such an algorithm does not exist, but rather different heuristics were suggested. This work aims at filling this gap, namely,to construct an algorithm that computes the optimal CI for small sample sizes in the two-sample case and to compare it to existing methods.

The rest of the work is organized as follows: in Sect. 2 the optimization problem is stated and the basic notation is introduced. The algorithm suggested in Blyth and Still (1983) finds the optimal solutions for the one-sample case. It is based on solving small and local optimization problems and then using an inversion step to find the global optimum solution. Section 3 presents the algorithm and discusses extensions to the two-sample case. It is shown that this approach fails in the two-sample case and therefore, in order to findan optimal CI, one needs to solve a global optimization problem, rather than small and local ones, which is computationally much harder. The global optimization problem is presented and discussed in Sect. 4. Using the Gurobi Optimization, LLC (2023) package, we find near-optimal solutions when the sample sizes are smaller than 15, and we compare these solutions to some existing methods, both approximate and exact in Sect. 5. We find that the improvement in terms of lengths with respect to the best competitor varies between 1.5 and 5% for different parameters of the problem. Section 6 concludes with some recommendations and future research directions.

## 2 Problem statement

Recall that $X$ and $Y$ are independent, $X \sim \text{Binomial}(n, p_1)$ and $Y \sim \text{Binomial}(m, p_2)$. We aim at constructing CIs for $p_1$ (respectively, $\Delta := p_1 - p_2$) for the one- (respectively, two-) sample cases. In the one-sample case, we define $C_1$ to be the collection of all confidence intervals, i.e.,

$$C_1 := \{[l_x, u_x]\}_{x \in \{0,1,\ldots,n\}},$$

where $l_x, u_x$ is the lower and upper limit of the confidence interval when $X = x$ is observed. Correspondingly, for the two-sample case, we define

$$C_2 := \{[l_{x,y}, u_{x,y}]\}_{x \in \{0,1,\ldots,n\}, y \in \{0,1,\ldots,m\}},$$

and here $[l_{x,y}, u_{x,y}]$ is the confidence interval for $\Delta$ when $(X = x, Y = y)$ is observed.

We aim to find an optimal exact CI, where optimality is with respect to the sum of all interval lengths. In the one-sample case, the length is

$$Length(C_1) = \sum_{x=0}^{n} (u_x - l_x),$$

and in the two-sample case, it is

$$Length(C_2) = \sum_{y=0}^{m} \sum_{x=0}^{n} (u_{(x,y)} - l_{(x,y)}).$$

For computational reasons, we define a grid $D$ for $\Delta$ values, and a grid $P$ for $p_1, p_2$ single proportions values, e.g., $P = \{0, 0.01, 0.02 \ldots, 1\}$, $D = \{-1, -0.99, \ldots 0, 0.01, 0.02 \ldots, 1\}$). The grids choices are connected to each other since only $(p_1, p_2) \in P \times P$ such that $p_1 - p_2 \in D$ are active in the problem. From a statistical point of view, the finer is the grid the better; however, a finer grid comes at the cost of computational burden.

The optimization problem we aim to solve for the one-sample case is

$$\begin{aligned} &\min_{C_1} Length(C_1) \text{ subject to } P_{p_1}(p_1 \in [l_X, u_X]) \\ &\geq 1 - \alpha \ \ \forall p_1 \in P, \end{aligned} \tag{1}$$

where the sub-index $p_1$ means that the probability is under $X \sim p_1$ (and similar notation is used for the two-sample case). For the two-sample case, the optimization problem is

$$\begin{aligned} &\min_{C_2} Length(C_2) \\ &\quad \text{subject to } P_{p_1,p_2}(\Delta \in [l_{(X,Y)}, u_{(X,Y)}]) \\ &\geq 1 - \alpha \text{ for all } (p_1, p_2) \\ &\quad \in P \times P \text{ such that } p_1 - p_2 = \Delta \in D. \end{aligned} \tag{2}$$

## 3 Generalization of the Blyth and Still algorithm to the two-sample case

The Blyth and Still algorithm finds all the solutions to the problem (1). In Sect. 3.1 the algorithm is described in detail. Generalization of the algorithm to the two-sample case is discussed in Sect. 3.2. It is shown that the generalized algorithm provides confidence regions rather than intervals.

### 3.1 The Blyth and Still algorithm

We consider the one-sample case, that is, Problem (1), and describe the Blyth and Still algorithm. First, a few definitions are given.

**Definition 3.1**

- A subset $S_1 = \{r, r+1, \ldots, t\}$ where $0 \le r < t \le n$ is an acceptance region with respect to $p_1$ if $P_{p_1}(X \in S_1) \ge 1-\alpha$.
- A subset $S_1$ is a minimal acceptance region (henceforth MAR) with respect to $p_1$, denoted by $MAR(p_1)$, if there is no other acceptance region with respect to $p_1$ that has fewer elements.
- Let $S_1, \tilde{S}_1$ be two MARs with respect to $p_1$ and $\tilde{p}_1$, where $p_1 \le \tilde{p}_1$. We say that the pair $(S_1, \tilde{S}_1)$ maintains monotonicity if $\min\{S_1\} \le \min\{\tilde{S}_1\}$ and $\max\{S_1\} \le \max\{\tilde{S}_1\}$.

The algorithm can be described as follows:

The Blyth and Still algorithm

Input: $P = \{\rho_1, \rho_2, \ldots, \rho_{|P|}\}$ - a grid of values in $[0,1]$ such that $\rho_1 \le \rho_2 \le \cdots \le \rho_{|P|}$; $n$ - sample size; $1-\alpha$ - desired level.
Output: $C_1$ - a collection of $n+1$ confidence intervals.

1. Find all MARs. For all $p_1 \in P$ calculate all MARs.
2. Remove MARs that do not maintain monotonicity. For all $i = 1, 2, \ldots, |P|-1|$ and for all $S_1 = MAR(\rho_i)$: if for all $\tilde{S}_1$ that is a MAR of $\rho_{i+1}$ the pair $(S_1, \tilde{S}_1)$ does not maintain monotonicity, then remove $S_1$. Also, for all $i = 2, 3, \ldots, |P|$ and for all $\tilde{S}_1 = MAR(\rho_i)$, if for all $S_1$ that is a MAR of $\rho_{i-1}$ the pair $(S_1, \tilde{S}_1)$ does not maintain monotonicity, then remove $\tilde{S}_1$.
3. Choose linear ordering. For $i = 1$ choose $MAR^*(\rho_1)$ from all the MARs of $\rho_1$ that remained after the previous step. For $i = 2, 3, \ldots, |P|$, choose $MAR^*(\rho_i)$ from all the remaining MARs of $\rho_i$ such that $(MAR^*(\rho_{i-1}), MAR^*(\rho_i))$ maintains monotonicity.
4. Invert. For all $x = 0, 1, \ldots, n$, define $CR(x) := \{p_1 \in P : x \in MAR^*(p_1)\}$ and $l_x := \min\{CR(x)\}$, and $u_x := \max\{CR(x)\}$.
5. Return $C_1 = \{[l_x, u_x]\}_{x \in \{0,1,\ldots,n\}}$.

We now discuss every step of the algorithm in detail.

1. **Find all MARs**.

Finding all MARs with respect to $p_1$ can be done in the following manner: set $r = 0$ and find the smallest $t_0$ that makes the interval $[0, t_0]$ cover $p_1$ with probability of at least $1-\alpha$, i.e., $P_{p_1}(X \in [0, t_0]) \ge 1-\alpha$. Then, repeat this procedure for $r = 1, 2, \ldots, n$: for each $r$, find the smallest integer $t_r$ such that $P_{p_1}(X \in [r, t_r]) \ge 1-\alpha$. Notice that there exists a critical value $R$ such that for $r \ge R$ there is no $t_r$ that provides coverage of $p_1$ with the desired probability, that is, even if we set $t_r = n$, the interval $S_1 = [r, n]$ is not an acceptance region for $p_1$, i.e., $P_{p_1}(X \in [r, n]) < 1-\alpha$. After calculating $t_0, t_1, \ldots$, the lengths of $[0, t_0], [1, t_1], \ldots$ are compared and the intervals with minimal length are chosen. Thus, for each $p_1 \in P$ there are $O(n^2)$ calculations, and the total number of calculations in this step is $|P|O(n^2)$.

2. **Remove solutions that do not maintain monotonicity**.

This step is needed to ensure that $CR(x)$ in the invert step (# 4) would be an interval rather than a confidence set. As mentioned in the introduction, the Sterne CI can lead to optimal confidence sets, which are optimal in terms of length, but they are not necessarily intervals. For a concrete example, suppose that for $p_1 = 0.1$ the only MAR is $MAR(0.1) = [1, 7]$ and for $p_1 = 0.11$ the MARs are $MAR(0.11) = [0, 7], [1, 8], [2, 9]$. Then, the first MAR $[0, 7]$ is removed as it violates the monotonicity assumption with respect to the MAR $[1, 7]$ of $p_1 = 0.1$. If for $p_1 = 0.1$ there was more than one MAR, $[0.7]$ is removed only if it violates the monotonicity assumption for any MAR of $p_1 = 0.1$.

3. **Choose linear ordering**.

There are different ways to choose a linear ordering that will lead to different CIs. However, all of them will be optimal in the sense of the optimization problem in (1). Blyth and Still explored a few options for choosing MARs that have other desired properties. For example, if one wants to avoid CIs where $l_x = l_{x+1}$ for some $x$'s, then certain linear orderings should be avoided.

4. **Invert**.

By the monotonicity property, the set $CR(x)$ is an interval, i.e., there are no holes in $CR(x)$. By the construction of $CR(x)$ we have that $\sum_{x=0}^{n} \#\{CR(x)\} = \sum_{p_1 \in P} \#\{MAR^*(p_1)\}$, where $\#A$ is the number of elements in set $A$. Since the number of elements in each $MAR^*(p_1)$ is minimal, so is $\sum_{x=0}^{n} \#\{CR(x)\}$. Minimizing $\sum_{x=0}^{n} \#\{CR(x)\}$ is equivalent to Problem (1) and hence the output of the algorithm is a solution to Problem (1). Moreover, by choosing different linear orderings in Step 3, all the optimal solutions can be found by this algorithm.

### 3.2 A generalization of the Blyth and Still algorithm to the two-sample case

In this section, we consider a generalization of the Blyth and Still algorithm that aims to address Problem (2). While the minimal length and the desired coverage probability are still preserved, we will show that the output of this generalized algorithm is not necessarily a confidence interval, but rather

a confidence set. We start with a definition that parallels Definition 3.1.

**Definition 3.2**

- A subset $S_2 \subseteq \{0, 1, \ldots, n\} \times \{0, 1, \ldots, m\}$ is an acceptance region with respect to $\Delta \in D$ if for all $(p_1, p_2) \in P \times P$ such that $p_1 - p_2 = \Delta$ we have that $P_{p_1,p_2}((X, Y) \in S_2) \geq 1 - \alpha$.
- A subset $S_2$ is a minimal acceptance region (henceforth MAR) with respect to $\Delta \in D$, denoted by $MAR(\Delta)$, if there is no other acceptance region with respect to $\Delta$ that has fewer elements.

Notice that here we define an acceptance region to be a subset of $\{0, 1, \ldots, n\} \times \{0, 1, \ldots, m\}$, without requiring that there are no holes (e.g., $S_2$ in which $(0, 2), (0, 4) \in S_2$, $(0, 3) \notin S_2$ is a possible acceptance region) as in the one sample definition of an acceptance region.The motivation is to allow for flexibility in the set of all possible MARs with the hope that a certain choice of MARs will lead to a confidence interval in the inversion step. However, later we demonstrate that there are cases in which all possible choices of MARs lead to confidence regions that are not intervals.

The generalized Blyth and Still algorithm

Input: $P$ - a grid of values in $[0, 1]$; $D$ - a grid of values in $[-1, 1]$; $n, m$ - sample sizes; $1-\alpha$ - desired level.
Output: $\tilde{C}_2$ - a collection of $(n+1)(m+1)$ confidence sets.

1. Find one MAR for each $\Delta \in D$. For all $\Delta \in D$ find one MAR, denoted by $MAR(\Delta)$.
2. Invert. For all $(x, y) \in \{0, 1, \ldots, n\} \times \{0, 1, \ldots, m\}$, define $CR(x, y) := \{\Delta \in D \text{ if } (x, y) \in MAR(\Delta)\}$.
3. Return $\tilde{C}_2 = \{CR(x, y)\}_{x \in \{0,1,\ldots,n\}, y \in \{0,1,\ldots,m\}}$.

Notice that in this algorithm the steps of removing MARs that do not maintain monotonicity and choosing linear ordering are not present. This will be explained below, but first, we describe how to find the MARs in Step 1.

Finding the MARs in the two-sample case is more complicated than the one-sample equivalent task because one needs to ensure $1-\alpha$ coverage for all $(p_1, p_2) \in P \times P$ that satisfy $p_1 - p_2 = \Delta$ and not for just one specific $p_1$. Also, in the one-sample case, the MARs are intervals but here the MARs are general sets. We found no simple algorithm to compute the MARs in the two-sample setting and this step is performed by solving Optimization Problem 1, which is given below. This optimization problem consists of $(n+1)(m+1)$ binary variables and has at most $|P|$ constraints for maintaining the coverage probability. In some instances, the solution is not unique. We show below that, unlike the one-sample case, here there is no way of choosing a solution that satisfies that $CR(x, y)$ in the invert step is always an interval. Therefore, we selected one solution arbitrarily. The optimal solution was computed by a procedure in the R software (R Core Team 2021) that uses the Gurobi package (Gurobi Optimization, LLC 2023).

**Optimization Problem 1** ***Problem parameters****: $D$ - a grid of values in $[-1, 1]$ for $\Delta$; $P$ - a grid of values in $[0, 1]$ for $p_1$ and $p_2$; $(n, m)$ - number of trials from each sample; confidence coefficient $1 - \alpha$.*
***Decision variables:*** *$r(x, y)$ - a binary variable that equals 1 iff $(x, y)$ belongs to the MAR.*
***Objective function:*** *Minimize $\sum_{y=0}^{m} \sum_{x=0}^{n} r(x, y)$.*
***Constraints:***

*(a) Maintain the coverage of $\Delta$:*

$$\sum_{y=0}^{m} \sum_{x=0}^{n} r(x, y) \binom{n}{x} \binom{m}{y} {p_1}^x (1 - p_1)^{n-x} {p_2}^y (1 - p_2)^{m-y} \geq 1 - \alpha, \quad (3)$$

*for all $(p_1, p_2) \in P \times P$ such that $p_1 - p_2 = \Delta$.*

*(b) The decision variables $r(x, y)$ are binary:*

$$r(x, y) \in \{0, 1\} \text{ for all } (x, y) \in \{0, 1, \ldots, n\} \times \{0, 1, \ldots, m\}.$$

Furthermore, we used the above program to find all possible solutions for the Optimization Problem 1. This allows us to show that there are examples in which no ordering of MARs will lead to confidence intervals as in the one-dimensional case. For example, when $n = 5, m = 5, \alpha = 0.1$, $P = \{0, 0.0001, 0.0002, \ldots, 1\}$ we find that:

(a) For $\Delta = -0.4$ the only MAR contains $(x, y) = (0, 5)$.
(b) For $\Delta = -0.37$ there are five MARs, all of them contain $(x, y) = (0, 5)$.
(c) For $\Delta = -0.38$ the only MAR does not contain $(x, y) = (0, 5)$.

This means that for any choice of MARs in this setting, if $(x, y) = (0, 5)$ is observed, the confidence set of the invert step will contain $\Delta = -0.4, -0.37$ but not $\Delta = -0.38$. That is, the optimal confidence set will be composed of at least two disjoint intervals. Furthermore, by examining the constraint (3) for continuous $p_1, p_2$ using analytical graphical tools, we observe that this phenomenon still occurs even for a finer grid. The full list of MARs for this example are presented in the appendix.

We found that this phenomenon occurs quite often: from six pairs of sample sizes $(n, m) \in \{(10, 5), (5, 5), (6, 4), (9, 6), (7, 7)\}$ and $\alpha = 0.05$, only $(10, 5)$ and $(6, 4)$ do not have MARs with this deficiency.

It follows that one cannot achieve CIs with minimal length using the Blyth and Still method. Rather, this method guarantees confidence sets (not necessarily intervals) that have a minimal number of elements in $D$ and have the desired coverage level $1 - \alpha$.

In Sect. 5 we examine the performance of this method where gaps in the confidence sets are simply filled in order to achieve a confidence interval.

## 4 Performing full optimization

In the previous section we showed that the generalized Blyth and Still algorithm to the two-sample case leads to confidence regions that are optimal in their size, but can be composed of several disjoint intervals, instead of one interval. The solution of filling the gaps between the disjoint intervals is later examined.

Therefore, a different optimization method should be considered in order to solve Problem (2). The aim is to find a set of confidence regions that are optimal in length, have the right coverage level, and are constrained to be intervals. This can be done by solving the following optimization problem. For some instances, the solution is not unique, and when this is the case, we select an arbitrary optimal solution.

**Optimization Problem 2** ***Problem parameters****: $D$ - a grid of values in $[-1, 1]$ for $\Delta$; $P$ - a grid of values in $[0, 1]$ for $p_1$ and $p_2$; $(n, m)$ - number of trials from each sample; confidence coefficient $1 - \alpha$.*
***Decision variables:*** *$l_{(x,y)}, u_{(x,y)}$ - the lower and upper limits for when $(x, y)$ is observed; $r(x, y, \Delta)$ - a binary variable that equals 1 iff the CI includes $\Delta$ when $(x, y)$ is observed.*
***Objective function:*** *Minimize $\sum_{y=0}^{m} \sum_{x=0}^{n} (u_{(x,y)} - l_{(x,y)})$.*
***Constraints:***

*(a) Maintain the coverage of $\Delta$:*

$$\sum_{y=0}^{m} \sum_{x=0}^{n} r(x, y, \Delta) \binom{n}{x} \binom{m}{y} {p_1}^x (1 - p_1)^{n-x} {p_2}^y (1 - p_2)^{m-y} \geq 1 - \alpha, \tag{4}$$

*for all $(p_1, p_2) \in P \times P$ such that $p_1 - p_2 = \Delta$.*

*(b) Connecting the variables $r(x, y, \Delta)$ and $l_{(x,y)}$ and $u_{(x,y)}$:*

$$r(x, y, \Delta) \leq \frac{\Delta - l_{(x,y)}}{2} + 1 \text{ and } r(x, y, \Delta) \leq \frac{u_{(x,y)} - \Delta}{2} + 1 \tag{5}$$

*for all $(x, y, \Delta) \in \{0, 1, \ldots, n\} \times \{0, 1, \ldots, m\} \times D$.*

*(c) Connecting further the variables $r(x, y, \Delta)$ and $l_{(x,y)}$ and $u_{(x,y)}$:*

$$\frac{u_{(x,y)} - l_{(x,y)}}{d_{max}} + 1 \leq \sum_{\Delta \in D} r(x, y, \Delta) \leq \frac{u_{(x,y)} - l_{(x,y)}}{d_{min}} + 1 \tag{6}$$

*for all $(x, y) \in \{0, 1, \ldots, n\} \times \{0, 1, \ldots, m\}$, where $d_{min}$ and $d_{max}$ are the minimal and maximal distances between successive elements in the sorted grid $D$.*

*(d) The variables $r(x, y, \Delta)$ are binary:*

$$r(x, y, \Delta) \in \{0, 1\} \text{ for all } (x, y, \Delta) \in \{0, 1, \ldots, n\} \times \{0, 1, \ldots, m\} \times D.$$

*(e) Interval limits are between $[-1, 1]$:*

$$-1 \leq l_{(x,y)} \leq 1 \text{ and } -1 \leq u_{(x,y)} \leq 1 \text{ for all } (x, y) \in \{0, 1, \ldots, n\} \times \{0, 1, \ldots, m\}.$$

A solution to Optimization Problem 2 finds the shortest CI that has $1 - \alpha$ coverage for every $\Delta \in D$, i.e., it solves Problem (2). The optimization problem consists of $2(n + 1)(m + 1)$ variables that assume values in $D$, and $|D|(n + 1)(m + 1)$ binary variables.

The constraint in (5) consists of two conditions, which force $r(x, y, \Delta)$ to be 0 if $\Delta < l(x, y)$ or $\Delta > u(x, y)$. This is because

$$\Delta < l(x, y) \iff \frac{\Delta - l_{(x,y)}}{2} < 0 \iff \frac{\Delta - l_{(x,y)}}{2} + 1 < 1.$$

Thus, by condition (5), if $\Delta < l(x, y)$ then $r(x, y, \Delta) < 1$. In addition, the expression $\frac{(\Delta - l_{(x,y)})}{2} + 1$ is non-negative, and hence $r(x, y, \Delta) < c$ for $c \geq 0$, which implies that $r(x, y, \Delta) = 0$, as it is a binary variable. Similarly, if $\Delta > u(x, y)$, then $r(x, y, \Delta) = 0$. If neither $\Delta < l(x, y)$ nor $\Delta > u(x, y)$ are satisfied, then Constraint (5) does not restrict $r(x, y, \Delta)$ to a certain value. This is where Constraint (6) comes into play. In the case where the grid $D$ is equally-spaced, Constraint (6) simplifies to

$$\sum_{\Delta \in D} r(x, y, \Delta) = \frac{u_{(x,y)} - l_{(x,y)}}{d} + 1, \tag{7}$$

where $d$ is the constant difference between successive elements in the sorted grid $D$. In this case, Eq. (7) implies that if $l_{(x,y)} \leq \Delta \leq u_{(x,y)}$, then $r(x, y, \Delta) = 1$. Combining this with (5), we have that the $r$ variables are fully determined by the $l$ and $u$ variables. Constraint (6) does not change the optimal value, but rather drastically decreases the number of

feasible solutions and thus reduces the number of computations needed to solve Optimization Problem 2.

Another way of forcing $r(x, y, \Delta)$ to be 1 if $l_{(x,y)} \le \Delta \le u_{(x,y)}$, even when $D$ is not equally-spaced, is to change the objective function to

$$\text{minimize} \sum_{y=0}^{m}\sum_{x=0}^{n}(u_{(x,y)} - l_{(x,y)}) - \frac{d_{min}}{2N}\sum_{x=0}^{n}\sum_{y=0}^{m} \sum_{\Delta\in D} r(x, y, \Delta),$$

where $N = (n+1)(m+1)|D|$ is the number of $r$ variables and $d_{min}$ is the minimal distance between consecutive elements in the sorted grid $D$.

If one wishes to find a solution that maintains the symmetry of the binomial distribution under the transformation $p \mapsto 1 - p$, then one can add the restriction

$$u_{(x,y)} = -l_{(n-x,m-y)} \text{ for all } (x, y) \in \{0, 1, \ldots, n\} \times \{0, 1, \ldots, m\}. \quad (8)$$

In the Generalized Blyth and Still algorithm that was given in Sect. 3.2, Optimization problem 1 is being solved $|D|$ times, each with $(n + 1)(m + 1)$ binary variables. Here, on the other hand, there are $|D|(n + 1)(m + 1)$ binary variables and the optimization problem is solved only once. Since the running time of the optimization problem solver is not linear in the number of the binary variables, Optimization Problem 2 is computationally much more difficult.

## 5 Comparisons

In this section, we compare the full optimization algorithm of Sect. 4 and the generalized Blyth and Still algorithm of Sect. 3.1 to several existing methods, both approximate and exact.

### 5.1 A list of methods

The existing methods we have compared are listed below.
1. The Wald CI, i.e.,

$$\hat{p}_1 - \hat{p}_2 \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}.$$

It is included in our comparison due to its widespread use even though it is known to perform poorly.
2. The adjusted Wald CI of Agresti and Caffo (2000) (AC) is given by

$$\bar{p}_1 - \bar{p}_2 \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n} + \frac{\bar{p}_2(1-\bar{p}_2)}{m}}.,$$

where $\bar{p}_1 = (x + 1)/(n + 2)$, $\bar{p}_2 = (y + 1)/(m + 2)$
3. The hybrid score (HS) of Newcombe (1998).

Newcombe hybrid score (HS)

Input: $n, m$ - sample sizes; $1 - \alpha$ - confidence coefficient.
Output: $C_2$ - a collection of $(n+1)(m+1)$ confidence intervals.

1. Calculate lower and upper bounds. Let $\hat{p}_1 = x/n$ and $\hat{p}_2 = y/n$. For each $x \in \{0, 1, \ldots, n\}$, let $l_x(1), u_x(1)$ be the two solutions for $p_1$ of $z_{1-\frac{\alpha}{2}} = \frac{|\hat{p}_1 - p_1|}{\sqrt{\frac{p_1(1-p_1)}{n}}}$, and for each $y \in \{0, 1, \ldots, m\}$, let $l_y(2), u_y(2)$ be the two solutions for $p_2$ of $z_{1-\frac{\alpha}{2}} = \frac{|\hat{p}_2 - p_2|}{\sqrt{\frac{p_2(1-p_2)}{m}}}$.
2. Hybrid score. For all$(x, y) \in \{0, 1, \ldots, n\} \times \{0, 1, \ldots, m\}$ define

$$l(x, y) = \hat{p}_1 - \hat{p}_2 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{l_x(1)(1-l_x(1))}{n} + \frac{u_y(2)(1-u_y(2))}{m}}$$

and

$$u(x, y) = \hat{p}_1 - \hat{p}_2 + z_{1-\frac{\alpha}{2}}\sqrt{\frac{u_x(1)(1-u_x(1))}{n} + \frac{l_y(2)(1-l_y(2))}{m}}.$$

3. Return $C_2 = \{[l_{(x,y)}, u_{(x,y)}]\}_{(x,y)\in\{0,1,\ldots,n\}\times\{0,1,\ldots,m\}}$.

The calculations of the HS CI can be found in the R software in the package 'DescTools' (Signorell 2024).
4. The exact method of Agresti and Min (2001) (AM)

The exact method of Agresti and Min (2001) (AM)

Input: $P$ - a grid of values in $[0, 1]$; $D$ - a grid of values in $[-1, 1]$; $n, m$ - sample sizes; $1-\alpha$ - confidence coefficient.Output: $C_2$ - a collection of $(n+1)(m+1)$ confidence intervals.

1. Calculate scores. For any triplet $(x, y, \Delta) \in \{0, 1, \ldots, n\} \times \{0, 1, \ldots, m\} \times D$ define

$$Z(x, y, \Delta) = \frac{\left(\frac{x}{n} - \frac{y}{m} - \Delta\right)^2}{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{m}},$$

where $\tilde{p}_1, \tilde{p}_2$ are the MLE for $p_1, p_2$ under $p_1 - p_2 = \Delta$, i.e., they maximize the likelihood ${p_1}^x(1-p_1)^{n-x}{p_2}^y(1-p_2)^{m-y}$, under the constraint $p_1 - p_2 = \Delta$.

2. Calculate $\lambda$ values. For any triplet $(x, y, \Delta) \in \{0, 1, \ldots, n\} \times \{0, 1, \ldots, m\} \times D$ define

$$\lambda(x, y, \Delta) = \max\left\{P_{p_1,p_2}\Big(Z(X, Y, \Delta) \geq Z(x, y, \Delta)\Big) : (p_1, p_2) \in P \times P \text{ s.t } p_1 - p_2 = \Delta\right\}.$$

3. Invert For all $(x, y) \in \{0, 1, \ldots, n\} \times \{0, 1, \ldots, m\}$ define $CR(x, y) := \{\Delta \in D \text{ if } \lambda(x, y, \Delta) > \alpha\}$ and $l_{(x,y)} := \min\{CR(x, y)\}, u_{(x,y)} := \max\{CR(x, y)\}$.
4. Return $C_2 = \{[l_{(x,y)}, u_{(x,y)}]\}_{(x,y)\in\{0,1,\ldots,n\}\times\{0,1,\ldots,m\}}$

Notice that similar to the generalized Blyth and Still algorithm, $CR(x, y)$ is not necessarily an interval. Therefore, the confidence interval is defined by the minimum and maximum value of $CR(x, y)$.

We could not find a code in R that implements the AM algorithm, and therefore we wrote our own code. For calculating the MLEs $\tilde{p}_1, \tilde{p}_2$ in Step 1 we used the function 'z2stat' in the package 'PropCIs' (Scherer 2022); an explicit expression for the MSE is given in Miettinen and Nurminen (1985).

We ran the AM algorithm under two modes, which we denote by AM1 and AM2. The first mode is with the grids $D = \{-1, -0.99, -0.98, \ldots, 1\}$ and $P = \{0, 0.01, 0.02, \ldots, 1\}$, and the second mode is with the grids $D = \{-1, -0.999, -0.998, \ldots, 1\}$ and $P = \{0, 0.001, 0.002, \ldots, 1\}$. The reason for considering the coarser grid of the first mode is to attain a better comparison to the full optimization method, in which the finer grid is computationally infeasible. The AM algorithm is sub-optimal but runs much faster than full optimization and therefore can be computed with a finer grid.

5. The generalized Blyth and Still algorithm that is given in Sect. 3.2 (BSG).

We ran the algorithm where the confidence sets are filled if they are not intervals. As in the AM method, we considered two possible modes, denoted by BSG1 and BSG2. In the first mode we used the grids $D = \{-1, -0.99, -0.98, \ldots, 1\}$ and $P = \{0, 0.02, 0.04, \ldots, 1\}$. In the second mode we used the grid $D = \{-1, -0.999, -0.998, \ldots, 1\}$ and a different grid $P$ for every $\Delta \in D$, a choice that improves the performance of the algorithm. Namely, for $\Delta \geq 0$ we define

$$P_\Delta = \{\Delta, \ldots, 1\} \text{ with equal jumps of } \frac{1-\Delta}{100}$$

and for $\Delta < 0$ we define

$$P_\Delta = \{0, \ldots, 1+\Delta\} \text{ with equal jumps of } \frac{1-\Delta}{100}.$$

The coverage condition of the algorithm in (3) is satisfied for any pair $(p_1, p_1 - \Delta)$ where $p_1 \in P_\Delta$.

6. The full optimization algorithm presented in Sect. 4 (Full).

We ran the algorithm of Sect. 4 with the grid $D = \{-1, -0.99, -0.98, \ldots, 1\}$, $P = \{0, 0.01, 0.02, \ldots, 1\}$ and denote it by Full1. Here we only considered the coarse grid since the computational complexity of the optimization problem is much greater. We ran the problem with the symmetric condition (8) and found that this restriction does not change the length of the CIs in the optimal solution.

The Gurobi software was given a time limit of two minutes. If the time limit is reached, the best solution is reported, as is the gap between this solution to the current lower bound in terms of percentage. The starting point of the algorithm is based on the output of the AM method.

Since the grid is relatively coarse, there are non-negligible amount of differences $p_1 - p_2$ for which $1-\alpha$ coverage is not preserved. We examined two ways to overcome this problem, where the updated limits are denoted by $l^*_{(x,y)}$ and $u^*_{(x,y)}$.

(a) Extending the CIs in each direction by adding or reducing 0.01 (which is the gap size in the grid we used) (Full2), i.e.,

$$l^*_{(x,y)} = l_{(x,y)} - 0.01 \text{ and } u^*_{(x,y)} = u_{(x,y)} + 0.01.$$

(b) Extending the CIs in each direction by adding or reducing 0.01/2 (Full3), i.e.,

$$l^*_{(x,y)} = l_{(x,y)} - 0.01/2 \text{ and } u^*_{(x,y)} = u_{(x,y)} + 0.01/2.$$

In these extensions, the new limits are truncated if they exceed the interval $[-1, 1]$.

### 5.2 Criteria of performance

We compare the methods listed in Sect. 5.1 according to the following six criteria.

*AVG length.* The average length in defined by

$$\frac{\sum_{y=0}^{m}\sum_{x=0}^{n}(u_{(x,y)}-l_{(x,y)})}{(n+1)(m+1)}.$$

*PCT of under-coverage.* Define the coverage probability function $CP(p_1, p_2) := P_{p_1,p_2}(\Delta \in [l_{(X,Y)}, u_{(X,Y)}])$. The percentage of under-coverage is

$$100 \times \int_0^1 \int_0^1 I(CP(p_1, p_2) < 1-\alpha) dp_1 dp_2.$$

*PCT of substantial under-coverage.* This is defined by

$$100 \times \int_0^1 \int_0^1 I(CP(p_1, p_2) < 1-\alpha-0.01) dp_1 dp_2.$$

*AVG deviation.* This is defined by

$$10,000 \times \int_0^1 \int_0^1 [1-\alpha-CP(p_1, p_2)] \, I(CP(p_1, p_2) < 1-\alpha) dp_1 dp_2.$$

This expression is the loss for an average pair $(p_1, p_2)$ (assuming a uniform distribution), where the loss for each pair is defined by the difference between the desired level $1-\alpha$ and the actual coverage level $CP(p_1, p_2)$ when $CP(p_1, p_2)$ is below $1-\alpha$ and zero otherwise. The factor 10,000 is used since this loss is relatively small in most of the methods we used.

Min CL. The minimum coverage probability is defined by $\min_{(p_1,p_2)\in[0,1]\times[0,1]} CP(p_1, p_2)$.

AVG CL. The average coverage probability is $\int_0^1 \int_0^1 CP(p_1, p_2) dp_1 dp_2$.

For calculating the above criteria (besides AVG length), we sampled 40,000 pairs $(p_1, p_2)$ from a uniform distribution on $[0, 1] \times [0, 1]$. This defines a grid $\mathcal{P}$ in $[0, 1] \times [0, 1]$. Then the above criteria are computed using this grid. For example, the percentage of under-coverage is evaluated by

$$100 \times \frac{1}{|\mathcal{P}|} \sum_{(p_1,p_2)\in\mathcal{P}} I(CP(p_1, p_2) < 1-\alpha).$$

### 5.3 Results

We calculated the resulting CIs of the methods listed in Sect. 5.1 for three cases of $(n, m)$, namely $(n, m) \in \{(9, 6), (14, 7), (10, 10)\}$. For each of them, three different confidence coefficients are considered, $\alpha \in \{0.01, 0.05, 0.1\}$. For each set of parameters and a CI method, we computed the six criteria of Sect. 5.2.

The results for $(n, m) = (9, 6), (14, 7), (10, 10)$ are given in Tables 1, 2, and 3, respectively. A few observations and conclusions are now given.

- The WALD CI performs poorly. Almost for all pairs, the coverage probability is below the desired level $1-\alpha$ and even below $1-\alpha-0.01$. Also, the average coverage is well below the desired level. This finding is not surprising as the WALD CI relies on asymptotic approximation, which is not valid for small sample sizes.
- We considered three non-exact methods: WALD, HS and AC. Comparing these methods in terms of average length, the order is usually WALD < HS < AC, but the same order holds under the under-coverage and substantial under-coverage criteria. This means that narrower CIs come with the price of under-coverage.
- The Full1 method produces CI with optimal length, or close to optimal; see the discussion below. As we expected, it has the shortest average length among all exact CIs. Compared to the approximate CIs it is longer by 2–10% than HS and WALD and it has a similar length as AC.
- The Full1 method does not guarantee exact coverage for any $(p_1, p_2)$, just for the pairs in the grid. For $(n, m) = (9, 6)$, the percentage of under-coverage pairs ranges from 6% for $\alpha = 0.01$, to 10% for $\alpha = 0.1$. For the two other sample sizes, it ranges from 14 to 18%. Examining Full1 by the criterion of percentage of substantial under-coverage, we can see that it has good performance, especially for small $\alpha$. Yet, for $(n, m, \alpha) = (10, 10, 0.1)$ the percentage of substantial under-coverage reaches 8%, which might be too high. Still, the Full1 has a smaller percentage of under-coverage and percentage of substantial under-coverage compared to the approximate CIs, including AC.
- The exact methods Full1, BSG1 and AM1 ran with the same grid for $\Delta$. Among these methods, the order of the average length is usually Full1 < BSG1 < AM1. The length improvement of Full1 compared to AM1 is about 2–5%. On the other hand, AM1 has better coverage than BSG1 and Full1.
- The modification of Full2 produces a CI that is exact for all pairs but it comes at the cost of a larger length of 0.02. Full3 does not guarantee exact coverage, but the percentage of under-coverage is decreased by about 90% compared to Full1, and the intervals are extended by half the amount compared to Full2.
- The BSG2 method achieves significant improvement in the coverage criteria compared to BSG1, at the cost of average length that is greater by about 2%. Similarly,

**Table 1** The performance measures of Sect. 5.2 for the different methods when $\alpha \in \{0.01, 0.05, 0.1\}$ and $(n, m) = (9, 6)$.

| | Method | WALD | AC | HS | AM1 | AM2 | BSG1 | BSG2 | Full1 | Full2 | Full3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha = 0.01$ | AVG length | 0.934 | 1.008 | **0.921** | 1.017 | 1.026 | 1.009 | 1.028 | 1.004 | 1.024 | 1.014 |
| | PCT of under-coverage | 100.00% | 27.11% | 57.73% | 4.13% | 0.23% | 5.11% | 0.22% | 6.25% | **0.00**% | 0.76% |
| | PCT of substantial under-coverage | 100.00% | 0.79% | 19.89% | 0.01% | **0.00**% | 0.01% | 0.00% | 0.01% | **0.00**% | 0.01% |
| | AVG deviation | 810.8 | 7.3 | 58.1 | 0.4 | **0.0** | 0.4 | **0.0** | 0.6 | **0.0** | **0.0** |
| | Min CL | 0.029 | 0.959 | 0.9 | 0.925 | 0.987 | 0.925 | 0.988 | 0.925 | 0.99 | 0.969 |
| | AVG CL | 0.909 | 0.993 | 0.987 | 0.994 | 0.995 | 0.994 | 0.995 | 0.994 | 0.994 | 0.994 |
| $\alpha = 0.05$ | AVG length | **0.728** | 0.776 | 0.745 | 0.797 | 0.807 | 0.789 | 0.802 | 0.779 | 0.799 | 0.789 |
| | PCT of under-coverage | 100.00% | 16.44% | 49.91% | 5.69% | 0.31% | 6.51% | 0.38% | 9.98% | **0.00**% | 0.92% |
| | PCT of substantial under-coverage | 100.00% | 3.69% | 21.68% | 0.53% | **0.00**% | 0.70% | 0.09% | 1.95% | **0.00**% | **0.00**% |
| | AVG deviation | 910.2 | 11.6 | 54.5 | 2.3 | 0.1 | 2.8 | 0.2 | 5.4 | **0.0** | 0.1 |
| | Min CL | 0.029 | 0.896 | 0.847 | 0.925 | 0.94 | 0.925 | 0.93 | 0.925 | 0.95 | 0.947 |
| | AVG CL | 0.859 | 0.963 | 0.954 | 0.965 | 0.968 | 0.964 | 0.967 | 0.961 | 0.966 | 0.964 |
| $\alpha = 0.1$ | AVG length | **0.616** | 0.653 | 0.639 | 0.691 | 0.7 | 0.67 | 0.683 | 0.662 | 0.682 | 0.672 |
| | PCT of under-coverage | 99.26% | 17.06% | 43.56% | 4.45% | 0.32% | 6.72% | 0.80% | 10.16% | **0.00**% | 0.63% |
| | PCT of substantial under-coverage | 98.46% | 8.90% | 28.22% | 0.58% | 0.06% | 1.15% | 0.22% | 1.94% | **0.00**% | **0.00**% |
| | AVG deviation | 904.1 | 31.7 | 71.3 | 2.2 | 0.2 | 3.9 | 0.6 | 6.1 | **0.0** | 0.1 |
| | Min CL | 0.029 | 0.731 | 0.808 | 0.851 | 0.885 | 0.851 | 0.874 | 0.851 | 0.9 | 0.894 |
| | AVG CL | 0.81 | 0.92 | 0.909 | 0.928 | 0.932 | 0.924 | 0.93 | 0.92 | 0.929 | 0.925 |

The best method for the criteria AVG length, PCT of under-coverage, PCT of substantial under-coverage and AVG deviation is bald-faced

**Table 2** The performance measures of Sect. 5.2 for the different methods when $\alpha \in \{0.01, 0.05, 0.1\}$ and $(n, m) = (14, 7)$.

| | Method | WALD | AC | HS | AM1 | AM2 | BSG1 | BSG2 | Full1 | Full2 | Full3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha = 0.01$ | AVG length | 0.851 | 0.9 | **0.832** | 0.892 | 0.903 | 0.888 | 0.905 | 0.88 | 0.9 | 0.89 |
| | PCT of under-coverage | 100.00% | 33.33% | 58.65% | 8.31% | 0.39% | 8.81% | 0.42% | 14.68% | **0.00**% | 1.67% |
| | PCT of substantial under-coverage | 98.58% | 0.95% | 15.50% | 0.02% | **0.00**% | 0.02% | **0.00**% | 0.02% | **0.00**% | 0.01% |
| | AVG deviation | 632.2 | 11.3 | 46.2 | 0.7 | **0.0** | 0.8 | **0.0** | 1.4 | **0.0** | 0.1 |
| | Min CL | 0.038 | 0.956 | 0.895 | 0.894 | 0.988 | 0.894 | 0.986 | 0.894 | 0.99 | 0.958 |
| | AVG CL | 0.927 | 0.992 | 0.987 | 0.993 | 0.993 | 0.993 | 0.994 | 0.992 | 0.993 | 0.993 |
| $\alpha = 0.05$ | AVG length | **0.658** | 0.69 | 0.666 | 0.704 | 0.716 | 0.693 | 0.708 | 0.682 | 0.702 | 0.692 |
| | PCT of under-coverage | 99.60% | 18.51% | 47.75% | 6.16% | 0.24% | 7.14% | 0.63% | 14.90% | **0.00**% | 1.47% |
| | PCT of substantial under-coverage | 97.75% | 3.34% | 13.88% | 0.40% | **0.00**% | 0.42% | 0.01% | 1.38% | **0.00**% | **0.00**% |
| | AVG deviation | 738.9 | 11.6 | 39.3 | 2.4 | 0.1 | 2.5 | 0.2 | 6.3 | **0.0** | 0.1 |
| | Min CL | 0.038 | 0.907 | 0.863 | 0.894 | 0.941 | 0.894 | 0.939 | 0.894 | 0.95 | 0.946 |
| | AVG CL | 0.876 | 0.961 | 0.954 | 0.963 | 0.966 | 0.961 | 0.965 | 0.958 | 0.963 | 0.961 |
| $\alpha = 0.1$ | AVG length | **0.555** | 0.58 | 0.568 | 0.61 | 0.619 | 0.585 | 0.604 | 0.578 | 0.597 | 0.588 |
| | PCT of under-coverage | 98.96% | 17.71% | 43.05% | 4.66% | 0.38% | 9.70% | 0.62% | 14.98% | **0.00**% | 1.16% |
| | PCT of substantial under-coverage | 97.48% | 7.05% | 23.90% | 0.52% | 0.02% | 1.75% | 0.08% | 3.64% | **0.00**% | **0.00**% |
| | AVG deviation | 742.5 | 21.0 | 50.4 | 2.2 | 0.1 | 5.6 | 0.3 | 11.2 | **0.0** | 0.2 |
| | Min CL | 0.038 | 0.802 | 0.824 | 0.875 | 0.889 | 0.864 | 0.885 | 0.842 | 0.9 | 0.891 |
| | AVG CL | 0.826 | 0.918 | 0.908 | 0.926 | 0.93 | 0.917 | 0.927 | 0.913 | 0.924 | 0.919 |

The best method for the criteria AVG length, PCT of under-coverage, PCT of substantial under-coverage and AVG deviation is bald-faced

**Table 3** The performance measures of Sect. 5.2 for the different methods when $\alpha \in \{0.01, 0.05, 0.1\}$ and $(n, m) = (10, 10)$.

| | Method | WALD | AC | HS | AM1 | AM2 | BSG1 | BSG2 | Full1 | Full2 | Full3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| alpha=0.01 | AVG length | 0.84 | 0.878 | **0.823** | 0.888 | 0.898 | 0.88 | 0.896 | 0.872 | 0.892 | 0.882 |
| | PCT of under-coverage | 100.00% | 33.22% | 55.63% | 7.34% | 0.43% | 7.42% | 0.61% | 14.01% | **0.00**% | 1.52% |
| | PCT of substantial under-coverage | 99.26% | 0.60% | 15.60% | 0.01% | **0.00**% | 0.01% | **0.00**% | 0.01% | **0.00**% | 0.01% |
| | AVG deviation | 499.7 | 8.2 | 45.8 | 0.7 | **0.0** | 0.6 | **0.0** | 1.1 | **0.0** | 0.1 |
| | Min CL | 0.043 | 0.969 | 0.892 | 0.906 | 0.987 | 0.906 | 0.989 | 0.906 | 0.99 | 0.957 |
| | AVG CL | 0.94 | 0.992 | 0.988 | 0.994 | 0.994 | 0.993 | 0.994 | 0.992 | 0.994 | 0.993 |
| $\alpha = 0.05$ | AVG length | 0.649 | 0.673 | **0.654** | 0.69 | 0.7 | 0.68 | 0.698 | 0.673 | 0.692 | 0.682 |
| | PCT of under-coverage | 100.00% | 21.56% | 42.91% | 9.98% | 0.76% | 11.27% | 1.24% | 14.73% | **0.00**% | 1.24% |
| | PCT of substantial under-coverage | 99.21% | 5.31% | 20.74% | 2.27% | 0.03% | 0.31% | 0.08% | 1.27% | **0.00**% | **0.00**% |
| | AVG deviation | 584.2 | 14.9 | 51.0 | 5.6 | 0.3 | 3.8 | 0.4 | 5.5 | **0.0** | 0.1 |
| | Min CL | 0.043 | 0.907 | 0.835 | 0.906 | 0.936 | 0.906 | 0.939 | 0.906 | 0.95 | 0.946 |
| | AVG CL | 0.892 | 0.96 | 0.954 | 0.962 | 0.965 | 0.961 | 0.965 | 0.959 | 0.964 | 0.961 |
| $\alpha = 0.1$ | AVG length | **0.547** | 0.566 | 0.556 | 0.588 | 0.598 | 0.575 | 0.594 | 0.566 | 0.586 | 0.576 |
| | PCT of under-coverage | 98.15% | 23.24% | 44.90% | 9.31% | 0.74% | 14.35% | 1.53% | 18.32% | **0.00**% | 1.18% |
| | PCT of substantial under-coverage | 92.87% | 11.73% | 26.32% | 2.72% | 0.13% | 4.84% | 0.47% | 7.94% | **0.00**% | 0.01% |
| | AVG deviation | 588.1 | 32.5 | 62.6 | 6.3 | 0.4 | 11.9 | 1.3 | 18.5 | **0.0** | 0.2 |
| | Min CL | 0.043 | 0.743 | 0.801 | 0.835 | 0.883 | 0.835 | 0.872 | 0.835 | 0.9 | 0.865 |
| | AVG CL | 0.841 | 0.916 | 0.907 | 0.919 | 0.924 | 0.917 | 0.927 | 0.912 | 0.923 | 0.918 |

The best method for the criteria AVG length, PCT of under-coverage, PCT of substantial under-coverage and AVG deviation is bald-faced

AM2 improves AM1 in terms of coverage, but the average length increases slightly.

- Out of all the exact methods examined, only BSG2, AM2, Full2 and Full3 have satisfactorily performance for the coverage criteria. The coverage probability of Full2 is always larger than $1-\alpha$ in the above parameters. Comparing BSG2, AM2 and Full3 we can observe that Full3 has the largest percentage of under-coverage, for most of the nine combinations of $n, m$ and $\alpha$ we considered, while AM2 has the lowest. On the other hand, Full3 has the smallest percentage of substantial under-coverage, smaller than 0.01% for all 9 cases. AM2 and BSG2 have slightly higher numbers, yet still very low, ranging from 0.13 to 0.47%.
  Considering the criterion of AVG deviation, all three methods have low scores, in comparison to the other methods. BSG2 has a slightly higher score than AM2 and Full3, which are mostly comparable.
- In addition we checked if the Full3 CIs maintain the nestedness property that was mentioned in the introduction. We checked for every sample sizes $3 \le m \le n \le 15$ and for each sample results $(x, y) \in \{0, 1, \ldots, n\} \times \{0, 1, \ldots, m\}$ that the 99% CI contains the 95% CI, and the latter contains the 90% CI. Overall there are 50 violations of nestedness out of 9, 191 comparisons (0.544%). For example, when the sample sizes are $(n, m) = (10, 7)$ and the observations are $(x, y) = (8, 1)$, the 90% CI is [0.185, 0.855], while the 95% CI is [0.195, 0.895], which does not contain the former interval.

To examine further the under-coverage of the different methods we plotted in Fig. 1 all pairs $(p_1, p_2) \in \mathcal{P}$ for which $CP(p_1, p_2)$ is below $1-\alpha$ when $(n, m, \alpha) = (9, 6, 0.05)$ for all methods excluding WALD.

We observe that AM2, BSG2 and Full3 have similar low under-coverage, but the pattern is a bit different. For AM2 the under-coverage is mostly for pairs $(p_1, p_2)$ that are close $(\frac{1}{2}, \frac{1}{2})$, while for Full3 it is mostly for large $\Delta = |p_1 - p_2|$. The graph of Full2 is empty as there is no under-coverage for this method.

Additionally, Fig. 2 plots the coverage probability as a function of $p_2$ when $p_1 = 0.5$. We can see that all the seven exact methods (AM1, AM2, BSG1, BSG2, Full1, Full2, Full3) exhibit a similar pattern, and the coverage probability is above $1-\alpha$ for almost all $p_2$. Notice that the graph of Full2 has a few short lines (looking like points) in the high-confidence area, which do not exist in the Full3 graph. This is due to the extension of the limits by $0.01/2$ in Full3 compared to the extension of 0.01 in Full2.

Table 4 reports the decrease in the average length of the solutions found by the Full1 method compared to the best lower bound that was computed. It is demonstrated that the gap between the best lower bound and the solution that was found is quite small. Even if the time limit of the algorithm is extended, we believe that it generally would not result in better performance. By observing the outputs of the optimization algorithm throughout the run, it seems that the solution found is optimal or very close to optimal, and more running time will mostly improve the computation of the lower bound, and not the solution itself. For example, for $(n, m, \alpha) = (10, 10, 0.05)$ the solution after 180 s, was the same one that was found after 30 s. The changes were only in the computation of the gap: from 1.67% to 0.87%.

Considering both coverage and length, it seems that Full3 is the best method among the ones we suggested, namely, Full1, Full2, Full3, BSG1 and BSG2. Among the other methods, AM2 has the best performance. Comparing Full3 and AM2, they perform similarly in the coverage criteria but Full3 has a smaller average length.

To examine further the decrease of length of Full3 compared to AM2, we considered 21 pairs of $(n, m)$, where $5 \le m \le n \le 10$. For each such pair and for $\alpha \in \{0.01, 0.05, 0.1\}$ we computed the relative improvement, which is defined by

$$100 \times \frac{\text{AVG length(AM2)-AVG length(Full3)}}{\text{AVG length (AM2)}}. \tag{9}$$

The results are plotted in Fig. 3. We observe that for all 21 pairs Full3 produced shorter intervals and the relative improvement varies from 0.5% to 5%. The larger the $\alpha$, the larger is the relative improvement. For $\alpha = 0.01$, the relative improvement is about 1%, and for $\alpha = 0.05$, the range is from 2.5% to 4%, respectively. It also seems that the relative improvement tends to increase with $n$. In all runs of Full3, the gap between the solution obtained to the lower bound is rather small and the largest gap is 1.35%. Figure 4, which appears in the appendix, extends Fig. 3 to sample sizes $(n, m)$ where $3 \le m \le n \le 15$.

Figures 3 and 4 demonstrate that larger sample sizes lead to more relative improvement. This is because the degrees of freedom of the optimization method are larger for larger sample sizes. Therefore, the advantage of performing full optimization is more significant.

### 5.4 Summary of the findings

The Full algorithm was shown to be computationally feasible for small $n, m$ using the rather coarse grid of $D = \{-1, -0.99, \ldots, 1\}$. While the resulting CIs do not have the right coverage probability for $p_1$ and $p_2$ that are not in the grid, simple adjustments can be made to improve the coverage at a small cost in the average length. The adjusted method, Full3, is comparable, in terms of coverage, to AM2 and BSG2, which are computed under a finer grid, but has shorter CIs.

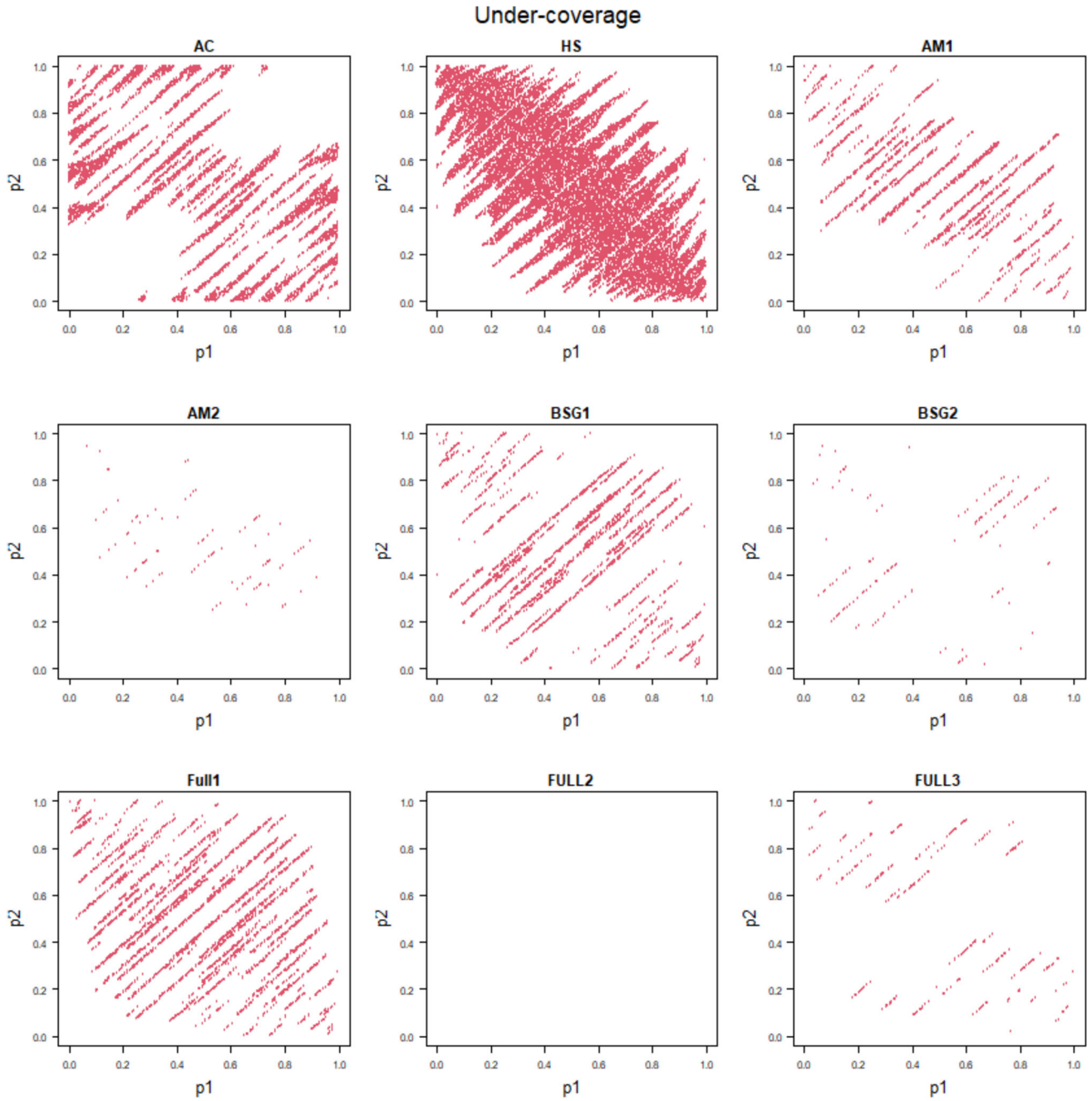


**Fig. 1** Plotting all pairs $(p_1, p_2) \in \mathcal{P}$ for which $CP(p_1, p_2)$ is below $1-\alpha$ when $(n, m, \alpha) = (9, 6, 0.05)$ for all methods listed in Sect. 5.1 besides WALD

**Table 4** A bound of the gap, in terms of percentage of length, between the optimal solution and the one found by Full1 as computed by the Gurobi package

| $\alpha \backslash (n, m)$ | $(n = 9, m = 6)$ (%) | $(n = 14, m = 7)$ (%) | $(n = 10, m = 10)$ (%) |
|---|---|---|---|
| 0.01 | 0.1 | 0.957 | 0.922 |
| 0.05 | 0.33 | 0.94 | 1.08 |
| 0.1 | 0.77 | 1.31 | 1.33 |

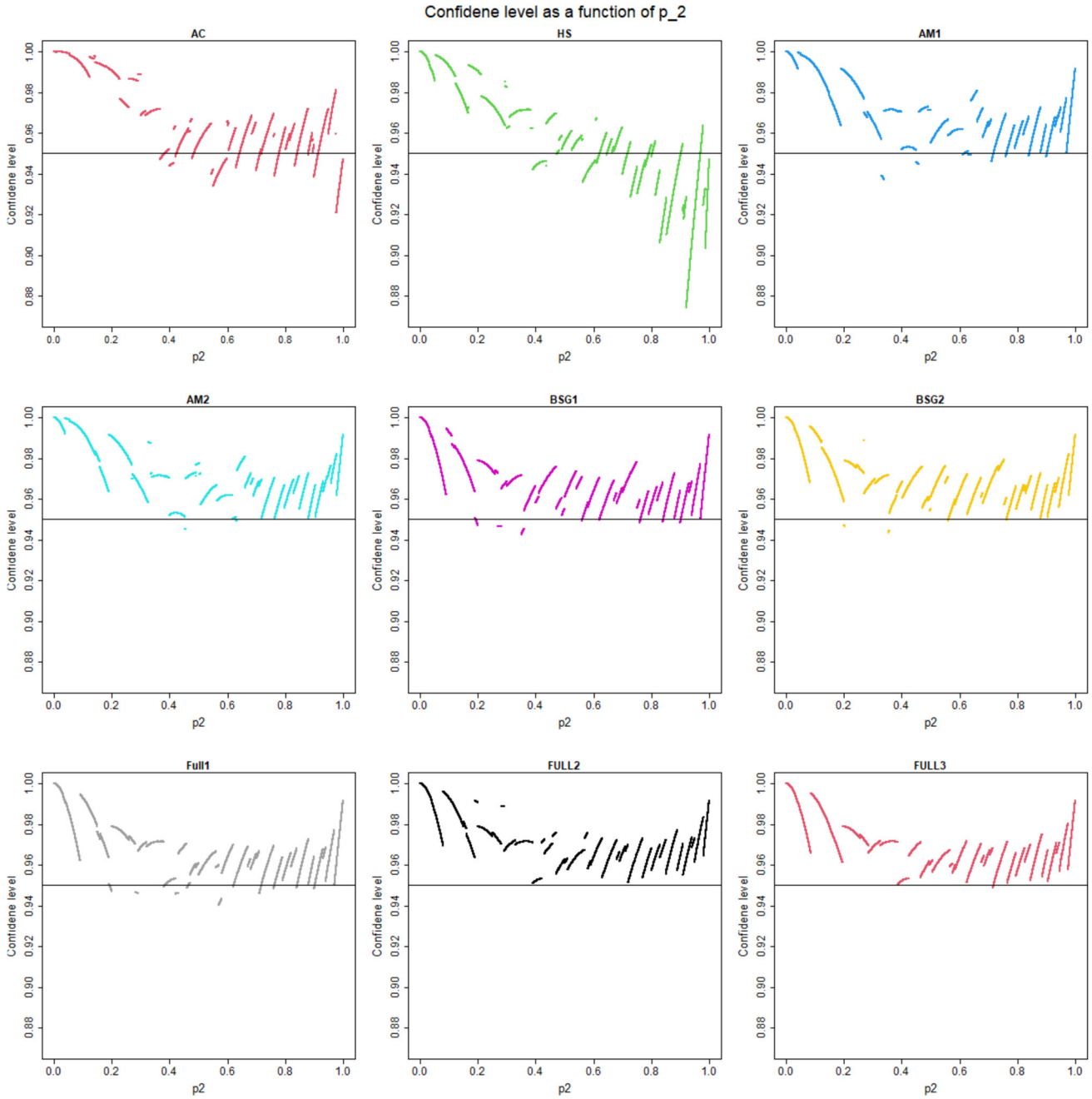


**Fig. 2** Plotting the coverage probability as a function of $p_2$ when $p_1 = 0.5$ and $(n, m, \alpha) = (9, 6, 0.05)$ for for all methods listed in Sect. 5.1 excluding WALD. The vertical line represents the $1 - \alpha = 0.95$ confidence coefficient

## 6 Discussion

For small $n, m$ $(n, m \leq 15)$ we recommended the use of the Full3 method, as it has good coverage and a small average length. Tables for various $(n, m, \alpha)$ of the Full3 method are presented in the following link https://technionmail-my.sharepoint.com/:f:/g/personal/ap_campus_technion_ac_il/El-213Kms51BhQxR8MmQJCYBDfIsvtrK9mQIey1sZnZWIQ?e=hxGunl. The second best method is the AM2 method, and it can be used when Full3 is not available.

We also tried several examples with larger sample sizes than 15. When both sample sizes were 25, the algorithm could not find feasible solutions. For smaller sample sizes (around 20) the results were similar to what was reported in Sect. 5.3. However, a more thorough study is required for larger sample sizes, and we leave this for future research.

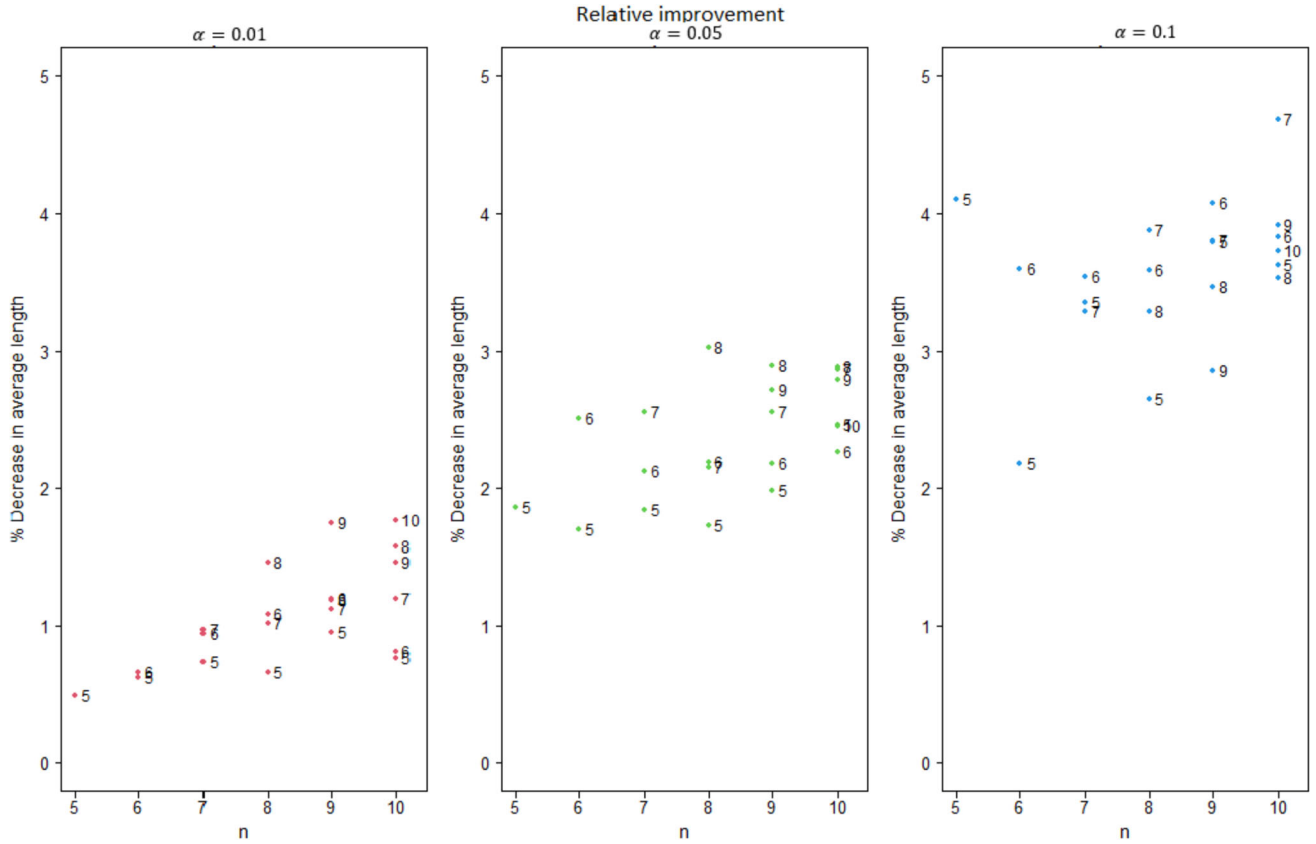


**Fig. 3** Plotting the relative improvement as defined in (9) for every pair $(n, m)$ where $n \in \{5, \ldots, 10\}$ and $m \in \{5, \ldots, n\}$ and for $\alpha \in \{0.01, 0.05, 0.1\}$. The $x$-axis in the graphs is $n$ and the number near each point is the corresponding $m$

Extensions of this work can go in several directions. One can consider extending the Full algorithm to other frequently used discrete distributions, like Poisson or Hyper-geometric. This amounts to changing the coverage criterion (4) according to the distribution used. One can also consider other related optimization problems, for example finding the shortest CIs that have an average confidence coefficient of $1-\alpha$, and the minimal coverage probability is above $1-\beta$ for some $\beta > \alpha$. The availability of powerful optimization algorithms and software allows one to investigate such problems.

**Author Contributions** A.P. and D.A. wrote the manuscript. A.P. developed the algorithm and performed the computations.

**Data availability** No datasets were generated or analysed during the current study.

## Declarations

**Conflict of interest** The authors declare no Conflict of interest.

# 7 Appendix

## A list of notation

- $p_1, p_2$—Proportions of binomial distribution
- $\Delta = p_1 - p_2$—Difference between two proportions
- $1-\alpha$—The stated confidence coefficient
- $n, m$—The sample sizes
- $z_{1-\frac{\alpha}{2}}$—The $1-\frac{\alpha}{2}$ quantile of standard normal distribution
- $X$—Random variable $\sim binomial(n, p_1)$, $Y$ - random variable $\sim binomial(m, p_2)$
- $x, y$—The sample results
- $l_x, u_x$—Lower and upper limit of the confidence interval for $p_1$ when X = x is observed. $l_{(x,y)}, u_{(x,y)}$ lower and upper limit of the confidence interval for $\Delta$ when $(X, Y) = (x, y)$ is observed
- CI—The confidence interval
- $C_1$—The collection of all confidence intervals for one sample case:

$$C_1 := \{[l_x, u_x]\}_{x\in\{0,1,\ldots,n\}}$$

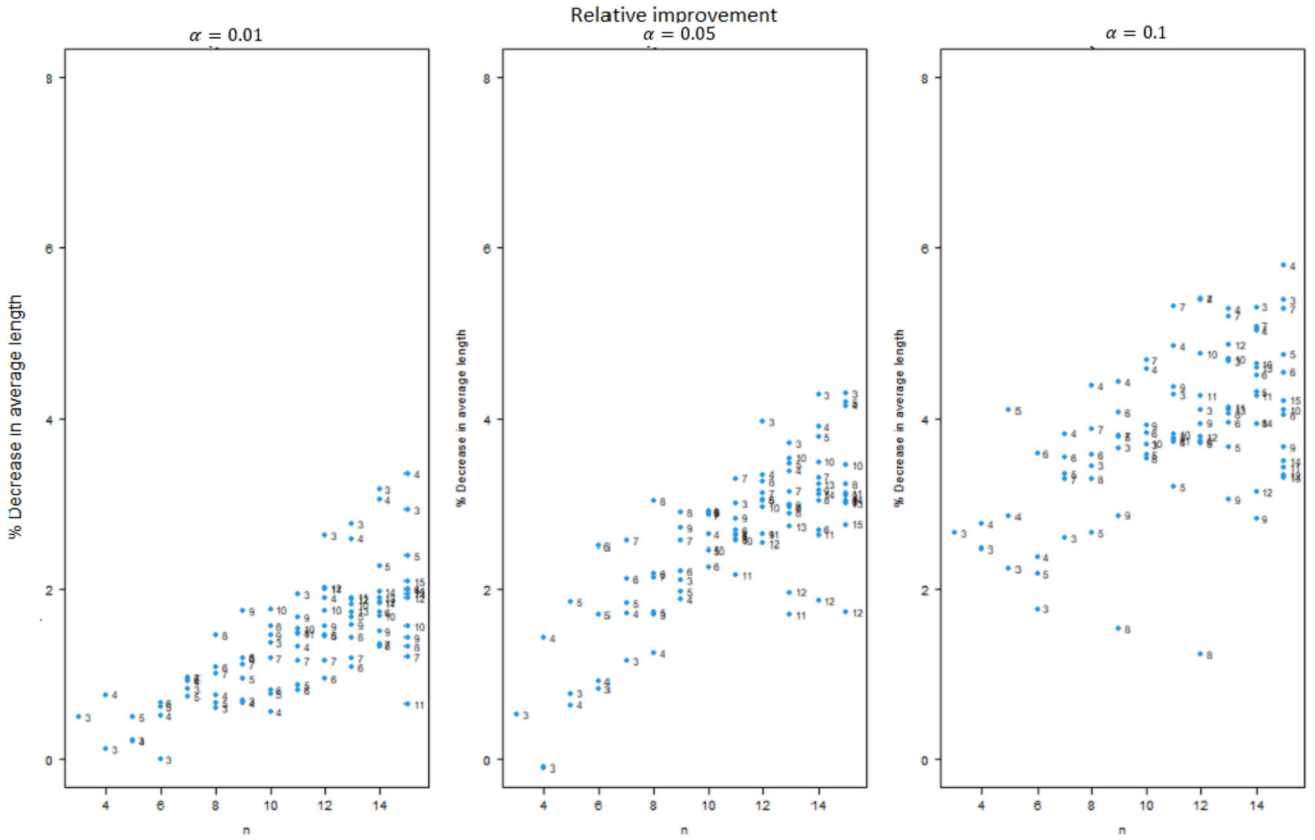


**Fig. 4** Plotting the relative improvement as defined in (9) for every pair $(n, m)$ where $n \in \{3, \ldots, 15\}$ and $m \in \{3, \ldots, n\}$ and for $\alpha \in \{0.01, 0.05, 0.1\}$. The $x$-axis in the graphs is $n$ and the number near each point is the corresponding $m$

- $C_2$—The collection of all confidence intervals for two sample case:

$$C_2 := \{[l_{x,y}, u_{x,y}]\}_{x\in\{0,1,\ldots,n\}, y\in\{0,1,\ldots,m\}}$$

- $D$—Grid for $\Delta$ values
- $P$—Grid for $p_1$, $p_2$ values.

## Relative improvement to additional pairs of sample sizes

Figure 4 displays the relative improvement as shown in Fig. 3 to sample sizes $(n, m)$ where $3 \le m \le n \le 15$.

## The full list of MARs for the example presented in Sect. 3.2

MAR for $\Delta = -0.37$:

$$\begin{aligned} MAR1(-0.37) = \{&(0,1),(0,2),(0,3),(0,4),(0,5),\\ &(1,1),(1,2),(1,3),(1,4),(1,5),\\ &(2,2),(2,3),(2,4),(2,5),(3,3),\\ &(3,4),(3,5),(4,4),(4,5)\} \end{aligned}$$

MARs for $\Delta = -0.4$:

$$\begin{aligned} MAR1(-0.4) = \{&(0,1),(0,2),(0,3),(0,4),(0,5),\\ &(1,1),(1,2),(1,3),(1,4),(1,5),\\ &(2,2),(2,3),(2,4),(2,5),(3,3),\\ &(3,4),(3,5),(4,4),(4,5)\} \end{aligned}$$

$$\begin{aligned} MAR2(-0.4) = \{&(0,1),(0,2),(0,3),(0,4),(0,5),\\ &(1,1),(1,2),(1,3),(1,4),(1,5),\\ &(2,3),(2,4),(2,5),(3,2),(3,3),\\ &(3,4),(3,5),(4,4),(4,5)\} \end{aligned}$$

$$\begin{aligned} MAR3(-0.4) = \{&(0,1),(0,2),(0,3),(0,4),(0,5),\\ &(1,1),(1,2),(1,3),(1,4),(1,5),\\ &(2,1),(2,3),(2,4),(2,5),(3,3),\\ &(3,4),(3,5),(4,4),(4,5)\} \end{aligned}$$

$$\begin{aligned} MAR4(-0.4) = \{&(0,1),(0,2),(0,3),(0,4),(0,5),\\ &(1,1),(1,2),(1,3),(1,4),(1,5),\\ &(2,2),(2,3),(2,4),(2,5),(3,2),\\ &(3,4),(3,5),(4,4),(4,5)\} \end{aligned}$$

$$
\begin{aligned}
&(3,4),(3,5),(4,4),(4,5)\}\\
MAR5(-0.4) = \{&(0,1),(0,2),(0,3),(0,4),(0,5),\\
&(1,1),(1,2),(1,3),(1,4),(1,5),\\
&(2,2),(2,3),(2,4),(2,5),(3,4),\\
&(3,5),(4,3),(4,4),(4,5)\}
\end{aligned}
$$

MAR for $\Delta = -0.38$:

$$
\begin{aligned}
MAR1(-0.38) = \{&(0,1),(0,2),(0,3),(0,4),\\
&(1,1),(1,2),(1,3),(1,4),(1,5),\\
&(2,2),(2,3),(2,4),(2,5),\\
&(3,3),(3,4),(3,5),(4,4),(4,5)\}
\end{aligned}
$$

## References


Agresti, A., Caffo, B.: Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. Am. Stat. **54**(4), 280–288 (2000)

Agresti, A., Coull, B.A.: Approximate is better than "exact" for interval estimation of binomial proportions. Am. Stat. **52**(2), 119–126 (1998)

Agresti, A., Min, Y.: On small-sample confidence intervals for parameters in discrete distributions. Biometrics **57**(3), 963–971 (2001)

Blaker, H.: Confidence curves and improved exact confidence intervals for discrete distributions. Can. J. Stat. **28**(4), 783–798 (2000)

Blyth, C.R., Still, H.A.: Binomial confidence intervals. J. Am. Stat. Assoc. **78**(381), 108–116 (1983)

Brown, L.D., Cai, T.T., DasGupta, A.: Interval estimation for a binomial proportion. Stat. Sci. **16**(2), 101–133 (2001)

Brumback, B., Berg, A.: On effect-measure modification: relationships among changes in the relative risk, odds ratio, and risk difference. Stat. Med. **27**(18), 3453–3465 (2008)

Chan, I.S., Zhang, Z.: Test-based exact confidence intervals for the difference of two binomial proportions. Biometrics **55**(4), 1202–1209 (1999)

Clopper, C.J., Pearson, E.S.: The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika **26**(4), 404–413 (1934)

Crow, E.L.: Confidence intervals for a proportion. Biometrika **43**(3/4), 423–435 (1956)

Fagerland, M.W., Lydersen, S., Laake, P.: Recommended confidence intervals for two independent binomial proportions. Stat. Methods Med. Res. **24**(2), 224–254 (2015)

Fay, M.P., Hunsberger, S.A.: Practical valid inferences for the two-sample binomial problem. Stat. Surv. **15**, 72–110 (2021)

Gurobi Optimization, LLC: Gurobi Optimizer Reference Manual (2023)

Keer, P.: Hypothesis testing in contingency tables: a discussion, and exact unconditional tests for r × c tables. https://repository.tudelft.nl/islandora/object/uuid:7102c72a-ab3e-49a6-9164-131de660053e?collection=education (2023)

Martín Andrés, A., Gayá Moreno, F., Álvarez Hernández, M., Herranz Tejedor, I.: Miettinen and Nurminen score statistics revisited. J. Biopharm. Stat. (2024). https://doi.org/10.1080/10543406.2024.2311242

Miettinen, O., Nurminen, M.: Comparative analysis of two rates. Stat. Med. **4**(2), 213–226 (1985)

Newcombe, R.G.: Interval estimation for the difference between independent proportions: comparison of eleven methods. Stat. Med. **17**(8), 873–890 (1998)

R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2021)

Santner, T.J., Snell, M.K.: Small-sample confidence intervals for p 1–p 2 and p 1/p 2 in 2 × 2 contingency tables. J. Am. Stat. Assoc. **75**(370), 386–394 (1980)

Scherer, R.: PropCIs: Various Confidence Interval Methods for Proportions (2022)

Signorell, A.: DescTools: Tools for descriptive statistics. R package version 0.99.55 (2024)

Sterne, T.E.: Some remarks on confidence or fiducial limits. Biometrika **41**(1/2), 275–278 (1954)

Wilson, E.B.: Probable inference, the law of succession, and statistical inference. J. Am. Stat. Assoc. **22**(158), 209–212 (1927)