INVARIANT SMALL SAMPLE CONFIDENCE INTERVALS FOR THE DIFFERENCE OF TWO SUCCESS PROBABILITIES

Thomas J. Santner
Department of Statistics
1958 Neil Avenue
Ohio State University
Columbus, OH 43210

Shin Yamagami
Information Technology Research Institute
Tokyo Gas Company, Ltd.
2-3 Nakase Chiba 260
Tokyo, Japan

Keywords: Berkson's simple difference; relative risk; confidence interval; comparative binomial experiment.

ABSTRACT

An algorithm is proposed for determining $100 \times (1-\alpha)\%$ small-sample confidence intervals for the difference $\Delta \equiv p_1$ - p_2 of two binomial success probabilities based on n_1 and n_2 trials, respectively. The interval covers the true Δ with probability at least $(1-\alpha)$ for all $0 < p_1,\ p_2 < 1$; it is invariant with respect to relabeling of the two populations and with respect to interchanging the outcomes of success and failure intervals. Coverage and expected length comparisons are made with the small sample $100 \times (1-\alpha)\%$ tail intervals of Santner and Snell (1980). A FORTRAN program implementing the algorithm is available from the authors.

1 Introduction and Summary

In medicine, biology, and other areas of scientific inquiry one is often faced with the problem of comparing two binomial success probabilities p_1 and p_2, $0 < p_1,\ p_2 < 1$, based on independent observations $X_1 \sim$ B(n_1,p_1) and $X_2 \sim$ B(n_2,p_2). Although the comparison is sometimes made by means of the hypothesis test H_0: $p_1 = p_2$ versus H_A: $p_1 \neq p_2$,

many practitioners often realize this formulation is inappropriate should they reject H_0 and then desire to make a stronger inference than $p_1 \neq p_2$. Thus many statisticians, beginning with Cornfield (1956), have suggested that scientists formulate their treatment comparison problems as one of interval estimation of an appropriate scalar function of p_1 and p_2. Three quantities often used to compare p_1 and p_2 are:

(i) the odds ratio $\psi \equiv p_1(1-p_2)/(1-p_1)p_2$,

(ii) the relative risk $\rho \equiv p_1/p_2$, and

(iii) the difference $\Delta \equiv p_1 - p_2$.

The odds ratio is perhaps the most difficult of (i) - (iii) to interpret although it is the easiest for which to form confidence intervals. Conversely, ρ and Δ are more difficult to construct confidence intervals for but easier to interpret.

This paper proposes an algorithm to construct invariant small- sample confidence intervals for Δ; the proposed intervals attain at least their nominal level for all (n_1, n_2, p_1, p_2), and are invariant with respect to the interchange of p_1 and p_2 as well as the definition of success and failure. Formally, the Δ intervals constructed below satisfy the coverage requirement

$$P_{\mathbf{p}}\{\underline{\Delta}(\mathbf{X}) < \Delta < \overline{\Delta}(\mathbf{X})\} \geq 1-\alpha \tag{1}$$

for *all* p_1, p_2 and n_1, n_2; here $\mathbf{p} = (p_1, p_2)$, $\mathbf{X} = (X_1, X_2)$, and $\underline{\Delta}(\mathbf{x})$ and $\overline{\Delta}(\mathbf{x})$ are the lower and upper limits of the confidence interval at $\mathbf{x} = (x_1, x_2)$, respectively. The notation $P_{\mathbf{p}}\{\mathbf{X} \in E\}$ indicates a probability calculated when $X_i \sim B(n_i, p_i)$ for i = 1, 2.

The computer intensive methods described in this paper complement the previous work on large-sample intervals for Δ. For example, Beal (1987) compares five large-sample Δ intervals with respect to their coverage probabilities. Of course, even the best large-sample interval is anticonservative for sufficiently small n_1, n_2 and extreme (p_1, p_2). To illustrate, Figure 1 displays the anticonservatism of the (large sample) nominal 95% Jeffrey-Perks (JP) interval (Beal, 1987; p. 945). The horizontal axis is Δ and the vertical axis is the lower envelope

$$\inf\left\{P_{\mathbf{p}}\left\{\Delta \in \text{JP interval}\right\} : \mathbf{p} \text{ such that } p_1 - p_2 = \Delta\right\}$$

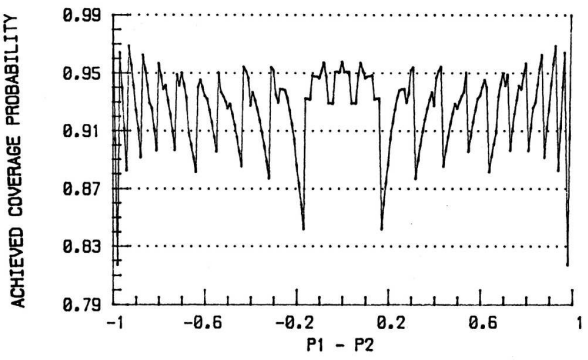

Figure 1: Plot of $(\Delta, \inf_Q P_{\boldsymbol{p}}\{\Delta \in \text{JP Interval}\}$ where $Q = \{\boldsymbol{p} | p_1 - p_2 = \Delta\}$ for nominal 95% Confidence Intervals when $n_1 = 10 = n_2$

for a grid of Δ values when $n_1 = 10 = n_2$. The JP interval's coverage probability is particularly affected for extreme p_1 or p_2. The iteratively computed Mee (1984) Δ intervals described in Beal are similarly anticonservative.

Santner and Snell (1980) and more recently Coe and Tamhane (1991) have considered the problem of determining *small-sample* Δ (and ρ) confidence intervals. Santner and Snell (1980) constructed two computationally feasible intervals (a 'conditional' interval and a 'tail' interval) and, in principle, a third interval which is too complex to be computationally competitive. For later comparison, the end of this section describes tail (T) intervals, the superior of the two computationally feasible intervals. Some comparisons with Coe and Tamhane (1991) intervals will be given.

The iteratively computed interval proposed in Section 2 is computationally feasible and generally less conservative than the Santner and Snell T intervals. Section 3 establishes the invariance properties of the proposed intervals and compares them with those of T intervals. Section 4 compares the expected length of the proposed and T intervals in a specific example and discusses our computational experience with a FORTRAN 77 program which implements the Section 2 algorithm. The remainder of this section reviews briefly aspects of the one- dimensional Crow (1956) algorithm for determining a confidence interval for a single binomial p; these ideas will be required in our analysis of the two-dimensional Δ problem.

Every conservative $100 \times (1-\alpha)\%$ confidence interval for p, $0 < p < 1$, based on single sample data $Y \sim B(n,p)$ results from inverting a family of *acceptance* regions $\{A(p_0) : 0 < p_0 < 1\}$; the region A(p_0) corresponds to a size $\leq \alpha$ test of H_0: $p = p_0$ versus $H_A : p \neq p_0$. Thus

$$P_{p_0}\{Y \in A(p_0)\} \geq 1-\alpha \tag{2}$$

for $0 < p_0 < 1$ where $P_{p_0}\{\cdot\}$ indicates a probability calculated under $Y \sim B(n,p_0)$.

Tail intervals for p use $A(p_0) = \{\ell(p_0), \ldots, u(p_0)\}$ to test H_0: $p = p_0$ where $\ell = \ell(p_0) \geq 0$ is the largest integer for which $P_{p_0}\{Y \leq \ell - 1\} \leq \alpha/2$ and $u = u(p_0) \leq n$ is the smallest integer for which $P_{p_0}\{Y \geq u+1\} \leq \alpha/2$ (Clopper and Pearson, 1934). It is straightforward to show inversion of the set of $A(p_0)$ yields $(\underline{p}(y), \overline{p}(y))$ defined by $\underline{p}(0) = 0$, $\underline{p}(n) = 1$ and otherwise by $P_{\underline{p}(y)}\{Y \geq y\} = \alpha/2 = P_{\overline{p}(y)}\{Y \leq y \mid\}$. The critical feature of the tail method is that for each outcome y, there be defined sets of outcomes ("upper and lower tails") which are "more extreme" than y and "less extreme" than y. In the binomial problem, $\{y, \ldots, n\}$ and $\{0, \ldots, y\}$ are the "upper" and "lower" tails corresponding to the outcome y, respectively.

For the two-sample binomial Δ interval problem, Santner and Snell (1980) define tails corresponding to the outcome $\mathbf{x}$ by $\mathcal{U}(\mathbf{x}) := \{\mathbf{w} = (w_1, w_2) : \hat{\Delta}(\mathbf{w}) \geq \hat{\Delta}(\mathbf{x})\}$ and $\mathcal{L}(\mathbf{x}) := \{\mathbf{w} = (w_1, w_2) : \hat{\Delta}(\mathbf{w}) \leq \hat{\Delta}(\mathbf{x})\}$ where $\hat{\Delta}(\mathbf{w}) = w_1/n_1 - w_2/n_2$. In words, $\mathcal{U}(\mathbf{x})$ ($\mathcal{L}(\mathbf{x})$) is the set of outcomes consistent with point estimates of Δ at least as large (small) as that suggested by $\mathbf{x}$.

Returning to the one-sample binomial p problem, Sterne (1954) proposed using "most probable" (MP) acceptance regions for H_0: $p = p_0$ versus H_A: $p \neq p_0$ defined by (2) and $P_{p_0}\{Y = i\} > P_{p_0}\{Y = j\}$ for $i \in A(p_0)$ and j $\notin A(p_0)$. It is straightforward to show his $A(p_0)$ are also of the form $\{\ell(p_0), \ldots, u(p_0)\}$ and have minimum cardinality among all $100(1-\alpha)\%$ acceptance regions. Sterne's intuitive motivation for this system was that *shorter* intervals should result from inverting *smaller* $A(p_0)$.

Crow (1956) shows, by example, that unfortunately Sterne's $A(p_0)$ need not invert to intervals. Crow also proves that a necessary and sufficient condition for intervals to result

from inverting acceptance regions of the form $\{\ell(p_0), \ldots, u(p_0)\}$ is Equation (3).

$$\ell(p_0) \text{ and } u(p_0) \text{ must be non-decreasing in } p_0. \tag{3}$$

Crow (1956) and Blyth and Still (1983) use algorithmic methods to construct acceptance sets which guarantee (3). They fix a partition $0 < p_1 < \ldots < p_i < \ldots < 1$ and iteratively construct $A(p_i) = \{\ell(p_i), \ldots, u(p_i)\}$ so that $\ell(p_i)$ and $u(p_i)$ are nondecreasing in p_i. In addition, with but a few exceptions, their $A(p_i)$ contain the same number of points as do Sterne's MP acceptance regions. Their algorithms force the $A(p_i)$ to be small by constructing $A(p_i)$ from $A(p_{i-1})$ by first trying to *eliminate* points from $A(p_{i-1})$ (set $\ell(p_i) = \ell(p_{i-1}) + 1$ and $u(p_i) = u(p_{i-1})$), or if not possible by *substitution* (set $\ell(p_i) = \ell(p_{i-1}) + 1$ and $u(p_i) = u(p_{i-1}) + 1$), or if not possible then only by the *addition* of a point to $A(p_i)$ (set $\ell(p_i) = \ell(p_{i-1})$ and $u(p_i) = u(p_{i-1}) + 1$). The exact details of the algorithm can vary. A linear order must exist among the sample outcomes to construct such an algorithm.

2 The Algorithm

2.1 Overview

The algorithm in Section 2.2 adapts Crow's technique to this nuisance parameter problem by constructing $100 \times (1-\alpha)\%$ acceptance regions for an increasing sequence of Δ values. Here we regard the parameters as Δ and (the nuisance quantity) p_1 rather than p_1 and p_2; for fixed Δ, p_1 ranges over $\mathcal{I}(\Delta)$ defined by

$$\mathcal{I}(\Delta) := \begin{cases} (0, \Delta + 1), & -1 < \Delta < 0 \\ (0, 1), & \Delta = 0 \\ (\Delta, +1), & 0 < \Delta < 1. \end{cases}$$

We let $P_{\Delta,p_1}\{\mathbf{X} \in E\}$ denote the probability that $\mathbf{X}$ is in E when $X_1 \sim B(n_1, p_1)$ and $X_2 \sim B(n_2, p_2)$.

The first decision that must be made to define the algorithm is to select an (arbitrary) linear order for the bivariate outcomes (x_1, x_2). This paper orders $\mathbf{x}$ corresponding to increasing values of $\hat{\Delta}(\mathbf{x}) := x_1/n_1 - x_2/n_2$, the estimated Δ value at $\mathbf{x}$. Second, for each Δ_0 it must be noted that the acceptance region $A(\Delta_0)$ for testing H_0: $\Delta = \Delta_0$ versus H_1:

$\Delta \neq \Delta_0$ must be constructed to have *minimum* coverage at least $(1-\alpha)$ over the nuisance parameter $p_1 \in \mathcal{I}(\Delta_0)$. Last, the algorithm below is used to compute intervals only when $n_1 \leq n_2$. When $n_1 > n_2$ we define the Δ interval at $\mathbf{x} = (x_1, x_2)$ by

$$\big(\underline{\Delta}(\mathbf{x};\mathbf{n}), \overline{\Delta}(\mathbf{x};\mathbf{n})\big) = \big(-\overline{\Delta}(\pi\mathbf{x};\pi\mathbf{n}), -\underline{\Delta}(\pi\mathbf{x};\pi\mathbf{n})\big) \tag{4}$$

where

$$\pi\mathbf{w} = \pi(w_1, w_2) = (w_2, w_1)$$

denotes the point with coordinates reversed, $\mathbf{n} = (\mathbf{n_1}, \mathbf{n_2})$, and the interval on the right hand side is produced by the algorithm. The intuitive motivation for (4) is that the point estimates $\hat{\Delta}$ at $\mathbf{x}$ and $\pi\mathbf{x}$ for the problems with sample sizes $\mathbf{n}$ and $\pi\mathbf{n}$, respectively, satisfy $\hat{\Delta}(\mathbf{x};\mathbf{n}) = x_1/n_1 - x_2/n_2 = -(x_2/n_2 - x_1/n_1) = -\hat{\Delta}(\pi\mathbf{x};\pi\mathbf{n})$ which suggests the interval should have the analogous property. It will be proved below that (4) and an appropriate construction when $n_1 = n_2$ guarantees that the intervals are consistent under interchange of population labels. Furthermore, the intervals are also constructed to be invariant when successes and failures are interchanged. This is made precise in Theorem 3.1 which shows that Step 4.1 of the algorithm and the construction of A(0) insure this property.

Formally, the algorithm first partitions the sample space $\mathcal{X} \equiv \{\mathbf{x} = (x_1, x_2) : x_1 = 0(1)n_1 \text{ and } x_2 = 0(1)n_2\}$ into equivalence classes as follows. First, observe the distinct values of $\hat{\Delta}(\mathbf{x})$ include $-1, 0, +1$ and are symmetrically distributed about zero since $\hat{\Delta}(x_1, x_2) = -\hat{\Delta}(n_1 - x_1, n_2 - x_2)$. Thus let

$$-1 = d_{-K} < \ldots < d_0 = 0 < \ldots < d_K = +1$$

be the *distinct* values of $\hat{\Delta}(\mathbf{x}) = x_1/n_1 - x_2/n_2$ for $\mathbf{x} \in \mathcal{X}$. Define the equivalence class $\mathcal{X}_i$ by

$$\mathcal{X}_i = \{\mathbf{x} \in \mathcal{X} : \hat{\Delta}(\mathbf{x}) = d_i\}$$

for $i = -K(1)K$. For example, $\mathcal{X}_{-K} = \{(0, n_2)\}$ corresponds to $d_{-K} = -1$ and $\mathcal{X}_K = \{(n_1, 0)\}$ corresponds to +1. By construction, $\mathcal{X} = \mathcal{X}_{-K} \cup \ldots \cup \mathcal{X}_K$ and $\mathcal{X}_i \cap \mathcal{X}_j = \phi$ when $i \neq j$.

The algorithm fixes a partition

$$0 = \Delta_0 < \Delta_1 < \ldots < \Delta_m = 1 \tag{5}$$

of [0,1] into sufficiently many pieces to guarantee the desired accuracy; e.g., the mesh $0 < .005 < .015 < \ldots < .995$ of 100 points guarantees 2- place accuracy. For $0 \leq i \leq m$ it constructs a $100 \times (1-\alpha)\%$ acceptance region $A(\Delta_i)$ corresponding to H_0: $\Delta = \Delta_i$ versus H_A: $\Delta \neq \Delta_i$; acceptance regions for $-1 \leq \Delta < 0$ are formed by symmetry. Each $A(\Delta_i)$ is defined to have the form

$$B_s \cup \mathcal{X}_{s+1} \cup \ldots \cup \mathcal{X}_{t-1} \cup C_t \text{ for some } -K \leq s \leq t \leq K \tag{6}$$

where

$$\phi \neq B_s \subset \mathcal{X}_s;\ C_t \subset \mathcal{X}_t;\ \text{and, when } s < t, \mathcal{X}_t - C_t \neq \phi. \tag{7}$$

Thus the estimated $\hat{\Delta}(\mathbf{x})$ values corresponding to $\mathbf{x} \in A(\Delta_i)$ are the consecutive sequence $d_s, \ldots, d_t$ (assuming $C_t \neq \phi$). Furthermore, the region $A(\Delta_{i+1})$ is determined by moving $A(\Delta_i)$ to the"right" in the sense that the endpoints satisfy

$$\min\{\hat{\Delta}(\mathbf{x}) : \mathbf{x} \in A(\Delta_i)\} \leq \min\{\hat{\Delta}(\mathbf{x}) : \mathbf{x} \in A(\Delta_{i+1})\} \tag{8}$$

$$\max\{\hat{\Delta}(\mathbf{x}) : \mathbf{x} \in A(\Delta_i)\} \leq \max\{\hat{\Delta}(\mathbf{x}) : \mathbf{x} \in A(\Delta_{i+1})\}; \tag{9}$$

(8)-(9) is the analog of (3).

The algorithm refers to a generic method, "rule R", of choosing subsets of B_s and $\mathcal{X}_t - C_t$ in Steps 0.2, 2.1, and 3.1. Several choices are considered in Section 2.3 and the properties of the resulting intervals are studied in Section 3. Throughout the remainder of the paper let

$$L(E, \Delta) := \inf_{p_1 \in \mathcal{I}(\Delta)} P_{\Delta, p_1}\{\mathbf{X} \in E\}$$

for $E \subset \mathcal{X}$ and $-1 < \Delta < 1$; also let $A - B \equiv \{w \in A : w \notin B\}$ be the set difference of A and B.

2.2 Statement of the Algorithm

Step 0. [Initialization]

0.1 Partition [0, 1] by $0 \equiv \Delta_0 < \ldots < \Delta_m = 1$.

0.2 Determine $A(\Delta_0) = A(0)$ by rule R and set i = 1.

0.3 Go to Step 1.

Step 1. [Induction]

1.1 If i = m+1, then go to Step 4. Otherwise assume

$$A(\Delta_{i-1}) = B_s \cup \mathcal{X}_{s+1} \cup \ldots \cup \mathcal{X}_{t-1} \cup C_t$$

satisfies (6)-(7) where $s = s(i-1) \leq t = t(i-1)$, and $L(A(\Delta_{i-1}), \Delta_{i-1}) \geq 1 - \alpha$.

1.2 Set $(s(i), t(i), B_{s(i)}, C_{t(i)}) \equiv (s(i-1), t(i-1), B_{s(i-1)}, C_{t(i-1)})$ and $A(\Delta_i) \equiv A(\Delta_{i-1})$.

1.3 If $L(A(\Delta_i), \Delta_i) \geq 1 - \alpha$ then go to Step 2, otherwise go to Step 3.

Step 2. [Elimination]

2.1 Set $A_T \equiv A(\Delta_i) - B$ where $B \subset B_{s(i)}$ is defined by Rule R.

2.2 If $L(A_T, \Delta_i) \geq 1 - \alpha$ then go to Step 2 after setting $A(\Delta_i) := A_T$ and modifying s(i) and $B_{s(i)}$ to satisfy (6)- (7).

2.3 If $L(A_T, \Delta_i) < 1 - \alpha$ then set $i = i + 1$ and go to Step 1.

Step 3. [Addition]

3.1 Set $A(\Delta_i) \equiv A(\Delta_i) \cup C$ where $C \subset \mathcal{X}_{t(i-1)} - C_{t(i-1)}$ is defined by Rule R and modifying $t(i)$ and $C_{t(i)}$ to satisfy (6)-(7).

3.2 If $L(A(\Delta_i), \Delta_i) \geq 1 - \alpha$ then go to Step 2; otherwise go to Step 3.

Step 4. [Completion and Inversion]

4.1 For $1 \leq i \leq m$ set

$$\begin{cases} \Delta_{-i} & \equiv -\Delta_i \\ A(\Delta_{-i}) & := \{\mathbf{n} - \mathbf{x} \in \mathcal{X} : \mathbf{x} \in A(\Delta_i)\} \end{cases}$$

4.2 Set

$$\begin{cases} \underline{\Delta}(\mathbf{x}) & := \min\limits_{-m \leq i \leq m} \{\Delta_i : \mathbf{x} \in A(\Delta_i)\} \\ \overline{\Delta}(\mathbf{x}) & := \max\limits_{-m \leq i \leq m} \{\Delta_i : \mathbf{x} \in A(\Delta_i)\}. \end{cases}$$

Several aspects of the algorithm deserve comment. First, in the Crow- Blyth-Still spirit, it tries to form small acceptance regions and thereby short confidence intervals by deleting (adding) points from (to) A(Δ_i) with small (large) probabilities consistent with (8)-(9).

Second, the acceptance regions defined in Step 4 are level $(1 - \alpha)$. By construction, $L(A(\Delta_i), \Delta_i) \geq 1 - \alpha$ for $\Delta_i \in (0, 1)$. As defined in Step 4.1, for $\Delta = \Delta_i \in [-1, 0]$ and for any $p_1 \in \mathcal{I}(\Delta)$

$$\begin{aligned} P_{\Delta,p_1}\{A(\Delta)\} &= \sum_{\mathbf{x} \in A(\Delta)} P_{\Delta,p_1}\{\mathbf{X} = \mathbf{x}\} \\ &= \sum_{\mathbf{x} \in A(\Delta)} P_{-\Delta,1-p_1}\{\mathbf{X} = \mathbf{n} - \mathbf{x}\} \\ &= \sum_{\mathbf{y} \in A(-\Delta)} P_{-\Delta,1-p_1}\{\mathbf{X} = \mathbf{y}\} \\ &= P_{-\Delta,1-p_1}\{A(-\Delta)\} \\ &\geq 1 - \alpha \end{aligned}$$

The second equality holds by calculation, the third by definition of A($-\Delta$), and the inequality holds since $-\Delta > 0$.

Third, we claim that the algorithm must terminate; this requires a technical analysis which is given in the Appendix. It is proven by showing that in Step 3 there always exists a set C to augment $A(\Delta_i)$ so that $L(A(\Delta_i) \cup C, \Delta_i) \geq 1 - \alpha$ for $1 \leq i \leq m$.

Lastly, the acceptance regions $\{A(\Delta_i) : -m \leq i \leq m\}$ must satisfy (8) and (9). This follows by construction for $0 \leq i \leq m$ and by symmetry for $-m \leq i \leq -1$ since

$$\min\{\hat{\Delta}(\mathbf{x}) : \mathbf{x} \in A(\Delta_i)\} = -\max\{\hat{\Delta}(\mathbf{x}) : \mathbf{x} \in A(-\Delta_i)\}$$

2.3 Construction of A(0), Elimination and Addition Sets

A rule for constructing $A(0)$ and addition-deletion, the invariant rule, is proposed. A computationally simpler alternative, the naive rule, will be mentioned at the end of the subsection; however, the naive rule generates wider intervals.

Invariant Rule R_I

The idea of this rule is to delete points which decrease $L(A_T, \Delta_i)$ as little as possible and add points which increase $L(A_T, \Delta_i)$ as much as possible.

Case 1: $n_1 < n_2$

$A(0)$: Choose $A(0)$ containing $\mathcal{X}_0$ and of the form (6)-(7) satisfying: (a) $\mathbf{x} \in A(0) \Rightarrow \mathbf{n} - \mathbf{x} \in A(0)$, (b) $A(0)$ contains as few additional points as possible, and (c)

$$L(A(0), 0) \geq 1 - \alpha. \tag{10}$$

Elimination: Given $\Delta > 0$ and $A_T = B_s \cup \mathcal{X}_{s+1} \cup \ldots \cup \mathcal{X}_{t-1} \cup C_t$ of form (6)-(7) satisfying $L(A_T, \Delta) \geq 1 - \alpha$, let $B \equiv \{\mathbf{x}\}$ where $\mathbf{x} \in B_s$ satisfies

$$L(A_T - \{\mathbf{x}\}, \Delta) = \max_{\mathbf{y} \in B_s} L(A_T - \{\mathbf{y}\}, \Delta). \tag{11}$$

Addition: Given $\Delta > 0$ and $A_T = B_s \cup \mathcal{X}_{s+1} \cup \ldots \cup \mathcal{X}_{t-1} \cup C_t$ of form (6)-(7) satisfying $L(A_T, \Delta) < 1 - \alpha$, let $C \equiv \{\mathbf{x}\}$ where $\mathbf{x} \in \mathcal{X}_t - C_t$ satisfies

$$L(A_T \cup \{\mathbf{x}\}, \Delta) = \max_{\mathbf{y} \in \mathcal{X}_t - C_t} L(A_T \cup \{\mathbf{y}\}, \Delta). \tag{12}$$

Case 2: $n_1 = n_2$

$A(0)$: Choose $A(0)$ containing $\mathcal{X}_0$ and of the form (6)-(7) satisfying: (a) $\mathbf{x} \in A(0) \Rightarrow \{\mathbf{n} - \mathbf{x}, \pi\mathbf{x}, \mathbf{n} - \pi\mathbf{x}\} \in A(0)$, (b) $A(0)$ contains as few additional points as possible, and (c)

$$L(A(0), 0) \geq 1 - \alpha. \tag{13}$$

Elimination: Given $\Delta > 0$ and $A_T = B_s \cup \mathcal{X}_{s+1} \cup \ldots \cup \mathcal{X}_{t-1} \cup C_t$ of form (6)-(7) satisfying $L(A_T, \Delta) \geq 1 - \alpha$, let $B \equiv \{\mathbf{x}, \mathbf{n} - \pi\mathbf{x}\}$ where $\mathbf{x} \in B_s$ satisfies

$$L(A_T - \{\mathbf{x}, \mathbf{n} - \pi\mathbf{x}\}, \Delta) = \max_{\mathbf{y} \in B_s} L(A_T - \{\mathbf{y}, \mathbf{n} - \pi\mathbf{y}\}, \Delta). \tag{14}$$

Addition: Given $\Delta > 0$ and $A_T = B_s \cup \mathcal{X}_{s+1} \cup \ldots \cup \mathcal{X}_{t-1} \cup C_t$ of form (6)-(7) satisfying $L(A_T, \Delta) < 1 - \alpha$, let $C \equiv \{\mathbf{x}, \mathbf{n} - \pi\mathbf{x}\}$ where $\mathbf{x} \in \mathcal{X}_t - C_t$ satisfies

$$L(A_T \cup \{\mathbf{x}, \mathbf{n} - \pi\mathbf{x}\}, \Delta) = \max_{\mathbf{y} \in \mathcal{X}_t - C_t} L(A_T \cup \{\mathbf{y}, \mathbf{n} - \pi\mathbf{y}\}, \Delta). \tag{15}$$

Break ties in the selection of **x** using criteria (11)-(16) by choosing the value of **x** with the smallest lexicographic order in (x_1, x_2).

The intervals produced by Algorithm 2.2 using R_I will be called CI intervals. When $n_1 < n_2$, $A(0)$ is constructed by setting $A_T = \mathcal{X}_0$, initially. If (10) holds for the trial set A_T then set $A(0)$ equal to A_T; it is straightforward to verify that $A(0)$ satisfies (6)-(7) (take $s = 0$, $t = 1$, $B_s = \mathcal{X}_0$, and $C_t = \phi$) as well as (b)-(c). For example, to prove (b) we need only note that if $\mathbf{x} \in \mathcal{X}_0$ then $\mathbf{n} - \mathbf{x} \in \mathcal{X}_0$. If (10) fails for A_T then, in the next and in all subsequent stages, A_T satisfies (6)-(7) with $s = -t$ unless $C_t = \phi$. In this case add $\{\mathbf{x}, \mathbf{n} - \pi\mathbf{x}\}$ to A_T where $\mathbf{x} \in \mathcal{X}_t - C_t$ satisfies

$$L(A_T \cup \{\mathbf{x}, \mathbf{n} - \pi\mathbf{x}\}, 0) = \max_{\mathbf{y} \in \mathcal{X}_t - C_t} L(A_T \cup \{\mathbf{y}, \mathbf{n} - \pi\mathbf{y}\}, 0) \tag{16}$$

until (10) holds. Note that B_s is not empty in the elimination steps (11) and (14) nor is $\mathcal{X}_t - C_t$ empty in the addition steps (12), (15), and (16). Thus none of the maxima (11)-(16) is vacuous.

When $n_1 = n_2$, then $\pi\mathbf{n} = \mathbf{n}$ and $\mathbf{x}$, $\pi\mathbf{x}$, $\mathbf{n} - \mathbf{x}$ and $\mathbf{n} - \pi\mathbf{x}$ are all legitimate outcomes. The set $\{\mathbf{x}, \pi\mathbf{x}, \mathbf{n} - \mathbf{x}, \mathbf{n} - \pi\mathbf{x}\}$ consists of either 1, 2, or 4 distinct points. For example, if $n_1 = 10 = n_2$ and $(x_1, x_2) = (4, 6)$ then $\{\mathbf{x}, \pi\mathbf{x}, \mathbf{n} - \mathbf{x}, \mathbf{n} - \pi\mathbf{x}\} = \{(4, 6), (6, 4)\}$ consists of 2 points. The set $A(0)$ is constructed in a similar fashion to that when $n_1 < n_2$. All maxima are well-defined since B_s and $\mathcal{X}_t - C_t$ are finite and nonempty $(s < t)$. When $\mathbf{x} \in \mathcal{X}_0$, the entire set $\{\mathbf{x}, \pi\mathbf{x}, \mathbf{n} - \mathbf{x}, \mathbf{n} - \pi\mathbf{x}\} \subset \mathcal{X}_0$.

Unless the nominal coverage is very low, $A(0)$ will contain points in addition to those in $\mathcal{X}_0$. Also at each step the rule R_I myopically attempts to force the modified version of A_T to be small by choosing a large set B and a small set C subject to insuring certain consistency properties.

Now consider properties of the acceptance regions produced by R_I. By construction R_I guarantees that $A(0)$ satisfies

$$\mathbf{x} \in A(0) \Rightarrow \mathbf{n} - \mathbf{x} \in A(0) \tag{17}$$

for all $n_1 \leq n_2$. Furthermore Step 4.1 guarantees for $i = 1, \ldots, m$ that $A(\Delta_i)$ satisfies

$$\mathbf{x} \in A(\Delta_i) \Rightarrow \mathbf{n} - \mathbf{x} \in A(-\Delta_i) \tag{18}$$

When $n_1 = n_2$, R_I guarantees

$$\mathbf{x} \in A(\Delta_i; \mathbf{n}) \Leftrightarrow \mathbf{n} - \pi\mathbf{x} \in A(\Delta_i; \mathbf{n}) \tag{19}$$

holds for all $\mathbf{x}$ and all Δ_i. It will be shown below that (19) is key for invariance (4) to hold when $n_1 = n_2$. However if the R_I-Case 1 rule is applied when $n_1 = n_2$, then (19) need *not* hold. What is true is that the equality

$$P_{\boldsymbol{p}}\{\mathbf{x}|\mathbf{n}\} = P_{\pi\boldsymbol{p}}\{\pi\mathbf{x}|\pi\mathbf{n}\}$$

shows that $\pi\mathbf{x}$ must enter or leave the region for the problem with sample sizes $\pi\mathbf{n}$ according as $\mathbf{x}$ enters or leaves the region for the problem with sample sizes $\mathbf{n}$. Thus the set of points tied to enter or leave the acceptance region at each stage is consistent under problem relabeling. However (19) can fail when there are ties in the choice of points satisfying Equations (11)-(14) since the selection is made according to lexicographic order. In particular when $n_1 = n_2$, the equality

$$P_{\boldsymbol{p}}\{\mathbf{x}|\mathbf{n}\} = P_{1-\pi\mathbf{p}}\{\mathbf{n} - \pi\mathbf{x}|\mathbf{n}\}$$

shows that two points ($\mathbf{x}$ and $\mathbf{n} - \pi\mathbf{x}$) corresponding the *same* Δ will definitely be tied for entrance into, and deletion from the acceptance region. Thus the the modifications in Case 2 of the definition of R_I are required so that (19) holds.

One other method of choosing $A(0)$-addition-elimination rules will be mentioned. This rule has the virtues of computational simplicity and invariance under both relabeling of successes-failures and populations at the cost of wider intervals. The rule adds and deletes only entire equivalence sets $\mathcal{X}_j$. Thus every $A(\Delta_i)$ is a union of consecutive equivalence sets $\mathcal{X}_j$.

Naive Rule R_N

$A(0)$: Use the minimal set of the form $\mathcal{X}_{-r} \cup \ldots \cup \mathcal{X}_r$ which satisfies

$$L(A(0), 0) \geq 1 - \alpha.$$

Elimination/Addition: Delete and add entire equivalence sets $\mathcal{X}_j$ (i.e., set $B = \mathcal{X}_s$ and $C = \mathcal{X}_t$ in the algorithm).

Formally the initial set $A(0)$ has $s = -r, B_s = \mathcal{X}_{-r}, t = r + 1$, and $C_t = \phi$. $A(0)$ is constructed by setting $A(0) = \mathcal{X}_0$ if $L(\mathcal{X}_0, 0) \geq 1 - \alpha$ or, if not, then successively adding pairs $\mathcal{X}_{-r}$ and $\mathcal{X}_r$ to the previous trial set until (10) holds.

R_N is simple to implement. It is straightforward to check that (18) and (19) hold for R_N acceptance regions. It will be shown below that (18) and (19) guarantee the corresponding confidence intervals, called CN intervals hereafter, have several desirable invariance properties. However, CN intervals are obviously conservative compared to CI intervals and thus are not considered a practical alternative. The next section compares invariance properties of T intervals with those of CI intervals.

3 Invariance Properties

We first study the invariance of T and CI intervals under the relabeling of the outcomes success and failure. Intuitively, since the difference of the two success probabilities changes sign when success and failure are switched, one would hope the confidence interval would exhibit the same property.

Theorem 3.1 T and CI intervals satisfy

$$(\underline{\Delta}(\mathbf{x}), \overline{\Delta}(\mathbf{x})) = (-\overline{\Delta}(\mathbf{n} - \mathbf{x}), -\underline{\Delta}(\mathbf{n} - \mathbf{x})) \tag{20}$$

for all $\mathbf{x} \in \mathcal{X}$.

Proof: Santner and Snell (1980) prove T intervals satisfy (20). To see that CI intervals also satisfy (20), observe that

$$\begin{aligned}
\underline{\Delta}(\mathbf{x}) &= \min\left\{\min_{0\leq i\leq m}\{\Delta_i : \mathbf{x} \in A(\Delta_i)\}, \min_{-m\leq i\leq 0}\{\Delta_i : \mathbf{x} \in A(\Delta_i)\}\right\} \\
&= \min\left\{\min_{-m\leq i\leq 0}\{-\Delta_i : \mathbf{x} \in A(-\Delta_i)\}, \min_{0\leq i\leq m}\{-\Delta_i : \mathbf{x} \in A(-\Delta_i)\}\right\} \\
&= \min\left\{-\max_{-m\leq i\leq 0}\{\Delta_i : \mathbf{n} - \mathbf{x} \in A(\Delta_i)\}, -\max_{0\leq i\leq m}\{\Delta_i : \mathbf{n} - \mathbf{x} \in A(\Delta_i)\}\right\} \\
&= -\max\left\{\max_{-m\leq i\leq 0}\{\Delta_i : \mathbf{n} - \mathbf{x} \in A(\Delta_i)\}, \max_{0\leq i\leq m}\{\Delta_i : \mathbf{n} - \mathbf{x} \in A(\Delta_i)\}\right\} \\
&= -\overline{\Delta}(\mathbf{n} - \mathbf{x})
\end{aligned}$$

The first equality holds by definition of $\underline{\Delta}$, the second since $-\Delta_{-i} = \Delta_i$ for all i, the third from the fact that for all $\mathbf{x} \in \mathcal{X}$, $\mathbf{x} \in A(-\Delta)$ if and only if $\mathbf{n} - \mathbf{x} \in A(\Delta)$, and the last two equalities by algebra and the definition of $\overline{\Delta}(\mathbf{n} - \mathbf{x})$. The characterization of $\mathbf{x} \in A(-\Delta)$ holds for $\Delta = 0$ by construction and for $\Delta \in (-1, 0)$ by Step 4.1. The lower limit follows by a similar argument. □

Remark 3.1 Equation (20) also holds for CN intervals since their acceptance regions satisfy $\mathbf{x} \in A(\Delta)$ if and only if $\mathbf{n} - \mathbf{x} \in A(\Delta)$ for all $\mathbf{x}$ and Δ.

Both T and CI intervals are consistent under relabeling of populations.

Theorem 3.2 T and CI intervals satisfy

$$\left(\underline{\Delta}(\mathbf{x}; \mathbf{n}), \overline{\Delta}(\mathbf{x}; \mathbf{n})\right) = \left(-\overline{\Delta}(\pi\mathbf{x}; \pi\mathbf{n}), -\underline{\Delta}(\pi\mathbf{x}; \pi\mathbf{n})\right) \tag{21}$$

for all $\mathbf{x} \in \mathcal{X}$ and all $n_1 \leq n_2$.

Proof: The proof for T intervals is straightforward from their definition. Equation (21) holds for CI intervals when $n_1 < n_2$ by definition (4). For CI intervals when $n_1 = n_2$ it suffices to consider one set of endpoints in (21) since the other is proved analogously. We have

$$\begin{aligned}
-\overline{\Delta}(\pi\mathbf{x}; \pi\mathbf{n}) &= -\max\left\{\max_{0 \leq i \leq m}\left\{\Delta_i : \pi\mathbf{x} \in A(\Delta_i, \pi\mathbf{n})\right\}, \max_{-m \leq i \leq 0}\left\{\Delta_i : \pi\mathbf{x} \in A(\Delta_i, \pi\mathbf{n})\right\}\right\} \\
&= -\max\left\{-\min_{-m \leq i \leq 0}\left\{\Delta_i : \pi(\mathbf{n} - \mathbf{x}) \in A(\Delta_i, \pi\mathbf{n})\right\}, -\min_{0 \leq i \leq m}\left\{\Delta_i : \pi(\mathbf{n} - \mathbf{x}) \in A(\Delta_i, \pi\mathbf{n})\right\}\right\} \\
&= -\max\left\{-\min_{-m \leq i \leq 0}\left\{\Delta_i : \mathbf{x} \in A(\Delta_i, \pi\mathbf{n})\right\}, -\min_{0 \leq i \leq m}\left\{\Delta_i : \mathbf{x} \in A(\Delta_i, \pi\mathbf{n})\right\}\right\} \\
&= \min\left\{\min_{-m \leq i \leq 0}\left\{\Delta_i : \mathbf{x} \in A(\Delta_i, \pi\mathbf{n})\right\}, \min_{0 \leq i \leq m}\left\{\Delta_i : \mathbf{x} \in A(\Delta_i, \pi\mathbf{n})\right\}\right\} \\
&= \underline{\Delta}(\mathbf{x}; \mathbf{n})
\end{aligned}$$

where the first equality is by definition of $\overline{\Delta}$, the second is by Step 4.1, the third is by (19), the fourth is by algebra, and the last is by definition of $\underline{\Delta}$ and the fact that $\pi\mathbf{n} = \mathbf{n}$. □

Remark 3.2 CI intervals could be alternatively constructed for $n_1 > n_2$ in such a way that Theorem 3.2 held when $n_1 < n_2$. This result could be proved for $n_1 \neq n_2$ by the same

argument as in the case $n_1 = n_2$, provided that the analogue of (19), $\mathbf{x} \in A(\Delta_i; \mathbf{n}) \leftrightarrow \pi(\mathbf{n} - \mathbf{x}) \in A(\Delta_i; \pi\mathbf{n})$, were true. The latter is used to establish the third equality in the proof of Theorem 3.2. This characterization would hold, for example, if the acceptance regions for $n_1 > n_2$ were *defined* in terms of those for the problem with sample sizes reversed by $\mathbf{x} \in A(\Delta_i; \mathbf{n}) \leftrightarrow \pi(\mathbf{n} - \mathbf{x}) \in A(\Delta_i; \pi\mathbf{n})$.

Both T and CN intervals satisfy

$$(\underline{\Delta}(\mathbf{x}), \overline{\Delta}(\mathbf{x})) = (\underline{\Delta}(\mathbf{y}), \overline{\Delta}(\mathbf{y})) \tag{22}$$

whenever **x** and **y** have the same point estimate; i.e., $\hat{\Delta}(\mathbf{x}) = \hat{\Delta}(\mathbf{y})$. Intuitively, (22) is a measure of the "coarseness" of the intervals produced by both systems. The same interval is produced whenever **x** and **y** have the same point estimate. The reason is that each system is based on the same set of acceptance regions when $\hat{\Delta}(\mathbf{x}) = \hat{\Delta}(\mathbf{y})$. CN intervals add and subtract only entire equivalence classes to acceptance regions. T intervals have the same acceptance sets since **x** and **y** have the same tails. CI intervals need not satisfy (22).

When $n_1 = n = n_2$ and $x_1 = x = x_2$, $\hat{\Delta}(\mathbf{x}) = x/n - x/n = 0$ *and* both binomials are estimated with equal precision. These suggest the interval should be symmetric about zero in this case.

Theorem 3.3 If $n_1 = n_2 = n$ then T and CI intervals satisfy

$$\underline{\Delta}(x, x) = -\overline{\Delta}(x, x) \tag{23}$$

for all outcomes $0 \leq x \leq n$.

Proof: Tail intervals satisfy (23) since

$$\begin{aligned} \underline{\Delta}(x, x) &= \underline{\Delta}(n - x, n - x) \text{ by (22)} \\ &= -\overline{\Delta}(x, x). \end{aligned}$$

where the second equality holds by Theorem 3.1. The first equality also holds for CI intervals by (19) (or equivalently by (20) and (21)) while the second is true by Theorem 3.1. □

Remark 3.3 Equation (23) also holds for CN intervals by the same steps proving the result for T intervals where the first equality holds by (22) and the second by Remark 3.1.

4 Comparisons

First some comparisons of T and R_I intervals will be made and then some brief comparisons of R_I and Coe and Tamhane intervals will be stated.

4.1 Comparison of Tail and Invariant Intervals

This section compares the operating characteristics of the T and CI intervals. Both achieve at least their nominal level for any **n** and **p**, and Section 3.1 proved both possess intuitive invariance properties. For two specific problems, one balanced and one unbalanced, the achieved coverage probability and lengths (expected and outcome-by-outcome) will be compared. Then some general considerations will be discussed.

Figures 2 and 3 plot the *difference* of the achieved coverage probabilities

$$P_{\boldsymbol{p}}\{\Delta \in \text{T interval}\} - P_{\boldsymbol{p}}\{\Delta \in \text{CI interval}\}$$

for nominal 95% intervals for $(n_1, n_2) = (10, 10)$ and $(n_1, n_2) = (15, 5)$, respectively, for the grid $p_1 = 0(.05)1$ and $p_2 = 0(.05)1$. Both systems of intervals achieve at least 95% coverage for all (p_1, p_2); in principle, the difference can either positive or negative. The figures show that the T system of intervals are substantially more conservative than the CI system intervals throughout the parameter space. Comparison of the Figures 2 and 3 shows that the T intervals are more conservative than the CI intervals in the balanced case because the differences are greater in this case.

The expected length of any system of outcomes is a weighted average of the lengths of the intervals over all outcomes. A comparison of the lengths of the T and CI intervals show that the latter are *uniformly* shorter than the former over *all* outcomes **y** when (n_1, n_2) = (15, 5). When (n_1, n_2) = (10,10), CI intervals are shorter than the T intervals for all but 8 out of 121 (central) outcomes; the reversals occur for "central" outcomes in the 11×11 lattice space of outcomes. In general, CI intervals will be shorter than T intervals for most outcomes when the nominal confidence level is large because T intervals must add entire equivalence classes.

To compare the expected lengths of the intervals define $E_{\boldsymbol{p}}\{L(T)\}$ and $E_{\boldsymbol{p}}\{L(CI)\}$ to be the expected length of the T and CI intervals, respectively, when **p** is the true vector of

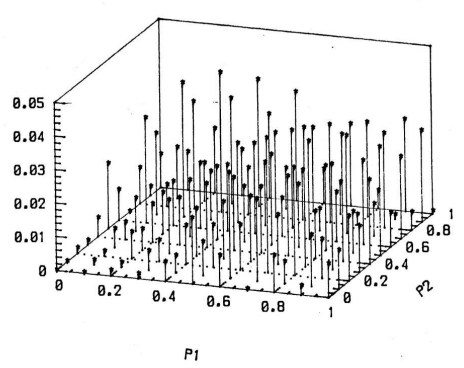

Figure 2: Plot of $(p_1, p_2, P_{\boldsymbol{p}}[\Delta \in \text{T interval}] - P_{\boldsymbol{p}}[\Delta \in \text{CI interval}])$ for nominal 95% Intervals when $n_1 = 10 = n_2$

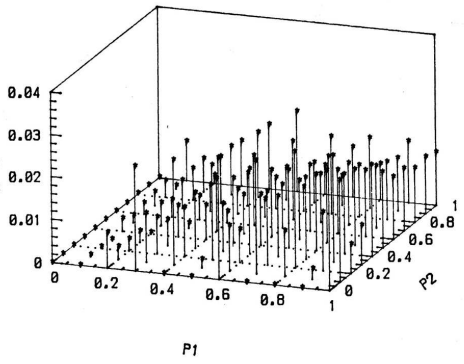

Figure 3: Plot of $(p_1, p_2, P_{\boldsymbol{p}}[\Delta \in \text{T interval}] - P_{\boldsymbol{p}}[\Delta \in \text{CI interval}])$ for nominal 95% Intervals when $(n_1, n_2) = (15, 5)$

probabilities. Figures 4 and 5 plot $E_{\boldsymbol{p}}\{L(T)\} - E_{\boldsymbol{p}}\{L(CI)\}$ versus (p_1, p_2) for the cases $(n_1, n_2) = (10, 10)$ and $(15, 5)$, respectively, at the same (p_1, p_2)) grid as Figures 2 and 3. In all cases CI intervals are superior to T intervals. In the balanced case CI intervals have the greatest improvement over T intervals at the *edges* of **p** space and the least improvement in the *center* of **p** space. The situation is qualitatively different for the unbalanced case with the least improvement at the *edges* and overall a more nearly constant difference in expected lengths.

The two cases discussed above are typical of those examined in detail by the authors. While both CI and T intervals have identical invariance properties, CI intervals are superior

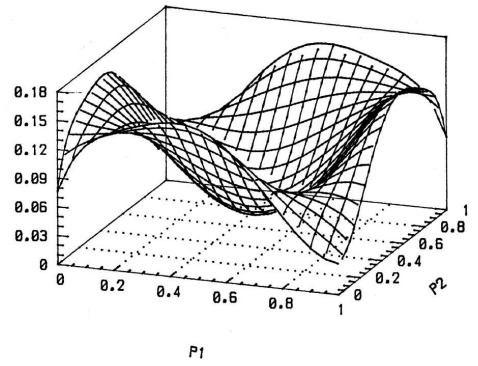

Figure 4: Plot of $(p_1, p_2, E_{\boldsymbol{p}}$ [Length of TM$] - E_{\boldsymbol{p}}$ [Length of CI]) for nominal 95% Intervals when $n_1 = 10 = n_2$

in terms of both coverage and length. However T intervals do have several attractive features. First, T intervals are computed for each $\mathbf{x} \in \mathcal{X}$ separately while CI intervals must essentially be simultaneously computed for all $\mathbf{x} \in \mathcal{X}$. Second, T intervals are simpler to compute than CI intervals since they require maximization of a polynomial in p_1 (for each Δ) which can be performed by determining the zeroes of the derivative polynomial. The latter is not a practical technique for constructing CI intervals. A third feature of T intervals not necessarily possessed by CI intervals is their monotonicity in α. Suppose $(\underline{\Delta}_i(\mathbf{x}), \overline{\Delta}_i(\mathbf{x}))$ is a $100 \times (1 - \alpha_i)\%$ interval at $\mathbf{x}$ for α_i for $i = 1, 2$ and that $\alpha_2 < \alpha_1$. Then one expects the intervals with large coverage should contain those with low coverage for all outcomes; i.e.,

$$\underline{\Delta}_2(\mathbf{x}) \leq \underline{\Delta}_1(x) \text{ and } \overline{\Delta}_1(\mathbf{x}) \leq \overline{\Delta}_2(\mathbf{x}) \text{ for all } \mathbf{x} \in \mathcal{X}. \tag{24}$$

Tail intervals satisfy (24) because of the monotonicity of L(E,Δ) for tail sets E. Table 1 is an example showing (24) need not hold for CI intervals.

The reason why CI intervals can violate (24) is that Algorithm 2.2 forms acceptance sets containing $\mathbf{x}$ with $\hat{\Delta}(\mathbf{x})$ as small as possible; it does so because it deletes points from B_s whenever possible, and adds points to C_t only when necessary. Geometrically the acceptance region tends to stay in the "northwest" portion of $\mathcal{X}$. At the expense of additional calculation the algorithm can be modified to make violations of (24) less likely. Some alternative proposals for acceptance regions will be discussed in Section 5.

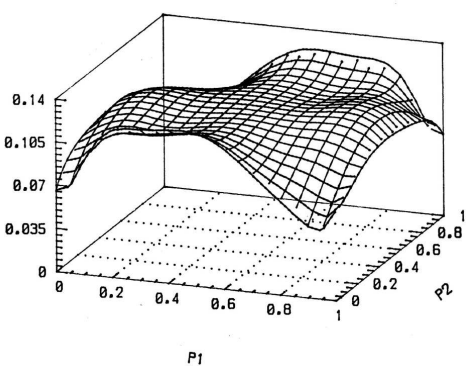

Figure 5: Plot of $(p_1, p_2, E_{\boldsymbol{p}}$ [Length of TM] $- E_{\boldsymbol{p}}$ [Length of CI]) for nominal 95% Intervals when $(n_1, n_2) = (15, 5)$

Table 1: Values of $\underline{\Delta}(0,0)$ and $\overline{\Delta}(0,0)$ for Nominal 95% CI Intervals When $(n_1, n_2) = (10, 5)$.

α	$\underline{\Delta}(0,0)$	$\overline{\Delta}(0,0)$
.05	-.545	.410
.10	-.426	.420

4.2 Comparison of R_I and Coe and Tamhane Intervals

Coe and Tamhane (1991) propose an algorithm for constructing a system of intervals (which we will denote CT intervals) for $p_1 - p_2$ satisfying the same invariance properties (stated in Theorem 3.1-3.3) as R_I intervals. While the algorithm for the R_I intervals begins with the acceptance set for the central value $\Delta = 0$ and modifies it for subsequent Δ_i values, the CT algorithm begins with the *smallest* Δ value in the selected grid of Δ_i values and modifies this acceptance for larger Δ values. The net result is that central outcomes (i.e., those consistent with estimated Δ values near zero) tend to occur more often in the acceptance sets for the R_I intervals while sample outcomes consistent with more extreme Δ outcomes tend to occur more often in the acceptance sets of the CT intervals. This effect is seen in the three examples (5×5,10×10 and 15×5) for which they report comparisons. CT intervals tend to be shorter than R_I intervals in the "central" part of the outcome space while R_I intervals tend to be shorted for outcomes in the "northwest" and "southeast" corners of the

outcome space. Coe and Tamhane report that their intervals have shorter averag length over all the outcomes than R_I intervals for these three examples.

On the other hand CT intervals require more time to compute than R_I intervals. Coe and Tamhane report that CT intervals require 4 times as much time as R_I intervals for the 5×5 problem and 8 times as much time for the 10×10 problem. They conjecture the difference will increase exponentially. Our conclusion is that CT intervals fall between the partition method intervals of Santner and Snell (1980) and the R_I intervals in computational complexity. The computational time of partition method intervals increases exponentially in $n_1 \times n_2$ and are not feasible for practical work.

5 Discussion and a Worked Example

First some comments are made regarding the use and running time of the FORTRAN subroutine for computing CI intervals available from the authors. Then an example is given and concluding remarks presented about possible modifications to the basic algorithm presented in Section 2.

5.1 Remarks about the FORTRAN Program

A FORTRAN subroutine available from the authors implements the Algorithm of Section 2 for (n_1, n_2) with $(n_1 + 1)(n_2 + 1) \leq 500$. The partition of the Δ axis (which determines the accuracy of the interval) must contain no more than 1000 points. The function $L(\cdot, \Delta)$ is approximated as the minimum value over $\mathcal{I}(\Delta)$; the code constrains this p_1 partition to contain no more than 1000 points. The program requires about 6 (23) seconds of Decstation 5000 CPU time to calculate the family of 95% CI intervals for all 121 (441) outcomes $\mathbf{x}$ when $n_1 = 10 = n_2$ ($n_1 = 20 = n_2$) and a partition of size 200 is used for both the p_1 and Δ axes.

5.2 An Example

Fisher (1935) studies the effect of genetics on criminal tendencies by recording the numbers of twins who are also criminals for 17 criminals who have dizygotic twins and 13 criminals who have monozygotic twins. The proportions of twins in the dizygotic and monozygotic

groups who are also criminals are 2/17 = .12 and 10/13 = .77, respectively, with an estimated difference of $\hat{\Delta}(2, 10)$= -.65 for the two groups. Table 2 lists the 95% T and CI intervals for the true Δ. The CI interval is interior to the T interval with the CI interval having length .51 and the T interval length .57.

Table 2: 95% T and CI Intervals for Δ Based on Fisher's Criminal Twin Data

Method	$\underline{\Delta}$	$\overline{\Delta}$
T	-.873	-.306
CI	-.848	-.335

5.3 Concluding Remarks

The algorithm presented in Section 2 does not perform substitution (checking for the existence of equal sized acceptance regions meeting the probability requirement and choosing among them based on a criteria such as balance of the 'tails'). Blyth and Still (1983) found that, for testing H_0: $p = p_0$, choosing among the family of equal sized acceptance regions the $A(p_0) = \{\ell(p_0), \ldots, u(p_0)\}$ which minimizes the tail difference

$$|P_{\boldsymbol{p_0}}\{Y \leq \ell - 1\} - P_{\boldsymbol{p_0}}\{Y \geq u + 1\}|$$

produced the system of intervals which most nearly had the additional intuitive properties that they were increasing in α, and had endpoints increasing in $\hat{p}$. Intuitively, the effect of not implementing substitution is that the interval endpoints change more slowly than if substitution is allowed.

Thus several analogs of substitution by 'tail balancing' were implemented for the acceptance regions $A(\Delta_i)$ before recommending the algorithm of Section 2. For example, if $A(\Delta_i) = B_s \cup \mathcal{X}_{s+1} \cup \ldots \cup \mathcal{X}_{t-1} \cup C_t$ with $-K \leq s \leq t \leq K$ is constructed by the algorithm in Section 2, then we considered alternative acceptance regions obtained by substituting **y** in the complement of C_t for **x** in B_s and choosing as $A(\Delta_i)$ the region which minimizes

$$\max\{|P_{\Delta,p_1}\{\text{lower tail of } A(\Delta_i)\} - P_{\Delta,p_1}\{\text{upper of } A(\Delta_i)\}| : p_1 \in \mathcal{I}(\Delta_i)\}\,.$$

However empirical comparisons of the systems of intervals produced by the Section 2 algorithm with these modifications showed none of the variants produced intervals with markedly superior monotonicity properties in α or $\hat{\Delta}$. Of course, all required greater computational time. In part, the difficulty is caused by the presence of the nuisance parameter p_1.

Santner and Snell (1980) show how tail intervals can be constructed for the relative risk $\rho = p_1/p_2$. The results of this paper suggest Crow ρ intervals can also be expected to be superior to the ρ tail intervals. One problem in constructing such intervals is that ρ ranges over the infinite interval $(0, \infty)$ rather than a finite one. However Crow type algorithms can directly be constructed for a monotone function of ρ which has bounded range such as

$$\tan^{-1}(\rho) : (0, \infty) \to (0, \pi/2), \text{ or}$$

$$\tanh^{-1}(\rho) : (0, \infty) \to (0, 1), \text{ or}$$

$$\frac{\rho}{1+\rho} : (0, \infty) \to (0, 1).$$

Finally, we note that there are circumstances in which it is reasonable to consider intervals satisfying a modification of the basic coverage requirement (1). One obvious case is when there is prior information about the location of p_1 and p_2 which can be written in the form of a density for (p_1, p_2) then it is straightforward to determine the corresponding prior density $h(\Delta, p_1)$. In this case it is reasonable to require confidence intervals to satisfy

$$\int_{-1}^{+1} \int_{\mathcal{I}(\Delta)} P_{\Delta,p_1}\{\underline{\Delta}(\mathbf{x}) < \Delta < \overline{\Delta}(\mathbf{x})\} h(\Delta, p_1) d\Delta dp_1 \geq 1 - \alpha.$$

Even in the absence of prior (p_1, p_2) information it might still be reasonable to weaken (1) by only requiring

$$\int_{-1}^{+1} \int_{\mathcal{I}(\Delta)} P_{\Delta,p_1}\{\underline{\Delta}(\mathbf{x}) < \Delta < \overline{\Delta}(\mathbf{x})\} d\Delta dp_1 \geq 1 - \alpha.$$

Fujino and Okuno (1984) develop systems satisfying the latter requirement for the one-sample binomial problem.

APPENDIX

We prove that, in Step 3.1, there exists a set C so that $L(A_T \cup C,\ \Delta_{i+1}) \geq 1-\alpha$. Suppose $A = A(\Delta_i)$ satisfies $L(A,\ \Delta_i) \geq 1-\alpha$; the set of points $\mathbf{x}$ *not* in A is of the form $B \cup C$ where all $\mathbf{x} \in B$ satisfy $\hat{\Delta}(\mathbf{x}) \leq \min\left\{\hat{\Delta}(\mathbf{y}) : \mathbf{y} \in A\right\}$ and all $\mathbf{x} \in C$ satisfy $\hat{\Delta}(\mathbf{x}) \geq \max\left\{\hat{\Delta}(\mathbf{y}) : \mathbf{y} \in A\right\}$ The points in B have been previously eliminated while all the points in C are candidates to be added. It suffices to show B must have small probability mass under any (Δ, p_1) with $\Delta > \Delta_i$; i.e.,

$$S(B, \Delta) \leq \alpha \text{ for all } \Delta > \Delta_i \tag{25}$$

where $S(E, \Delta) = \sup_{p_1 \in I(\Delta)} P[\mathbf{X} \in E | \Delta, p_1]$. Since $S(B, \Delta_i) < \alpha$, it suffices to prove that $S(B, \Delta)$ is nonincreasing in Δ.

Theorem 5.1 For any $\Delta_1 < \Delta_2$

$$S(B, \Delta_1) \geq S(B, \Delta_2)$$

Proof: It suffices to show

$$P_{p_1,p_2}\{B\} \geq P_{p_1+\delta,p_2}\{B\} \tag{26}$$

and

$$P_{p_1,p_2}\{B|p_1,p_2\} \geq P_{p_1,p_2-\delta}\{B\} \tag{27}$$

for any $\delta > 0$ provided that p_1, p_2, $p_1 + \delta$, and $p_2 + \delta$ are all in [0, 1]. The proof of (27) is similar to that of (26) and hence only (26) will be given. For $i = 0, \ldots, n_2$ define $B_i = \{(x, i) \in B\}$. If $(x,\ i) \in B$ then $(y,\ i) \in B_i$ for every $y \in \{0, \ldots, x\}$; i.e., each B_i is a vertical lattice of consecutive points beginning at the horizontal axis. It is well known that for each x, $P_{p_1}\{X_1 \in \{0, \ldots, x\}\}$ is nonincreasing in p_1. Thus, if $x_i = \max\{x : (x,\ i) \in B\}$, then

$$P_{p_1,p_2}\{B_i\} = P_{p_1}\{X_1 \in \{0, \ldots, x_i\}\} \times P_{p_2}\{X_2 = i\}$$

is also nonincreasing in p_1. It follows that

$$P_{\boldsymbol{p}}\{B|p_1,p_2\} = \sum_{i=0}^{n_2} P_{\boldsymbol{p}}\{B_i\}$$

is nonincreasing in p_1 and therefore (26) holds. □

ACKNOWLEDGMENTS

The first author's research was partially supported by the U.S. Army Research Office through the Mathematical Science Institute of Cornell University and by the Cornell University Biomechanical Engineering program. The second author's research was supported by the Tokyo Gas Company Ltd. of Tokyo, Japan.

BIBLIOGRAPHY

Costello, P.S., and Wolfe, D.A. (1985), "A new nonparametric approach to the problem of agreement between two groups of judges," *Communications in Statistics - Simulation and Computation,* **B 14(4),** 791-805.

Beal, S.L. (1987). Asymptotic confidence intervals for the difference between two binomial parameters for use with small samples. *Biometrics* **43**, 941-950.

Blyth, C. and Still, H. (1983). Binomial confidence intervals. *J. Am. Statist. Assoc.* **76**, 108-116.

Clopper, C.J. and Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**, 404-13.

Coe, P.R. and Tamhane, A.C. (1991). Small sample confidence intervals for the difference, ratio, and odds ratio of two success probabilities. preprint.

Cornfield, J. (1956). A statistical problem arising from retrospective studies in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **(Vol. IV.)**, ed. Jerzey Newman, Berkeley: University of California Press, 135-148.

Crow, E.L. (1956). Confidence intervals for a proportion. *Biometrika* **43**, 423-435.

Duffy, D.E. and Santner, T.J. (1987). Confidence intervals for a binomial parameter based on multistage tests. *Biometrics* **43**, 81-93.

Fisher, R.A. (1935). The logic of inductive inference. *J. Roy. Statist. Soc.* **98**, 39-54.

Fujino, Y. and Okuno, T. (1984). The minimax average confidence intervals for a binomial probability: one-sided case. *Rep. Statist. Appl. Res. JUSE* 31, 1-7.

Gart, J.J. (1971). The comparison of proportions. A review of significance tests, confidence intervals and adjustments for stratification. *Review of the International Statistical Institute* **39**, 148-169. Addenda and errata (1972) 40, 221.

Katz, D., Baptista, J., Azen, S.P., and Pike, M.C. (1978). Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics* **34**, 469-474.

Mee, R.W. (1984). Confidence bounds for the difference between two probabilities (letter). *Biometrics* **40**, 1175-76.

Santner, T.J., and Snell, M.K. (1980). Small-sample confidence intervals for $p_1 - p_2$ and p_1/p_2 in 2x2 contingency tables. *J. Am. Statist. Assoc.* **75**, 386-394.

Sterne (1954). Some remarks on confidence or fiducial limits. *Biometrics* **40**, 961-971.