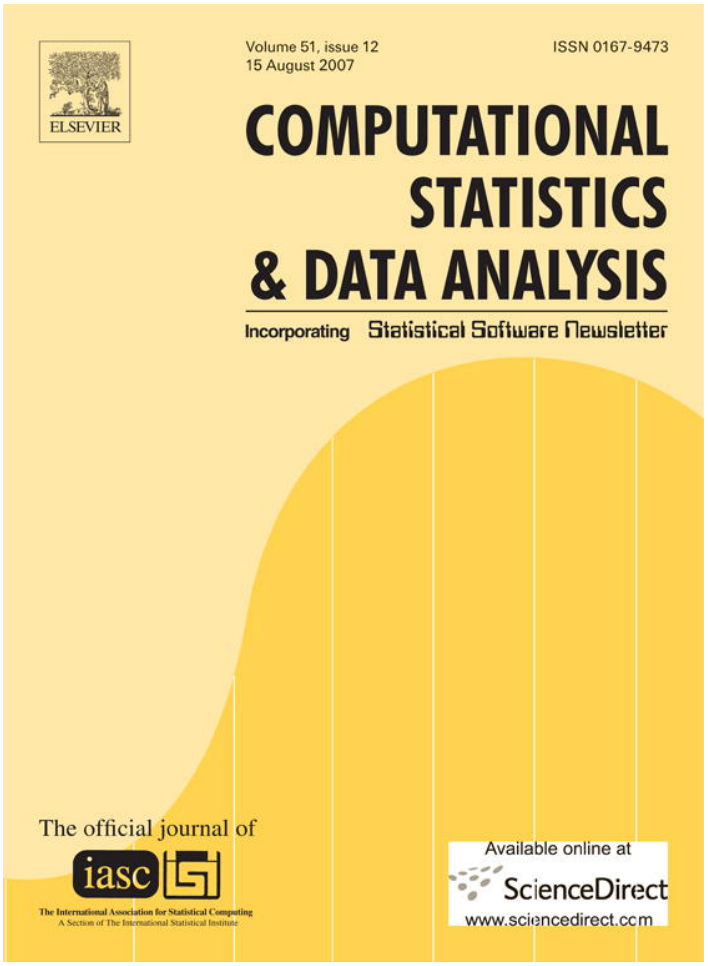

Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 5791–5799
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Small-sample comparisons of confidence intervals for the difference of two independent binomial proportions

Thomas J. Santner[a,*], Vivek Pradhan[b], Pralay Senchaudhuri[b], Cyrus R. Mehta[b,c], Ajit Tamhane[d,e]

[a] *Department of Statistics, The Ohio State University, USA*
[b] *Cytel Inc., USA*
[c] *Harvard School of Public Health, USA*
[d] *Department of Industrial Engineering and Management Sciences, Northwestern University, USA*
[e] *Department of Statistics, Northwestern University, USA*

Received 11 July 2006; received in revised form 6 October 2006; accepted 10 October 2006
Available online 13 November 2006

Abstract

This paper compares the exact small-sample achieved coverage and expected lengths of five methods for computing the confidence interval of the difference of two independent binomial proportions. We strongly recommend that one of these be used in practice. The first method we compare is an asymptotic method based on the score statistic (AS) as proposed by Miettinen and Nurminen [1985. Comparative analysis of two rates. Statist. Med. 4, 213–226.]. Newcombe [1998. Interval estimation for the difference between independent proportions: comparison of seven methods. Statist. Med. 17, 873–890.] has shown that under a certain asymptotic set-up, confidence intervals formed from the score statistic perform better than those formed from the Wald statistic (see also [Farrington, C.P., Manning, G., 1990. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. Statist. Med. 9, 1447–1454.]). The remaining four methods compared are the exact methods of Agresti and Min (AM), Chan and Zhang (CZ), Coe and Tamhane (CT), and Santner and Yamagami (SY). We find that the CT has the best small-sample performance, followed by AM and CZ. Although AS is claimed to perform reasonably well, it performs the worst in this study; about 50% of the time it fails to achieve nominal coverage even with moderately large sample sizes from each binomial treatment.

Keywords: Confidence interval; Exact coverage; *P*-value; Wald statistic; Score statistic; Small-sample; Two-armed bandit

1. Introduction

In biomedical research the difference of two independent binomial proportions is frequently of research interest. Suppose that $X_1 \sim B(n_1, p_1)$ is independent of $X_2 \sim B(n_2, p_2)$. Numerous methods have been proposed for computing a confidence interval for $\Delta \equiv p_2 - p_1$. Katz et al. (1978) proposed inverting an unstandardized test statistic for the

* Corresponding author. Tel.: +1 614 292 3593; fax: +1 614 292 2096.
E-mail address: santner.1@osu.edu (T.J. Santner).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.10.018

family of hypotheses

$$H_0: \Delta = \Delta^\star \quad \text{versus} \quad H_A: \Delta \neq \Delta^\star,$$

where $\Delta^\star \in (-1, +1)$, to construct a Δ confidence interval. Agresti and Caffo (2000) and Brown et al. (2002) showed that inverting the Wald test statistic, the most popular method to compute the confidence interval for Δ, has poor coverage properties. Agresti and Caffo (2000) adjusted the Wald test statistic by adding two successes and two failures to both samples and showed that the resulting "adjusted" Wald statistic Δ interval gives good 95% coverage. Newcombe (1998) compared several asymptotic methods for computing Δ confidence intervals and concluded that the interval based on the score statistic, as proposed by Wilson (1927) for a single binomial proportion, has superior performance characteristics compared with the Δ interval based on the Wald statistic. Nurminen (1986) and Miettinen and Nurminen (1985) proposed a Δ interval, denoted by AS, based on a score statistic for testing H_0 that is computed using restricted maximum likelihood estimation. Farrington and Manning (1990) reviewed several asymptotic methods and recommended use of the AS Δ confidence interval. The AS method is based on the asymptotic normality of the binomial sample proportions and thus may not perform well for small size cases or extreme p_1 and p_2. Santner and Snell (1980) suggested an exact method using an unstandardized test statistic for H_0. Coe and Tamhane (1993) and Santner and Yamagami (1993) proposed exact methods similar to the Blyth and Still (1983) method for single binomial proportion. Lee et al. (1997) introduced a likelihood weighted method for constructing large- and moderate-sample Δ intervals and noted that in their operating characteristic studies the Coe–Tamhane intervals have the best small-sample performance. Chan (1998), Chan and Zhang (1999), and Agresti and Min (2001) proposed a common exact method, referred to as the "tail method" that uses the standardized test statistic. More recently, Chen (2002) proposed a quasi-exact method. The AM, CZ, and AS intervals are currently computed by StatXact.

This paper compares the small-sample performance of Δ confidence intervals. We consider four exact methods, the methods by Agresti and Min (AM), Chan and Zhang (CZ), Coe and Tamhane (CT) and Santner and Yamagami (SY), and their asymptotic counterpart AS. We compared the performance of these methods by finding the achieved coverage and expected lengths for nominal 90% intervals.

2. Statistical methods

Below we describe the various confidence intervals compared in this paper. Each interval system is obtained by inverting a family of tests

$$H_0: \Delta = \Delta^\star \quad \text{versus} \quad H_1: \Delta \neq \Delta^\star \tag{1}$$

corresponding to an arbitrary $\Delta^\star \in (-1, +1)$.

2.1. The AS method

The score statistic for testing (1) is

$$S(\boldsymbol{X}) = \frac{\widehat{p}_2 - \widehat{p}_1 - \Delta^\star}{\sqrt{\widetilde{p}_1(1-\widetilde{p}_1)/n_1 + \widetilde{p}_2(1-\widetilde{p}_2)/n_2}},$$

where $\boldsymbol{X} \equiv (X_1, X_2)$, $\widehat{p}_j \equiv X_j/n_j$ for $j = 1, 2$, and $\widetilde{p}_1$ and $\widetilde{p}_2$ are the maximum likelihood estimates of p_1 and p_2, respectively, under the restriction that $p_2 - p_1 = \Delta^\star$. The test statistic must be defined separately if $x_1 = x_2 = 0$ and $\Delta^\star = 0$. Miettinen and Nurminen (1985) have shown that the restricted maximum likelihood estimates $\widetilde{p}_1$ and $\widetilde{p}_2$ can be obtained by solving the cubic equation

$$\sum_{k=0}^{3} L_k p_1^k = 0 \quad \text{for } p_1 \in [\max\{0, -\Delta^\star\}, \min\{1, 1-\Delta^\star\}],$$

for $\widetilde{p}_1$ and setting $\widetilde{p}_2 = \widetilde{p}_1 + \Delta^\star$ where $L_3 = N$, $L_2 = (n_2 + 2n_1)\Delta^\star - N - x_1 - x_2$, $L_1 = (n_1\Delta^\star - N - 2x_1)\Delta^\star + x_1 + x_2$, and $L_0 = x_1\Delta^\star(1 - \Delta^\star)$; here $N = n_1 + n_2$. Under H_0 the test statistic $S(\boldsymbol{X})$ has mean 0 and variance 1.

The asymptotic $100(1-\alpha)\%$ AS confidence interval $(\underline{\Delta}^{\mathrm{AS}}, \overline{\Delta}^{\mathrm{AS}})$ at $\boldsymbol{x} \equiv (x_1, x_2)$ is obtained by inverting one-sided tests based on $S(\boldsymbol{x})$; this leads to the interval endpoints defined by

$$1-\Phi\left(\frac{\widehat{p}_2-\widehat{p}_1-\underline{\Delta}^{\mathrm{AS}}}{\sqrt{\widetilde{p}_1(1-\widetilde{p}_1)/n_1+\widetilde{p}_2(1-\widetilde{p}_2)/n_2}}\right)=\frac{\alpha}{2}=\Phi\left(\frac{\widehat{p}_2-\widehat{p}_1-\overline{\Delta}^{\mathrm{AS}}}{\sqrt{\widetilde{p}_1(1-\widetilde{p}_1)/n_1+\widetilde{p}_2(1-\widetilde{p}_2)/n_2}}\right).$$

2.2. The CZ method

Chan and Zhang's method is based on using one-sided exact tests of (1) based on the score statistic $S(\boldsymbol{x})$. Let $\boldsymbol{\Omega}=\{\boldsymbol{x}=(x_1,x_2) : 0\leqslant x_j\leqslant n_j,\ j=1,2\}$ denote the set of all possible binomial outcomes. For the given $\Delta^\star$ let $p_1 \in I(\Delta^\star) \equiv (\max(0,-\Delta^\star), \min(1,1-\Delta^\star))$. When $X_1 \sim B(n_1, p_1)$ and $X_2 \sim B(n_2, p_2=p_1+\Delta^\star)$, the probability of observing $\boldsymbol{x}$ is

$$f(\boldsymbol{x})=f(\boldsymbol{x}|p_1,\Delta^\star)=\prod_{j=1}^{2}\binom{n_j}{x_j}p_j^{x_j}(1-p_j)^{n_j-x_j}.$$

Let

$$P_{p_1,\Delta^\star}(S(\boldsymbol{x}))\equiv\sum_{\boldsymbol{y}:S(\boldsymbol{y})\leqslant S(\boldsymbol{x})}f(\boldsymbol{y}|p_1,\ \Delta^\star)\left(Q_{p_1,\Delta^\star}(S(\boldsymbol{x}))\equiv\sum_{\boldsymbol{y}:S(\boldsymbol{y})\geqslant S(\boldsymbol{x})}f(\boldsymbol{y}|p_1,\Delta^\star)\right) \tag{2}$$

denote the probability of obtaining a value of the score statistic that is less than or equal (greater than or equal) to $S(\boldsymbol{x})$. In Eq. (2), p_1 is viewed as a nuisance parameter. The nuisance parameter is eliminated by considering the worst-case p_1 scenario of obtaining small (large) values of $S(\boldsymbol{x})$:

$$P_{\Delta^\star}(S(\boldsymbol{x}))=\sup\{P_{p_1,\Delta^\star}(S(\boldsymbol{x})) : p_1\in I(\Delta^\star)\}\quad\text{and}\quad Q_{\Delta^\star}(S(\boldsymbol{x}))=\sup\{Q_{p_1,\Delta^\star}(S(\boldsymbol{x})) : p_1\in I(\Delta^\star)\}.$$

The level $100(1-\alpha)\%$ CZ confidence interval $(\underline{\Delta}^{\mathrm{CZ}}, \overline{\Delta}^{\mathrm{CZ}})$ at $\boldsymbol{x}$ is the solution of

$$P_{\overline{\Delta}^{\mathrm{CZ}}}(S(\boldsymbol{x}))=\frac{\alpha}{2}=Q_{\underline{\Delta}^{\mathrm{CZ}}}(S(\boldsymbol{x})).$$

2.3. The AM method

The AM method is similar in spirit to the CZ method, but is based on a two-sided test of (1). Set

$$R_{p_1,\Delta}(S(\boldsymbol{x}))=\sum_{\{\boldsymbol{y}:|S(\boldsymbol{y})|\leqslant|S(\boldsymbol{x})|\}}f_{p_1,\Delta^\star}(\boldsymbol{y}).$$

The p_1 nuisance parameter is again eliminated by taking the supremum of $R_{p_1,\Delta^\star}$ over all possible values of p_1 in $I(\Delta^\star)$ and we set

$$R_{\Delta^\star}(S(\boldsymbol{x}))=\sup\{R_{p_1,\Delta^\star}(S(\boldsymbol{x})) : p_1\in I(\Delta^\star)\}.$$

The level $100(1-\alpha)\%$ AM confidence interval $(\underline{\Delta}^{\mathrm{AM}}, \overline{\Delta}^{\mathrm{AM}})$ at $\boldsymbol{x}$ is obtained as follows. Set $\underline{\Delta}^{\mathrm{AM}}$ to be that $\Delta^\star$ value obtained by starting at $\Delta^\star=-1$ and increasing $\Delta^\star$ until

$$R_{\underline{\Delta}^{\mathrm{AM}}}(S(\boldsymbol{x}))=\alpha.$$

Similarly $\overline{\Delta}^{\mathrm{AM}}$ is that value obtained by starting at $\Delta^\star=+1$ and decreasing $\Delta^\star$ until

$$R_{\overline{\Delta}^{\mathrm{AM}}}(S(\boldsymbol{x}))=\alpha.$$

2.4. The CT method

Both the Santner/Yamagami and Coe/Tamhane confidence intervals use greedy heuristics to construct acceptance sets that contain as few $\boldsymbol{x}$ points as possible for testing (1). Both of their algorithms perform this computation for a fine, but finite grid of $\Delta^\star \in (-1, +1)$. The acceptance sets are required to satisfy additional properties to insure that (1) their inversion results in intervals, and (2) the resulting intervals have certain symmetry properties. For example, both systems of intervals are invariant under relabeling of treatments and also under relabeling of the outcomes of "success" and "failure." The methods differ in their choice of greedy heuristic; there is no theory that guarantees either produces an "optimal" set of acceptance sets but the small-sample coverage and length characteristics of these methods differ, as will be seen below. The Santner/Yamagami intervals $(\underline{\Delta}(\boldsymbol{x}), \overline{\Delta}(\boldsymbol{x}))$ satisfy

$$P\{\underline{\Delta}(\boldsymbol{X}) < p_2 - p_1 < \overline{\Delta}(\boldsymbol{X})|\ (p_1, p_2)\} \geqslant 1 - \alpha \tag{3}$$

for all $n_1, n_2 \geqslant 1$ and $0 \leqslant p_1, p_2 \leqslant 1$ and the Coe/Tamhane intervals also satisfy (3) except for rare combinations of $(n_1, n_2, p_1, p_2, \alpha)$ (see Section 3.3). This subsection summarizes the algorithm used to produce the CT intervals while Section 2.5 summarizes the steps used by the SY intervals.

1. Partition the Δ-space, $[-1, +1]$, by the equi-spaced grid $-1 \leqslant \Delta_{-M} < \Delta_{-M+1} < \cdots < 0 = \Delta_0 < \Delta_1 < \cdots < \Delta_M \leqslant +1$, where $\Delta_{-i} = -\Delta_i$, for $1 \leqslant i \leqslant M$, and the number of cut points, M, determines the desired decimal place accuracy of the resulting interval. Set $i = 1$.
2. Partition the p_1-space $[\Delta_i, 1]$ by a grid $0 \leqslant p_{i0} < p_{i1} < \cdots < p_{iN_i}$ symmetrically about the midpoint $(1 + \Delta_i)/2$. As usual, we regard (p_1, p_2) probabilities and their equivalent (p_1, Δ) interchangeably.
3. For each $j = 0, \ldots, N_i$, construct the (non-randomized) acceptance set A_{ij} for testing H_0: $\Delta = \Delta_i$ at level α to be those outcomes $\boldsymbol{x}$ which are most probable when $p_1 = p_{ij}$, i.e., A_{ij} consists of $\boldsymbol{x}$ for which

$$f(\boldsymbol{x}|\ p_1, p_2) \geqslant f(\boldsymbol{y}|\ p_1, p_2) \quad \text{for all } \boldsymbol{x} \in A_{ij} \text{ and } \boldsymbol{y} \notin A_{ij},$$

where $p_1 = p_{ij}$, $p_2 = p_{ij} - \Delta_i$, and

$$\sum_{\boldsymbol{x} \in A_{ij}} f(\boldsymbol{x}|\ p_1, p_2) = P\{\boldsymbol{X} \in A_{ij}|\ p_1 = p_{ij}, \Delta = \Delta_i\} \geqslant 1 - \alpha.$$

4. Construct the (combined) acceptance region $A_i = \bigcup_{j=0}^{N_i} A_{ij}$ of H_0: $\Delta = \Delta_i$.
5. Add any $\boldsymbol{x}$ points to A_i that are required to eliminate "holes" in either the x_1-direction or in the x_2-direction.
6. Let $\widehat{\Delta}(\boldsymbol{x}) = x_1/n_1 - x_2/n_2$ for any outcome $\boldsymbol{x}$. Eliminate any $\boldsymbol{x}^*$ from A_i for which $\boldsymbol{x}^* \notin A_{i-1}$ and $\widehat{\Delta}(\boldsymbol{x}^*) \leqslant \min_{x \in A_{i-1}} \widehat{\Delta}(\boldsymbol{x})$.
7. For each $0 \leqslant i \leqslant M$, let

$$P\{A_i|\Delta_i\} \equiv \inf_{p_1 \in I(\Delta_i)} P\{\boldsymbol{X} \in A_i|\ p_1, \Delta_i\}.$$

then $P\{A_i|\Delta_i\} \geqslant 1 - \alpha$ by construction. Also let

$$\mathscr{D} = \{\boldsymbol{x} \in A_i | P\{A_i - \{\boldsymbol{x}\}|\ \Delta_i\} \geqslant 1 - \alpha\}.$$

(When $n_1 = n_2$ then $\mathscr{D}$ is modified to be the set of pairs of points

$$\mathscr{D} = \{\{\boldsymbol{x}, \boldsymbol{n} - \pi\boldsymbol{x}\} | \{\boldsymbol{x}, \boldsymbol{n} - \pi\boldsymbol{x}\} \in A_i \text{ and } P\{A_i - \{\boldsymbol{x}, \pi\boldsymbol{x}\}|\ \Delta_i\} \geqslant 1 - \alpha\}.)$$

Here $\pi\boldsymbol{x} = (x_2, x_1)$ is the permutation of $\boldsymbol{x} = (x_1, x_2)$. The separate definition for $n_1 = n_2$ insures that the final intervals have certain invariance properties. The set $\mathscr{D}$ can be thought of as "candidate points" for elimination from the acceptance set A_i. Eliminate the point $\boldsymbol{x}^* \in \mathscr{D}$ from A_i where $P\{A_i - \{x^*\}|\ \Delta_i)\} = \max_{x \in \mathscr{D}} P\{A_i - \{x\}|\ \Delta_i\}$ ($\boldsymbol{x}$ is the myopically optimal point to eliminate). (If $n_1 = n_2$, then pairs of points $(\boldsymbol{x}, \boldsymbol{n} - \pi\boldsymbol{x})$ are eliminated from A_i in a similar fashion.)
8. Construct acceptance sets $A_{-i} \equiv \{\boldsymbol{n} - \boldsymbol{x}|\ \boldsymbol{x} \in A_i\}$ corresponding to H_0: $\Delta = \Delta_{-i}$ for $i = 1, \ldots, M$.

9. Invert the acceptance sets $\{A_i\}$ to form the confidence interval $(\underline{\Delta}^{\mathrm{CT}}, \overline{\Delta}^{\mathrm{CT}})$ at $\boldsymbol{x}$ as follows:

$$\underline{\Delta}^{\mathrm{CT}} \equiv \min_{-M \leqslant i \leqslant M} \{\Delta_i : \boldsymbol{x} \in A_i\} \quad \text{and} \quad \overline{\Delta}^{\mathrm{CT}} \equiv \max_{-M \leqslant i \leqslant M} \{\Delta_i : \boldsymbol{x} \in A_i\}.$$

2.5. The SY method

The SY intervals $(\underline{\Delta}^{\mathrm{SY}}, \overline{\Delta}^{\mathrm{SY}})$ use a different myopic algorithm than do the CT intervals to construct a level α acceptance set of the hypothesis H_0: $\Delta = \Delta_i$ where $-M \leqslant i \leqslant M$ and again $-1 \leqslant \Delta_{-M} < \Delta_{-M+1} < \cdots < 0 = \Delta_0 < \Delta_1 < \cdots < \Delta_M \leqslant +1$ is a partition of $[-1, +1]$ satisfying $\Delta_{-i} = -\Delta_i$. They invert the corresponding acceptance sets as do the CT intervals.

The key elements of the SY method are as follows. They let $\widehat{\Delta}(\boldsymbol{x}) = x_1/n_1 - x_2/n_2$, $\mathscr{X}$ denote the sample space of (X_1, X_2), and $-1 = d_{-K} < \cdots < d_0 = 0 < d_{+K} = +1$ denote the distinct values of $\{\widehat{\Delta}(\boldsymbol{x}) | \ \boldsymbol{x} \in \mathscr{X}\}$. They partition $\mathscr{X}$ into the disjoint subsets $\mathscr{X}_i = \{\boldsymbol{x} \in \mathscr{X} | \ \widehat{\Delta}(\boldsymbol{x}) = d_i\}$ for $-K \leqslant i \leqslant +K$. Every level α acceptance set A_i constructed by their algorithm is of the form

$$A_i = \mathscr{B}_i \cup \mathscr{X}_{i+1} \cup \cdots \cup \mathscr{X}_{t-1} \cup \mathscr{C}_t,$$

where $\mathscr{B}_i \subset \mathscr{X}_i$ satisfies $\mathscr{B}_i \neq \emptyset$ and $\mathscr{C}_t \subset \mathscr{X}_t$ satisfies $\mathscr{X}_t - \mathscr{C}_t \neq \emptyset$. Intuitively, the acceptance sets for larger Δ_i correspond to $\boldsymbol{x}$ associated with large $\widehat{\Delta}(\boldsymbol{x})$ point estimates of Δ. The steps of their algorithm are as follows:

1. *Initialization*: Construct the level α acceptance set A_0 for testing H_0: $\Delta = 0$ which satisfies $\boldsymbol{x} \in A_0$ if and only if $\boldsymbol{n} - \boldsymbol{x} \in A_0$ and so that A_0 has as few points as possible. Set $i = 0$.
2. *Update*: Construct A_{i+1} from A_i by tentatively setting $A_{i+1} = A_i$, removing one or more points $\boldsymbol{x}$ from $\mathscr{B}_i$ as long as the remaining points continue to form a level α acceptance set for testing $\mathrm{H}_0 : \Delta = \Delta_{i+1}$. If no points can be removed and A_{i+1} violates coverage requirement (3) for some (p_1, p_2) satisfying $p_2 - p_1 = \Delta_{i+1}$, then add one or more points $\boldsymbol{x}$ from $\mathscr{X}_t - \mathscr{C}_t$ to A_{i+1}.
3. *Termination*: For $1 \leqslant i \leqslant M$, set A_{-i} to be the permutation A_i as in Step 8 of CT and invert the$\{A_i\}_{i=-M}^{M}$ as in Step 9 of CT.

The CT and SY algorithms produce different systems of intervals with substantially different properties. The CT intervals tend to be shorter for outcomes $\boldsymbol{x}$ that are in the "center" portion of the sample space $\mathscr{X}$ while SY intervals tend to be shorter for $\boldsymbol{x}$ on the edge of the $\mathscr{X}$ sample space. Section 3 makes additional, detailed comparisons of the two systems of intervals.

3. Empirical comparisons

3.1. Methods used to compare performance

We compared the performance of the five interval systems described in Section 2 for 90% nominal level intervals. We take the nearness of the achieved coverage as the primary criterion and the expected length as the secondary criterion. For each method and each (n_1, n_2) studied, the 90% confidence interval $(\underline{\Delta}(\boldsymbol{x}), \overline{\Delta}(\boldsymbol{x}))$ was computed for all $(n_1 + 1)(n_2 + 1)$ outcomes $\boldsymbol{x}$. Then for each (p_1, p_2) used in the study, the exact achieved coverage was computed by

$$\zeta(p_1, p_2) = \sum_{\boldsymbol{x} \in \Gamma} \prod_{j=1}^{2} \binom{n_j}{x_j} p_j^{x_j} (1 - p_j)^{n_j - x_j},$$

where $\Gamma = \Gamma(p_1, p_2) = \{\boldsymbol{x} = (x_1, x_2) : \underline{\Delta}(\boldsymbol{x}) \leqslant p_2 - p_1 \leqslant \overline{\Delta}(\boldsymbol{x})\}$, and the exact expected length was computed by

$$\xi(p_1, p_2) = \sum_{\boldsymbol{x} \in \Gamma} \prod_{j=1}^{2} \binom{n_j}{x_{ij}} p_j^{x_j} (1 - p_j)^{n_j - x_j} (\overline{\Delta}(\boldsymbol{x}) - \underline{\Delta}(\boldsymbol{x})).$$

The AS, CZ, AM intervals were computed using StatXact PROCs. The SY interval was computed using the FORTRAN 77 code written by Yamagami (available from the authors). The SAS code implementing the CT method

(available from the authors) is not always reliable (see Sun et al., 2002). We implemented the CT method in C++ for this paper. After computing the 90% confidence intervals for each system, we determined the achieved coverage and expected length for a 100×100 uniform grid of $\boldsymbol{p} = (p_1, p_2)$ values in $[0, 1] \times [0, 1]$ ($100^2 = 10,000$ points $\boldsymbol{p}$). The distribution of coverages and expected lengths of each system and each (n_1, n_2) are displayed below using BliP plots (Lee and Tu, 1997). The vertical bars in each plotting symbol show the deciles of each achieved coverage or expected length distribution.

3.2. Results

We studied seven $\boldsymbol{n} = (n_1, n_2)$ sample size cases; three of these cases were balanced ($\boldsymbol{n} \in \{(5, 5),(15, 15), (30, 30)\}$) and four cases were unbalanced ($\boldsymbol{n} \in \{(5, 15), (15, 25), (25, 35), (20, 50\}$). Each row of two panels in the figure below corresponds to one of the seven $\boldsymbol{n}$ configurations. Within each row, the left and right panels display the distributions of the achieved coverages and expected lengths for the five methods over the $10,000\,\boldsymbol{p}$ cases, respectively. The vertical line in each left panel is the 90% nominal coverage level. In each panel, the distributions are denoted AS for the asymptotic score statistic method, CZ for the Chan and Zhang method, AM for the Agresti and Min method, CT for the Coe and Tamhane method and SY for the Santner and Yamagami method (Fig. 1).

3.3. Discussion

The AM, CZ, SY intervals achieved at least the 90% nominal coverage for all choices of $\boldsymbol{n}$ and $\boldsymbol{p}$ while the CT intervals have undercoverage for a few $\boldsymbol{n}$ and $\boldsymbol{p}$ cases. For example, when $n_1 = 15 = n_2$, CT intervals failed to achieve the nominal 90% coverage in 36 of the $10,000\,\boldsymbol{p}$ combinations computed and when $n_1 = 30 = n_2$, they failed to achieve the nominal 90% coverage in 56 of the $10,000\,\boldsymbol{p}$ combinations computed. This very occasional undercoverage occurs because Step 6 can eliminate outcomes that cause non-monotonicity in the estimated Δ values for the acceptance sets of two adjacent null hypothesis values, Δ_i and Δ_{i+1}. In contrast, the asymptotic AS method fell short nearly 50% of the time. In particular, when $n_1 = 15 = n_2$, over 65% of the $\boldsymbol{p}$ cases fell below 90% and when $n_1 = 30 = n_2$, over 58% of the $\boldsymbol{p}$ cases fell below the nominal level. In the unbalanced situations, the coverage of AS method also failed to achieve nominal coverage in nearly 50% of $\boldsymbol{p}$ cases.

When the achieved expected lengths for each method by each $\boldsymbol{n}$ are examined (see the right panels, above), the CT method consistently has the shortest achieved expected length of all methods followed by the AM and CZ methods. The SY method is very conservative and most of the time it has the highest expected length of all five methods.

4. Conclusion and recommendation

Based on the comparison of the distribution of achieved coverages and expected lengths for the five methods compared in this paper, we recommend the CT method be used to compute confidence intervals for the difference of two binomial proportions. If CT method is not available, we recommend that either the AM or CZ be used. We also recommend that neither the AS nor the SY methods be used in practice. The asymptotic AS method is supposed to work well for the large sample cases, however, even with the largest sample size case we studied, this was not true. The AS method failed to achieve the nominal coverage roughly 50% of the time. While having good achieved coverage, the SY method has inferior expected length compared to some of the other methods.

In some applications, confidence bounds for $p_1 - p_2$ (one-sided confidence intervals) are of primary scientific interest. For example, in a drug superiority trial, one can compare the lower bound for $p_1 - p_2$ with zero to make inference about which group is better. In non-inferiority trials with a prespecific margin, one can compare the upper bound for $p_1 - p_2$ with the margin to evaluate whether the non-inferiority criterion is met. In such applications, one-sided hypothesis tests are the relevant quantities whose acceptance regions are to be inverted. Of the two-sided tests considered in this paper, the AS and CZ tests are devised from one-sided tests and hence of interest in such cases. However, we note that in cases where the scientific application at hand does require a two-sided Δ confidence interval, intervals based on tests that directly minimize the number of points in the acceptance set for testing (1) can be substantially superior to intervals formed from tests obtained by intersecting two one-sided tests.

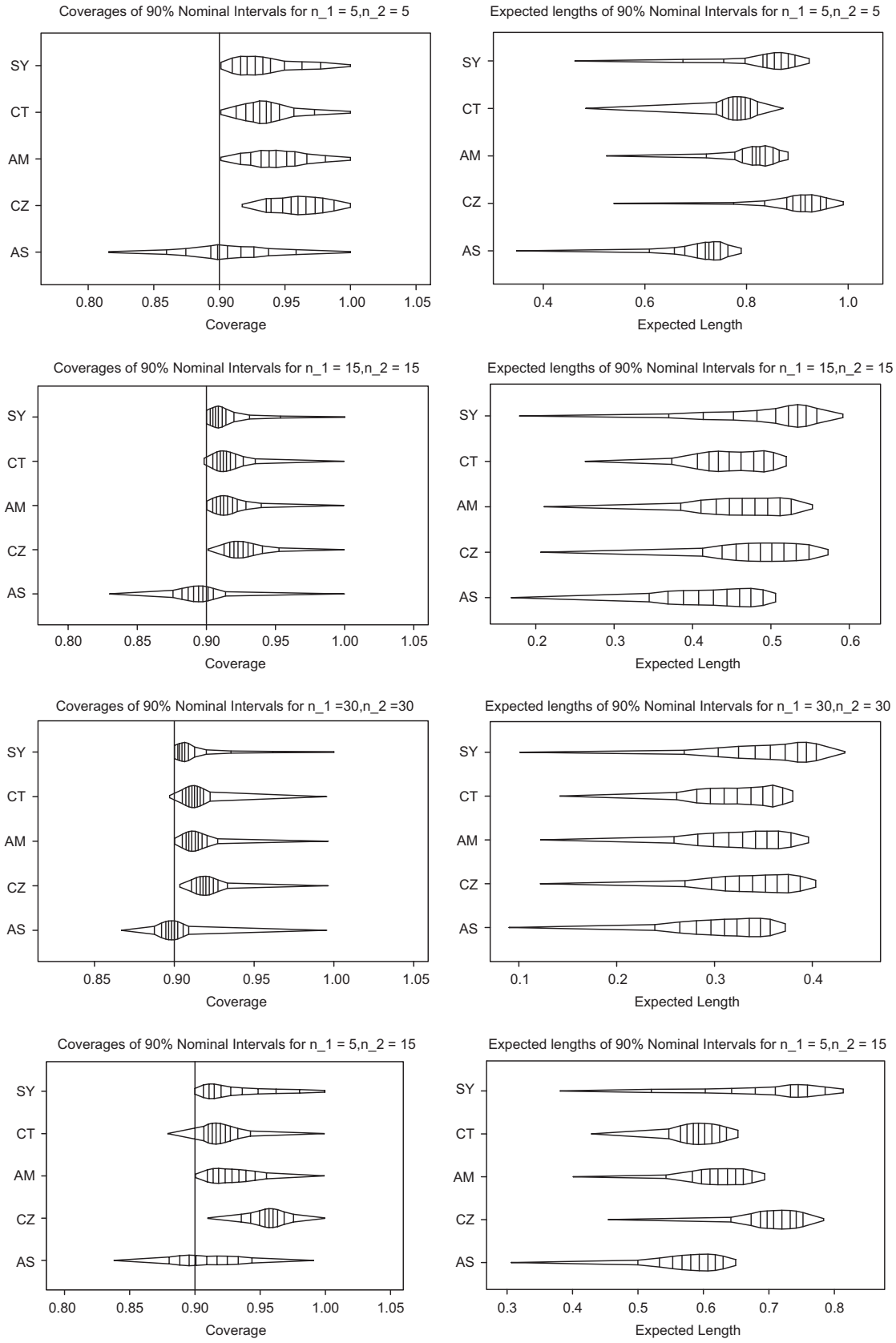

Fig. 1. Distributions of expected lengths and coverages for SY, CT, AM, CZ, and AS over 10,000 equally-spaced (p_1, p_2) values in $[0, 1] \times [0, 1]$ for seven (n_1, n_2) small- and medium-sized pairs.

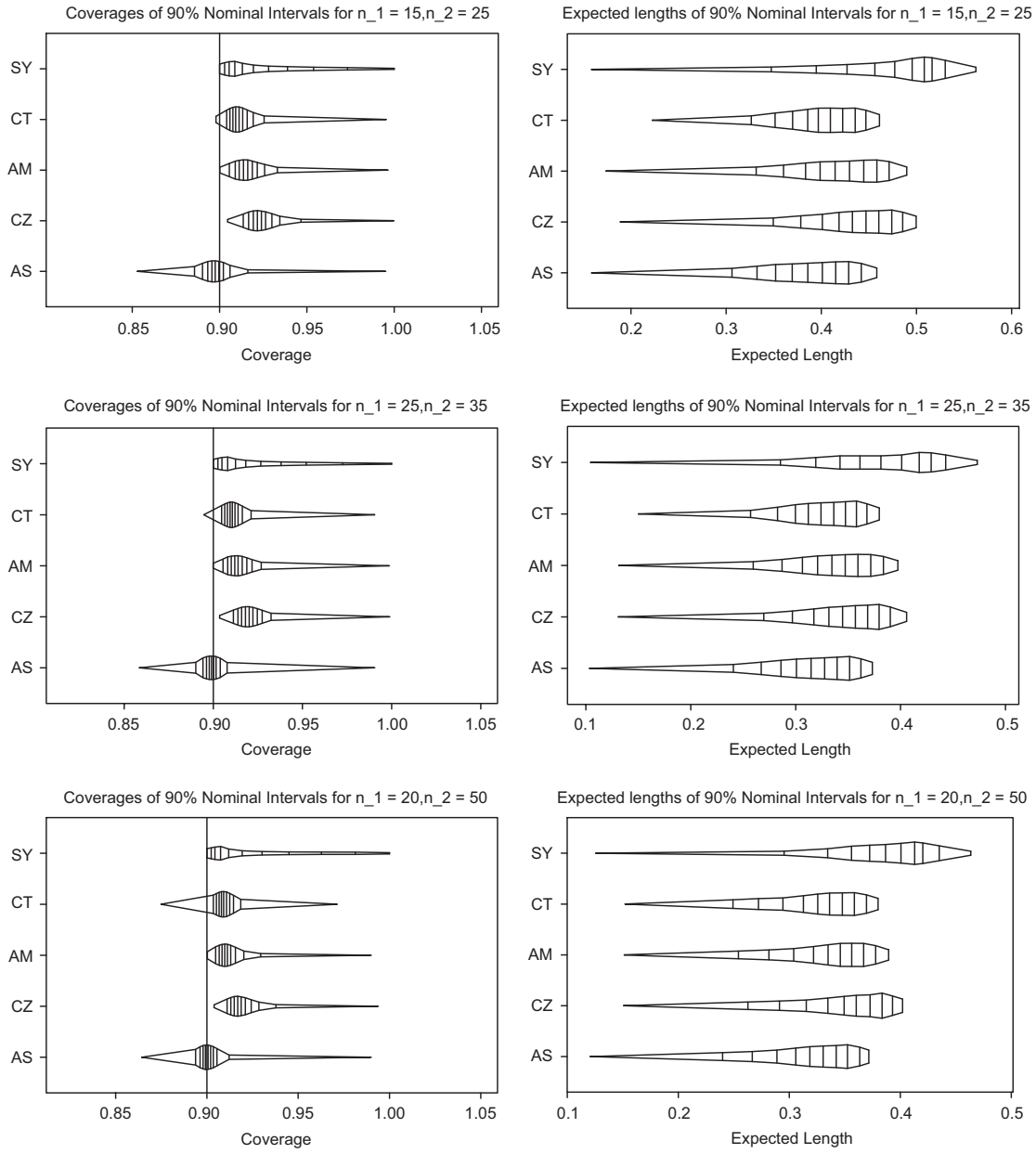

Fig. 1. *Continued*

Acknowledgements

The authors would like to thank Cheryl Dingus and Junfeng Sun, who performed some early comparisons of the methods and Paul Coe for help in understanding his SAS macro implementing the CT method. We would also like to thank the referee for suggestions that led to the improvement of this paper. The work of the first author was partially supported by the National Science Foundation under Grant DMS 0406026.

References

Agresti, A., Caffo, B., 2000. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. Amer. Statist. 54, 280–288.

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

Blyth, C., Still, H., 1983. Binomial confidence intervals. J. Amer. Statist. Assoc. 78, 108–116.
Brown, L.D., Cai, T.T., DasGupta, A., 2002. Confidence Intervals for a Binomial Proportion and Asymptotic Expansion. Ann. Statist. 30, 160–201.
Chan, I., 1998. Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies. Statist. Med. 17, 1403–1413.
Chan, I.S.F., Zhang, Z., 1999. Test-based exact confidence intervals for the difference of two binomial proportions. Biometrics 55, 1201–1209.
Chen, X., 2002. A quasi-exact method for the confidence intervals of the difference of two independent binomial proportions in small sample cases. Statist. Med. 21, 943–956.
Coe, P.R., Tamhane, A.C., 1993. Small sample confidence intervals for the difference, ratio, and odds ratio of two success probabilities. Comm. Statist. Part B–Simulation Comput. 22, 925–938.
Farrington, C.P., Manning, G., 1990. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. Statist. Med. 9, 1447–1454.
Katz, D., Baptista, J., Azen, S.P., Pike, M.C., 1978. Obtaining confidence intervals for the risk ratio in cohort studies. Biometrics 34, 469–474.
Lee, J.J., Serachitopol, D.M., Brown, B.W., 1997. Likelihood-weighted confidence intervals for the difference of two binomial proportions. Biometrical J. 39, 387–407.
Lee, J.J., Tu, Z.N., 1997. A versatile one-dimensional distribution plot: the BLiP plot. Amer. Statist. 51, 353–358.
Miettinen, O.S., Nurminen, M., 1985. Comparative analysis of two rates. Statist. Med. 4, 213–226.
Newcombe, R.G., 1998. Interval estimation for the difference between independent proportions: comparison of seven methods. Statist. Med. 17, 873–890.
Nurminen, M., 1986. Analysis of trends in proportions with an ordinally scaled determinant. Biometrical J. 28, 965–974.
Santner, T.J., Snell, M.K., 1980. Small-sample confidence intervals for $p_1 - p_2$ and p_1/p_2 in 2×2 contingency tables. J. Amer. Statist. Assoc. 75, 386–394.
Santner, T.J., Yamagami, S., 1993. Invariant small sample confidence intervals for the difference of two success probabilities. Comm. Statist. Part B–Simulation Comput. 22, 33–59.
Sun, J.F., Santner, T.J., Venard, C., 2002. An empirical comparison of small sample confidence intervals for $p_1 - p_2$. Technical Report #681, Ohio State University.
Wilson, E.B., 1927. Probable inference, the law of succession, and statistical inference. J. Am. Stat. Assoc. 22, 209–212.