COMMUN. STATIST.—SIMULA., 22(4), 925-938 (1993)

SMALL SAMPLE CONFIDENCE INTERVALS FOR THE DIFFERENCE, RATIO AND ODDS RATIO OF TWO SUCCESS PROBABILITIES

Paul R. Coe
Department of Mathematics
Rosary College
River Forest, IL 60305

Ajit C. Tamhane
Department of Statistics
Northwestern University
Evanston, IL 60208

Keywords and Phrases: Sterne method; binomial distribution; $p_1 - p_2$; exact confidence intervals; Bernoulli treatments

ABSTRACT

This paper discusses the problem of constructing small sample confidence intervals for the difference of success probabilities of two independent Bernoulli distributions. An algorithm is given based on an extension of Sterne's (1954) method for constructing small sample confidence intervals for a single success probability. These confidence intervals have several invariance and other desirable properties such as short lengths and monotonicity. A comparison is made with an algorithm due to Santner and Yamagami (1993) which is also based on an extension of Sterne's method. Our algorithm is found to yield shorter intervals for a majority of outcomes, and these outcomes are located in the central portion of the sample space. Santner and Yamagami's algorithm gives shorter intervals for outcomes in the northwest and southeast corners of the sample space (corresponding to large differences in the observed sample proportion of successes), and is computationally faster. Modifications of the algorithm for obtaining confidence intervals for the ratio and odds ratio are indicated.

1. INTRODUCTION

A common problem arising especially in biostatistical applications is how to compare two "success" probabilities, p_1 and p_2, based on the observed values of two independent binomial random variables, $X_1 \sim B(n_1, p_1)$ and $X_2 \sim B(n_2, p_2)$. The common functions used for comparing p_1 and p_2 are: Berkson's simple difference $\Delta = p_1 - p_2$, the relative risk $\rho = p_1/p_2$ and the odds ratio $\psi = p_1(1-p_2)/p_2(1-p_1)$. In this article we present a new method (referred to here as the CT method) for constructing small sample confidence intervals (C.I.'s) for Δ (Section 2.3). We also show how this method can be modified to obtain small sample C.I.'s for ρ and ψ (Section 4). For reasons of space, only the method for Δ is discussed in detail.

Our method can be thought of as an extension of Sterne's (1954) method (as modified by Blyth and Still (1983)) for constructing $(1-\alpha)$-level small sample C.I.'s for a single binomial "success" probability p. Sterne's method constructs short C.I.'s for p by first constructing, subject to certain constraints, smallest possible acceptance regions having probability contents at least $1-\alpha$ for given values of p, and then inverting these acceptance regions to obtain the desired C.I.'s. The extension to the present problem is complicated because of two reasons: (*i*) The sample space is two-dimensional, and hence no simple linear order exists among the sample outcomes. (*ii*) There are two parameters, namely Δ, the parameter of interest, and p_1, a nuisance parameter. Therefore, when finding the smallest possible acceptance region for a given value of Δ in the Sterne-spirit (with probability content at least $1-\alpha$), it is necessary to consider the infimum of the probability content over the range of possible values of p_1. Nonetheless, as in the single p problem, the method yields suitably short C.I.'s for Δ. The algorithm for the method is readily programmable and computationally feasible.

Other methods for constructing small sample C.I.'s for Δ have been developed by Thomas and Gart (1977), Santner and Snell (1980), and more recently by Santner and Yamagami (1993). After pointing out the liberal nature of the Thomas and Gart method, Santner and Snell proposed three new methods for constructing C.I.'s for Δ. Their first method (called the *conditional method* or CM) involves obtaining C.I.'s for Δ (or ρ) from the conditional C.I.'s for ψ derived by Cornfield (1956). Their second method (called the *tail method* or TM) is an extension of Clopper and Pearson's (1934) tail intervals for a single p. Their third method (called the *partition method* or PM) is an extension of the Sterne method for a single p. Of these three, only TM is a viable method; it is computationally feasible and gives C.I.'s having the desired invariance properties (discussed in the sequel), but is somewhat conservative for the same reason that the Clopper and Pearson method is conservative for the single p problem. CM is computationally easiest of the three, but the resulting C.I.'s are far too wide; e.g., CM yields C.I.'s for Δ (ρ) that always include 0 (1) regardless of the

data or the confidence level. Thus the null hypothesis $H_0 : p_1 = p_2$ will always be accepted if one uses these C.I.'s for testing purposes. PM does not appear to be computationally feasible nor programmable because of its complexity; the only example Santner and Snell give for this method is for the case $n_1 = n_2 = 2$, which they solve by hand.

Santner and Yamagami (1993) have given a computationally feasible extension of the Sterne method (as modified by Crow (1956)) for which they have written a FORTRAN program. Their method (referred to here as the SY method) is described briefly in Section 3. Both our proposed CT method and the SY method improve upon TM. We shall explain in Section 3 how the SY intervals are short for "extreme" values of Δ (close to ± 1) but long for "middle" values of Δ (close to 0) for the same reason (as pointed out by Blyth and Still) that the Crow intervals for p are short when p is close to 0 or 1, and are long when p close to 1/2. This has the consequence that the maximum as well as the average interval lengths are increased for the SY intervals, as they are for the Crow intervals. Blyth and Still have pointed out another drawback of the Crow method, namely that it produces rather irregular C.I.'s for p (C.I.'s whose endpoints do not increase in a regular fashion with x for fixed n, and do not necessarily decrease with n for fixed x where x is the observed value of $X \sim B(n,p)$). Similar drawbacks may be expected with the SY method. However, the SY method is found to be computationally much faster than the CT method. A comparison between the CT method and the SY method is given in Section 3.

We remark that our method can be readily embedded into a group sequential scheme to obtain small sample *repeated* C.I.'s (Jennison and Turnbull 1984, 1990; Lai 1984) for the measure of interest (Δ, ρ or ψ). This problem is considered in Coe and Tamhane (1993).

2. CONFIDENCE INTERVALS FOR Δ

2.1 Preliminaries

Let $\mathcal{X}$ denote the sample space of $\boldsymbol{X} = (X_1, X_2)$,

$$\mathcal{X} = \{\boldsymbol{x} = (x_1, x_2) : 0 \le x_1 \le n_1, 0 \le x_2 \le n_2\},$$

and let $f(\boldsymbol{x})$ denote the joint probability mass function of $\boldsymbol{X}$,

$$f(\boldsymbol{x}) = f(\boldsymbol{x}|p_1, p_2) = \prod_{i=1}^{2} \binom{n_i}{x_i} p_i^{x_i}(1-p_i)^{n_i - x_i} \quad \text{for } \boldsymbol{x} \in \mathcal{X}. \tag{2.1}$$

For fixed Δ, p_1 lies in the interval $I(\Delta)$ where $I(\Delta) = [0, 1+\Delta]$ for $-1 \le \Delta \le 0$ and $I(\Delta) = [\Delta, 1]$ for $0 \le \Delta \le 1$. We will denote by $P\{\boldsymbol{X} \in \mathcal{E}|p_1, \Delta\}$, the probability of event $\mathcal{E} \subseteq \mathcal{X}$ computed under $f(\boldsymbol{x})$ with "success" probabilities p_1 and $p_2 = p_1 - \Delta$.

In order that a rule, which assigns a set $C_{\boldsymbol{n}}(\boldsymbol{x})$ as a confidence set for Δ when outcome $\boldsymbol{x} \in \mathcal{X}$ is observed using sample sizes $\boldsymbol{n} = (n_1, n_2)$, be a $(1-\alpha)$-level C.I. rule, it must have the following property:

(*i*) For every $\boldsymbol{x} \in \mathcal{X}$, $C(\boldsymbol{x}) = C_{\boldsymbol{n}}(\boldsymbol{x})$ should be interval-valued, i.e., $C_{\boldsymbol{n}}(\boldsymbol{x}) = [L_{\boldsymbol{n}}(\boldsymbol{x}), U_{\boldsymbol{n}}(\boldsymbol{x})]$ for some real numbers $-1 \le L(\boldsymbol{x}) = L_{\boldsymbol{n}}(\boldsymbol{x}) \le U(\boldsymbol{x}) = U_{\boldsymbol{n}}(\boldsymbol{x}) \le 1$ (when it causes no confusion, we shall suppress the dependence on $\boldsymbol{n}$ of the confidence limits and the other related quantities, as is done here), and

$$P\{L(\boldsymbol{X}) \le \Delta \le U(\boldsymbol{X}) \mid p_1, \Delta\} \ge 1-\alpha \ \ \forall\, \Delta \text{ and } p_1.$$

In addition, we would like our C.I.'s to possess the following desirable properties:

(*ii*) The C.I.'s should be invariant with respect to the labelling of the populations. This implies that if the population labels are interchanged so that the new outcome is $\pi\boldsymbol{x} = (x_2, x_1)$ and the new sample sizes are $\pi\boldsymbol{n} = (n_2, n_1)$ (where π is the permutation operator) then the C.I. for the new $\Delta = p_2 - p_1$ is given by

$$[L_{\pi\boldsymbol{n}}(\pi\boldsymbol{x}), U_{\pi\boldsymbol{n}}(\pi\boldsymbol{x})] = [-U_{\boldsymbol{n}}(\boldsymbol{x}), -L_{\boldsymbol{n}}(\boldsymbol{x})]. \tag{2.2}$$

(*iii*) The C.I.'s should be invariant with respect to the labelling of "successes" and "failures." This implies that under the transformation $p_i \to 1-p_i$ $(i = 1, 2)$, the C.I. for the new $\Delta = p_2 - p_1$ is given by

$$[L_{\boldsymbol{n}}(\boldsymbol{n}-\boldsymbol{x}), U_{\boldsymbol{n}}(\boldsymbol{n}-\boldsymbol{x})] = [-U_{\boldsymbol{n}}(\boldsymbol{x}), -L_{\boldsymbol{n}}(\boldsymbol{x})]. \tag{2.3}$$

(*iv*) For fixed $\boldsymbol{n}$, the C.I.'s should be monotone in $\boldsymbol{x}$ as follows: For fixed x_2, both $L(\boldsymbol{x})$ and $U(\boldsymbol{x})$ should be non-decreasing in x_1. Similarly for fixed x_1, both $L(\boldsymbol{x})$ and $U(\boldsymbol{x})$ should be non-increasing in x_2.

(*v*) Finally, for fixed $\boldsymbol{n}$, we would like our C.I.'s to be short in some overall sense, e.g., the average interval length, $\sum_{\boldsymbol{x}\in\mathcal{X}}\{U(\boldsymbol{x}) - L(\boldsymbol{x})\}/\text{card}(\mathcal{X})$ should be as small as possible.

2.2 Considerations in the Development of the Algorithm

To obtain a $(1-\alpha)$-level C.I. for Δ, we first construct $(1-\alpha)$-level acceptance regions $\mathcal{A}_i$ for different values of $\Delta = \Delta_i$ (i.e., acceptance regions for α-level two-sided tests of the hypotheses $H_{0i} : \Delta = \Delta_i$) for $i = 1, 2, \ldots, M$; here $-1 < \Delta_1 < \Delta_2 < \cdots < \Delta_M < +1$ are prespecified. By inverting these acceptance regions in the usual manner (Lehmann 1986, p. 90)), a C.I. for Δ can be obtained.

We now explain what properties the acceptance regions $\mathcal{A}_i$ must have in order that the resulting confidence sets $C(\boldsymbol{x})$ for Δ have the properties stated above.

(*i*) Consider Property (*i*). In order that $C(\boldsymbol{x})$ be interval-valued, each outcome $\boldsymbol{x}$ must be contained only in a consecutive set of acceptance regions $\mathcal{A}_i$. In order that the resulting

confidence intervals $[L(\boldsymbol{x}), U(\boldsymbol{x})]$ have a confidence level $(1-\alpha)$, the $\mathcal{A}_i$ must satisfy

$$P_0(\mathcal{A}_i, \Delta_i) \equiv \inf_{p_1 \in I(\Delta_i)} P\{\boldsymbol{X} \in \mathcal{A}_i | p_1, \Delta_i\} \ge 1-\alpha \quad (1 \le i \le M). \tag{2.4}$$

(*ii*) The invariance property (Property (*ii*)) with respect to the labelling of the populations can be guaranteed as follows: First consider the case $n_1 = n_2 = n$ (say). Then, using the fact that $\pi\boldsymbol{n} = \boldsymbol{n}$ in this case, from (2.2) and (2.3) it is easy to see that we need

$$[L(\boldsymbol{x}), U(\boldsymbol{x})] = [L(\boldsymbol{n}-\pi\boldsymbol{x}), U(\boldsymbol{n}-\pi\boldsymbol{x})] \ \ \forall\ \boldsymbol{x} \in \mathcal{X}.$$

Hence the acceptance regions $\mathcal{A}_i$ that produce these C.I.'s must satisfy

$$\boldsymbol{x} \in \mathcal{A}_i \Longleftrightarrow \boldsymbol{n}-\pi\boldsymbol{x} = (n-x_2, n-x_1) \in \mathcal{A}_i \ \ \forall\ \mathcal{A}_i,$$

i.e., each $\mathcal{A}_i$ must be symmetric about the line $x_1 + x_2 = n$. This can be guaranteed by adding or deleting the sample points (x_1, x_2) and $(n-x_2, n-x_1)$ in pairs when constructing each $\mathcal{A}_i$. It does not seem possible to give a similar simple condition on the acceptance regions $\mathcal{A}_i$ when $n_1 \ne n_2$. However, the desired invariance can still be guaranteed by using a given method to find the C.I.'s for $n_1 < n_2$, and applying (2.2) to find the corresponding C.I.'s for $n_1 > n_2$.

(*iii*) The invariance property concerning the labelling of the "successes" and "failures" (Property (*iii*)) can be satisfied as follows: Having constructed a $(1-\alpha)$-level acceptance region $\mathcal{A}_i$ for $\Delta = \Delta_i$, use the acceptance region $\mathcal{A}_i'$ for $\Delta = -\Delta_i$ given by

$$\mathcal{A}_i' = \{\boldsymbol{x} \in \mathcal{X} : \boldsymbol{n}-\boldsymbol{x} \in \mathcal{A}_i\}.$$

The fact that $\mathcal{A}_i'$ also has probability content at least $1-\alpha$ follows from the identity

$$f(\boldsymbol{x}|p_1, p_2) = f(\boldsymbol{n}-\boldsymbol{x}|1-p_1, 1-p_2) \ \ \forall\, \boldsymbol{x} \in \mathcal{X},$$

which implies that $P_0(\mathcal{A}_i', -\Delta_i) = P_0(\mathcal{A}_i, \Delta_i) \ge 1-\alpha$.

(*iv*) In order that the C.I.'s possess the monotonicity property (Property (*iv*)), each $\mathcal{A}_i$ must contain no "holes" either in the x_1-direction for any fixed x_2 or in the x_2-direction for any fixed x_1. Because if there is a "hole," say, $(x_1-1, x_2) \in \mathcal{A}_i, (x_1+1, x_2) \in \mathcal{A}_i$ but $(x_1, x_2) \notin \mathcal{A}_i$, then $\Delta_i \in [L(x_1-1, x_2), U(x_1-1, x_2)]$, $\Delta_i \in [L(x_1+1, x_2), U(x_1+1, x_2)]$ but $\Delta_i \notin [L(x_1, x_2), U(x_1, x_2)]$, thus violating the monotonicity property.

(*v*) The narrowness property (Property (*v*)) is difficult to ensure formally because it involves solving a highly complicated discrete optimization problem. However, this property can be achieved at least approximately by constructing the $\mathcal{A}_i$ containing as few points as possible.

2.3 Algorithm

Step 0 (Initialization)

0.1 Partition the Δ-space, i.e., the interval $[-1,1]$, into a grid $-1 \le \Delta_{-M} < \Delta_{-M+1} < \cdots < \Delta_0 < \Delta_1 < \cdots < \Delta_M \le +1$ where $\Delta_{-i} = -\Delta_i$ for $i = 1,\ldots,M$ and $\Delta_0 = 0$.

0.2 Set $i = 1$. Go to Step 1.

Step 1 (Addition)

1.1 Partition the p_1-space, i.e., the interval $[\Delta_i, 1]$, into a grid $\Delta_i \le p_{i1} < p_{i2} < \cdots < p_{iN_i} \le 1$ symmetrically around the midpoint $(1+\Delta_i)/2$.

1.2 For $j = 1,\ldots,N_i$, construct $\mathcal{A}_{ij}$ such that

$$f(\boldsymbol{x}|p_1 = p_{ij}, p_2 = p_{ij} - \Delta_i) \ge f(\boldsymbol{x}'|p_1 = p_{ij}, p_2 = p_{ij} - \Delta_i) \quad \forall\, \boldsymbol{x} \in \mathcal{A}_{ij}, \boldsymbol{x}' \notin \mathcal{A}_{ij}$$

and

$$\begin{aligned} \sum_{\boldsymbol{x}\in\mathcal{A}_{ij}} f(\boldsymbol{x}|p_1 = p_{ij}, p_2 = p_{ij} - \Delta_i) &= P\{\boldsymbol{X} \in \mathcal{A}_{ij}|p_1 = p_{ij}, \Delta = \Delta_i\} \\ &\ge 1 - \alpha. \end{aligned}$$

1.3 Set $\mathcal{A}_i = \cup_{j=1}^{N_i} \mathcal{A}_{ij}$.

1.4 Augment $\mathcal{A}_i$ as necessary with additional sample points $\boldsymbol{x}$ to ensure that $\mathcal{A}_i$ has no "holes" either in the x_1-direction or in the x_2-direction.

1.5 Go to Step 2.

Step 2 (Elimination)

2.1 Let $\hat{\Delta}(\boldsymbol{x}) = x_1/n_1 - x_2/n_2$. Eliminate any $\boldsymbol{x}^* \in \mathcal{A}_i$ such that

$$\hat{\Delta}(\boldsymbol{x}^*) \le \min_{\boldsymbol{x}\in\mathcal{A}_{i-1}} \hat{\Delta}(\boldsymbol{x}) \text{ and } \boldsymbol{x}^* \notin \mathcal{A}_{i-1}.$$

2.2 Let $\mathcal{D} = \{\boldsymbol{x}|\boldsymbol{x} \in \mathcal{A}_i \text{ and } P_0(\mathcal{A}_i - \{\boldsymbol{x}\}, \Delta_i) \ge 1-\alpha\}$. (If $n_1 = n_2$ then let $\mathcal{D} = \{\{\boldsymbol{x}, \boldsymbol{n} - \pi\boldsymbol{x}\}|\{\boldsymbol{x}, \boldsymbol{n} - \pi\boldsymbol{x}\} \in \mathcal{A}_i \text{ and } P_0(\mathcal{A}_i - \{\boldsymbol{x}, \boldsymbol{n} - \pi\boldsymbol{x}\}, \Delta_i) \ge 1-\alpha\}$. We refer to $\mathcal{D}$ as the set of "dispensable" outcomes in the sense that they can be removed from $\mathcal{A}_i$ without violating equation (2.4). If $\mathcal{D} = \phi$, go to Step 2.5.

2.3 Find $\boldsymbol{x}^* \in \mathcal{D}$ such that

$$P_0(\mathcal{A}_i - \{\boldsymbol{x}^*\}, \Delta_i) = \max_{\boldsymbol{x}\in\mathcal{D}} P_0(\mathcal{A}_i - \{\boldsymbol{x}\}, \Delta_i),$$

and set $\mathcal{A}_i = \mathcal{A}_i - \{\boldsymbol{x}^*\}$. (If $n_1 = n_2$ then replace $\boldsymbol{x}$ and $\boldsymbol{x}^*$ in the above by $\{\boldsymbol{x}, \boldsymbol{n} - \pi\boldsymbol{x}\}$ and $\{\boldsymbol{x}^*, \boldsymbol{n} - \pi\boldsymbol{x}^*\}$, respectively.)

2.4 Go to Step 2.2.

2.5 Set $i = i + 1$. If $i \le M$, go to Step 1; otherwise go to Step 3.

Step 3 (Completion and Inversion)

3.1 Set the acceptance region

$$\mathcal{A}_{-i} = \{\boldsymbol{n} - \boldsymbol{x} \in \mathcal{X} : \boldsymbol{x} \in \mathcal{A}_i\} \text{ for } \Delta_{-i} = -\Delta_i, i = 1,2,\ldots,M. \tag{2.5}$$

3.2 Set

$$L(\boldsymbol{x}) = \min_{-M\le i\le M} \{\Delta_i : \boldsymbol{x} \in \mathcal{A}_i\} \tag{2.6}$$

and

$$U(\boldsymbol{x}) = \max_{-M\le i\le M} \{\Delta_i : \boldsymbol{x} \in \mathcal{A}_i\}. \tag{2.7}$$

Several comments about the algorithm are in order.

(*i*) Note that we use a "greedy" heuristic for obtaining the minimum cardinality $\mathcal{A}_i$ by removing that $\boldsymbol{x}$ which maximizes $P_0(\mathcal{A}_i - \{\boldsymbol{x}\}, \Delta_i)$, and so on (thus enabling us to remove as many "dispensable" sample points as possible). This strategy is not necessarily optimal for constructing the smallest $\mathcal{A}_i$ in all cases, but other strategies are computationally costlier.

(*ii*) The guarantee of (2.3) (invariance with respect to the interchange of "successes" and "failures") follows from (2.5).

(*iii*) In Step 1.1, the grid on p_1 is chosen symmetrically around the midpoint $(1+\Delta)/2$ because

$$f(\boldsymbol{x}|p_1 = (1+\Delta)/2 + \epsilon, p_2 = p_1 - \Delta) = f(\boldsymbol{n} - \pi\boldsymbol{x}|p_1 = (1+\Delta)/2 - \epsilon, p_2 = p_1 - \Delta)$$

for $0 < \epsilon < (1-\Delta)/2$ when $n_1 = n_2$. Therefore $\boldsymbol{x} \in \mathcal{A}_i \Leftrightarrow \boldsymbol{n} - \pi\boldsymbol{x} \in \mathcal{A}_i$ is automatically satisfied.

(*iv*) It is not difficult to show that $f(\boldsymbol{x})$ is unimodal; hence the $\mathcal{A}_{ij}$ obtained in Step 1.2 are convex, and thus have no "holes" in them. However, it is possible that the $\mathcal{A}_i$ obtained in Step 1.3 may be non-convex and therefore any "holes" in them must be filled in Step 1.4. If the grid on p_1 is sufficiently fine (as was the case with our algorithm), we found that such "holes" do not arise, and hence Step 1.4 is unnecessary.

(*v*) Step 2.1 ensures that each outcome $\boldsymbol{x}$ is in a consecutive set of acceptance regions $\mathcal{A}_i$, which in turn ensures Property (*i*) of interval-valued confidence sets.

(*vi*) A listing of the FORTRAN program for the above algorithm is given in Coe (1989). The program is designed to handle $n_1, n_2 \le 20$. This program actually computes acceptance regions for the negative half of the Δ-space, i.e., for Δ_{-i}, $i = 1,\ldots,M$, and then obtains the acceptance regions for positive Δ-values using (2.5). For three decimal place accuracy,

we use an equispaced grid with $\Delta_{-M} = -0.9995, \Delta_{i+1} - \Delta_i = 0.0010$ and $M = 1000$. Thus suppose from (2.6) we get $L(\boldsymbol{x}) = 0.3075$ and from (2.7) we get $U(\boldsymbol{x}) = 0.7415$; then we know that the "true" (assuming unlimited degree of precision) $L(\boldsymbol{x})$ is in the interval (0.3065,0.3075) and the the "true" $U(\boldsymbol{x})$ is in the interval (0.7415,0.7425). Therefore, to three decimal place accuracy, the C.I. is given by [0.307, 0.742]. Note that this interval is found from the calculated values of the confidence limits by rounding down the lower limit and rounding up the upper limit. All computations were done on a Macintosh SE/30 machine running Absoft MacFortran II version 2.1.

For the partition on p_1 for fixed $\Delta_i < 0$, an equispaced grid of 20 to 25 points, including 0 and $1 + \Delta_i$, is found to be sufficient to yield

$$\min_{p_{ij}} P\{\boldsymbol{X} \in \mathcal{A}_i | p_1 = p_{ij}, p_2 = p_{ij} - \Delta_i\} \approx \inf_{p_1 \in I(\Delta_i)} P\{\boldsymbol{X} \in \mathcal{A}_i | p_1, \Delta_i\} \equiv P_0(\mathcal{A}_i, \Delta_i), \quad (2.8)$$

which is needed in the checking of (2.4).

3. COMPARISON WITH THE SANTNER-YAMAGAMI METHOD

3.1 A Brief Review of the SY Method

The SY method first partitions the sample space $\mathcal{X}$ into disjoint equivalence classes, $\mathcal{X}_i = \{\boldsymbol{x} \in \mathcal{X} : \hat{\Delta}(\boldsymbol{x}) = d_i\}$ $(-K \le i \le K)$ where $-1 = d_{-K} < \cdots < d_0 = 0 < \cdots < d_K = +1$ are the distinct values of $\hat{\Delta}(\boldsymbol{x}) = x_1/n_1 - x_2/n_2$. This puts a partial linear order on the $\mathcal{X}_i$. For a selected grid, $0 = \Delta_0 < \cdots < \Delta_M = 1$, the SY method constructs acceptance regions $\mathcal{A}_i$ for Δ_i $(0 \le i \le M)$ of the form:

$$\mathcal{A}_i = \mathcal{B}_s \bigcup \mathcal{X}_{s+1} \bigcup \cdots \bigcup \mathcal{X}_{t-1} \bigcup \mathcal{C}_t \text{ for some } s = s(i) \text{ and } t = t(i) \ (-K \le s \le t \le K), \quad (3.9)$$

where $\mathcal{B}_s \subseteq \mathcal{X}_s, \mathcal{C}_t \subset \mathcal{X}_t$, and $\mathcal{B}_s \neq \phi, \mathcal{X}_t - \mathcal{C}_t \neq \phi$ when $s < t$. It begins by constructing $\mathcal{A}_0$ with the following properties: $\boldsymbol{x} \in \mathcal{A}_0 \Leftrightarrow \boldsymbol{n} - \boldsymbol{x} \in \mathcal{A}_0, \mathcal{A}_0$ has as few points as possible, and $P_0(\mathcal{A}_0, 0) \ge 1 - \alpha$. (The infimum over p_1 required in evaluating P_0 (see (2.4)) is approximated by the discrete minimum over $p_1 = p_{i1}, \ldots, p_{iN}$, as in (2.8).) In general, having determined $\mathcal{A}_i$ $(i = 0, 1, \ldots, M-1)$ of the form (3.9), it builds $\mathcal{A}_{i+1}$ from $\mathcal{A}_i$ as follows: As a first try, set $\mathcal{A}_{i+1} = \mathcal{A}_i$. If condition (2.4) is satisfied for $\mathcal{A}_{i+1}$ then remove subsets $\mathcal{B}$ from $\mathcal{B}_s$ (if $\mathcal{B}_s = \phi$, relabel $\mathcal{X}_{s+1}$ as $\mathcal{B}_s$) as long as that condition continues to be satisfied. If condition (2.4) is not satisfied for $\mathcal{A}_{i+1}$ then add subsets $\mathcal{C}$ from $\mathcal{X}_t - \mathcal{C}_t$ in steps to $\mathcal{C}_t$ (if $\mathcal{C}_t = \mathcal{X}_t$, relabel $\mathcal{X}_{t+1}$ as $\mathcal{X}_t$ and $\mathcal{C}_t = \phi$) until that condition is satisfied, and then try deleting subsets $\mathcal{B}$.

Santner and Yamagami (1993) considered several different rules for choosing subsets $\mathcal{B}_s$ and $\mathcal{C}_t$. Their preferred rule is the so-called invariant rule, R_I, which deletes single sample points $\boldsymbol{x}$ that decrease $P_0(\mathcal{A}_i, \Delta_i)$ as little as possible, and adds single sample points $\boldsymbol{x}$ that increase $P_0(\mathcal{A}_i, \Delta_i)$ as much as possible, just as does our method.

TABLE 1

95% CT and SY Confidence Intervals for $(n_1, n_2) = (5, 5)$

(x_1, x_2)	CT Interval	SY Interval	CT Length	SY Length	Shorter Interval
(0,0)	(−0.451, 0.451)	(−0.462, 0.462)	0.902	0.904	CT
(0,1)	(−0.657, 0.268)	(−0.658, 0.268)	0.925	0.926	CT
(1,1)	(−0.488, 0.488)	(−0.489, 0.489)	0.976	0.978	CT
(0,2)	(−0.811, 0.100)	(−0.811, 0.189)	0.911	1.000	CT
(1,2)	(−0.700, 0.339)	(−0.700, 0.462)	1.039	1.162	CT
(2,2)	(−0.555, 0.555)	(−0.556, 0.556)	1.110	1.112	CT
(0,3)	(−0.924, 0.000)	(−0.927, 0.001)	0.924	0.928	CT
(1,3)	(−0.825, 0.238)	(−0.826, 0.239)	1.063	1.065	CT
(2,3)	(−0.645, 0.393)	(−0.647, 0.462)	1.038	1.109	CT
(3,3)	(−0.555, 0.555)	(−0.556, 0.556)	1.110	1.112	CT
(0,4)	(−0.990, −0.100)	(−0.990, −0.189)	0.890	0.801	SY
(1,4)	(−0.926, 0.100)	(−0.904, 0.189)	1.026	1.093	CT
(2,4)	(−0.825, 0.238)	(−0.826, 0.239)	1.063	1.065	CT
(3,4)	(−0.700, 0.339)	(−0.700, 0.462)	1.039	1.162	CT
(4,4)	(−0.488, 0.488)	(−0.489, 0.489)	0.976	0.978	CT
(0,5)	(−1.000, −0.339)	(−1.000, −0.462)	0.661	0.538	SY
(1,5)	(−0.990, −0.100)	(−0.990, −0.189)	0.890	0.801	SY
(2,5)	(−0.924, 0.000)	(−0.927, 0.001)	0.924	0.928	CT
(3,5)	(−0.811, 0.100)	(−0.811, 0.189)	0.911	1.000	CT
(4,5)	(−0.657, 0.268)	(−0.658, 0.268)	0.925	0.926	CT
(5,5)	(−0.451, 0.451)	(−0.462, 0.462)	0.902	0.924	CT

3.2 Numerical Examples

We tried the CT and SY methods on three examples, two with equal sample sizes, $(n_1, n_2) = (5, 5)$ and $(10, 10)$, and one with unequal sample sizes, $(n_1, n_2) = (15, 5)$. The 95% CT and SY C.I.'s (to three decimal place accuracy) for $(n_1, n_2) = (5, 5)$ are shown in Table 1 for the upper half of the sample space $\mathcal{X} = \{(x_1, x_2) : 0 \le x_i \le 5\}$ (those for the lower half being obtained by symmetry). Note that the actual coverage probabilities of both these methods depend on true (p_1, p_2) or equivalently (p_1, Δ), but they are at least 95%. We see that the CT interval is shorter than the SY interval for 30 out of 36 outcomes. Space does not permit us to give similar tables of confidence intervals for the other two examples. However, Table 2 gives a summary comparison between the CT and SY methods for all three examples. We see that the CT method gives shorter intervals for a majority of the outcomes in each example. The average interval width is also shorter for the CT method. But the computing time is longer for the CT method, and seems to increase faster with the size of the problem. For example, for the smallest problem with $(n_1, n_2) = (5, 5)$ having 36 outcomes, the CT method requires four times as much time as does the SY method, while for the largest problem with $(n_1, n_2) = (10, 10)$ having 121 outcomes, the CT method

TABLE 2

Comparison of CT and SY 95% Confidence Intervals

(n_1, n_2)	Average Interval Length		No. of Sample Points Interval is Shorter		Computing Time in mins:secs	
	CT	SY	CT	SY	CT	SY
(5, 5)	.9565	.9732	30	6	4:15	1:04
(15, 5)	.7509	.8290	74	22	15:36	2:08
(10,10)	.6865	.7263	75	46	20:38	2:36

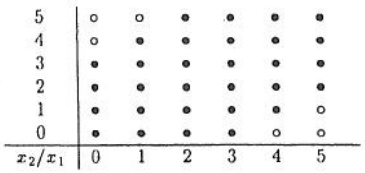

FIGURE 1

• Indicates Sample Points for Which CT Interval is Shorter
o Indicates Sample Points for Which SY Interval is Shorter
$(n_1, n_2) = (5, 5)$

requires eight times as much time as does the SY method. We shall discuss this issue further in the following subsection.

Figures 1–3 show the sample spaces $\mathcal{X}$ for the three examples with the outcomes x for which the CT intervals are shorter indicated by •'s, and those for which the SY intervals are shorter indicated by o's. We see that the CT intervals are shorter in the entire central portion of the sample space $\mathcal{X}$, while the SY intervals are shorter in the edge portion.

Santner and Yamagami (1993) compared the SY and TM methods for two examples studied here, namely, $(n_1, n_2) = (10, 10)$ and $(15, 5)$. They report that the TM intervals are longer than the SY intervals for all outcomes when $(n_1, n_2) = (15, 5)$, and for all outcomes except 8 (central) out of 121 when $(n_1, n_2) = (10, 10)$. This indicates that the TM intervals will also be longer than the CT intervals for almost all sample outcomes, although we have not actually carried out the computations for the TM method.

3.3 Discussion

We see from the above examples that the CT intervals are shorter than the SY intervals for x-values in the central portion of $\mathcal{X}$, these outcomes being most likely when Δ is close to 0, and are longer when Δ is close to ± 1. In practice, the Δ-values closer to 0 are more

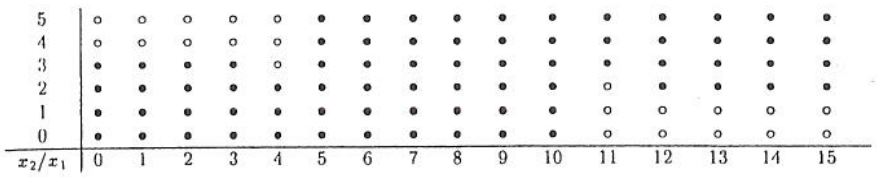

FIGURE 2

• Indicates Sample Points for Which CT Interval is Shorter
o Indicates Sample Points for Which SY Interval is Shorter
$(n_1, n_2) = (15, 5)$

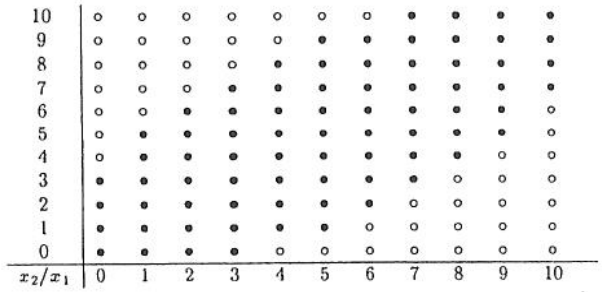

FIGURE 3

• Indicates Sample Points for Which CT Interval is Shorter
o Indicates Sample Points for Which SY Interval is Shorter
$(n_1, n_2) = (10, 10)$

common than those closer to ± 1. The CT intervals are also superior to the SY intervals in terms of the maximum and the average length criteria.

The reason for the long SY intervals for "middle" Δ-values is similar to that for the long Crow intervals when p is close to 1/2. From the description of the SY method given in Section 3.1, we see that it begins by constructing the acceptance region $\mathcal{A}_0$ for $\Delta = 0$ (which is symmetric in x and $n - x$ and falls in the central portion of $\mathcal{X}$), and tries to stay as close to it as possible when obtaining $\mathcal{A}_i$ from $\mathcal{A}_{i-1}$ for $i \geq 1$. Hence the x-values in the central portion of $\mathcal{X}$ are included in many more acceptance regions $\mathcal{A}_i$ than are the x-values near the edge of $\mathcal{X}$. As a result, the C.I.'s for the outcomes in the central portion of $\mathcal{X}$ are much longer than for those near the edge of $\mathcal{X}$.

Our method requires longer computing times because it begins the construction of the acceptance region $\mathcal{A}_i$ for each Δ_i from scratch. On the other hand, the SY method proceeds

in a stepwise manner, modifying $\mathcal{A}_{i-1}$ to obtain $\mathcal{A}_i$. It should be realized, of course, that computing times depend on the grids chosen for Δ and p_1. Note that both methods produce C.I.'s for *all* the outcomes in the sample space for given (n_1, n_2), and not just for a particular observed outcome $\boldsymbol{x}$ of interest. It may be possible to modify the methods so that they stop after the C.I. for the desired $\boldsymbol{x}$ has been calculated, thus cutting down the computing time.

4. CONFIDENCE INTERVALS FOR ρ AND ψ

4.1 Confidence Intervals for ρ

Small sample C.I.'s for ρ can be constructed using an algorithm similar to that for Δ described in Section 2. The main differences are as follows:

(i) In this case there is no invariance requirement with regard to the labelling of "successes" and "failure." However, the following invariance requirement with regard to the exchange of labels of populations is imposed:

$$[L_{\pi \boldsymbol{n}}(\pi \boldsymbol{x}), U_{\pi \boldsymbol{n}}(\pi \boldsymbol{x})] = [1/U_{\boldsymbol{n}}(\boldsymbol{x}), 1/L_{\boldsymbol{n}}(\boldsymbol{x})].$$

Here $[L_{\boldsymbol{n}}(\boldsymbol{x}), U_{\boldsymbol{n}}(\boldsymbol{x})]$ is a given $(1-\alpha)$-level C.I. for ρ when sample sizes $\boldsymbol{n} = (n_1, n_2)$ are used and the outcome $\boldsymbol{x} = (x_1, x_2)$ is observed.

(ii) The parameter ρ ranges over the infinite interval $[0, \infty]$. One could map this interval into a finite interval using one of the monotone transformations of ρ ($\tan^{-1}(\rho)$, $\tanh^{-1}(\rho)$ or $\rho/(1+\rho)$) suggested by Santner and Yamagami (1993), and apply the algorithm by forming an equispaced grid in this new interval. Our program works directly with ρ with an equispaced grid $\{\rho_{-i},\ (1 \le i \le M)\}$ for $\rho \in (0, 1)$ (specifically, the program uses a grid starting at 0.005 with steps of 0.01, which gives a two decimal place accuracy between 0 and 1), and $\rho_i = 1/\rho_{-i}$ for $\rho \in (1, \infty)$.

A FORTRAN program for implementing this algorithm is given in Coe (1989).

4.2 Confidence Intervals for ψ

The methods for calculating the C.I.'s for Δ and ρ are based on the unconditional probability distribution (2.1) of (X_1, X_2) (and thus yield unconditional C.I.'s). On the other hand, the method for ψ is based on the following conditional distribution of X_1, conditioned on $X_1 + X_2 = k$ (and thus yields conditional C.I.'s):

$$f(x_1|k, \psi) = \binom{n_1}{x_1}\binom{n_2}{k-x_1}\psi^{x_1} \Big/ \sum_{i=\ell}^{m}\binom{n_1}{i}\binom{n_2}{k-i}\psi^{i}, \tag{4.10}$$

where $\ell = \max(0, k-n_2)$ and $m = \min(k, n_1)$.

The following points may be noted about this algorithm.

(i) The C.I.'s for ψ are required to be invariant with respect to the labelling of the populations as follows:

$$[L_{\pi \boldsymbol{n}}(\pi \boldsymbol{x}), U_{\pi \boldsymbol{n}}(\pi \boldsymbol{x})] = [1/U_{\boldsymbol{n}}(\boldsymbol{x}), 1/L_{\boldsymbol{n}}(\boldsymbol{x})],$$

where $[L_{\boldsymbol{n}}(\boldsymbol{x}), U_{\boldsymbol{n}}(\boldsymbol{x})]$ is a given $(1-\alpha)$-level C.I. for ψ when sample sizes $\boldsymbol{n} = (n_1, n_2)$ are used and the outcome $\boldsymbol{x} = (x_1, x_2)$ is observed. Similarly, the C.I.'s are required to be invariant with respect to the labelling of "successes" and "failures" as follows:

$$[L_{\boldsymbol{n}}(\boldsymbol{n}-\boldsymbol{x}), U_{\boldsymbol{n}}(\boldsymbol{n}-\boldsymbol{x})] = [1/U_{\boldsymbol{n}}(\boldsymbol{x}), 1/L_{\boldsymbol{n}}(\boldsymbol{x})].$$

(ii) The parameter ψ ranges over the infinite interval $[0, \infty]$. Our program works with an equispaced grid $\{\psi_{-i}, (1 \le i \le M)\}$ for $\psi \in (0, 1)$ (specifically the program uses a grid starting at 0.005 with steps of 0.01, which gives two decimal place accuracy between 0 and 1), and $\psi_i = 1/\psi_{-i}$ for $\psi \in (1, \infty)$. Note that if $n_1 = n_2$, then it is sufficient to partition only $(0, 1)$ since the other half of the parameter space, viz. $(1, \infty)$, can be handled by symmetry.)

(iii) It should be noted that the construction of the acceptance regions $\mathcal{A}_i$ is easier here because we are in a one-dimensional sample space, and $f(x_1|k, \psi)$ is unimodal in x_1.

A FORTRAN program for implementing this algorithm is given in Coe (1989). The numerical results obtained with this program were found to be comparable to those obtained using Baptista and Pike's (1977) method. Note that their method does not involve inversion of acceptance regions as does our method. Also, our method seems more convenient for embedding into a group sequential scheme.

BIBLIOGRAPHY

Baptista, J. and Pike, M. C. (1977). "Exact two-sided confidence limits for the odds ratio in a 2×2 table," *Applied Statistics*, **26**, 214–220.

Blyth, C. and Still, H. A. (1983). "Binomial confidence intervals," *Journal of the American Statistical Association*, **78**, 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*, **47**, 381–391.

Coe, P. R. (1989). *Exact Repeated Confidence Intervals for Binomial Parameters in Group Sequential Experiments*, Doctoral Dissertation, Northwestern University.

Coe, P. R. and Tamhane, A. C. (1993). "Exact repeated confidence intervals for Bernoulli parameters in a group sequential clinical trial," *Controlled Clinical Trials*, **14**, 19–29.

Cornfield, J. (1956). "A statistical problem arising from retrospective studies," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, **4**, 135–148.

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

Fisher, R. A. (1935). "The logic of inductive inference," *Journal of the Royal Statistical Society*, **98**, 39–54.

Jennison, C. and Turnbull, B. W. (1984). "Repeated confidence intervals for group sequential clinical trials," *Controlled Clinial Trials*, **5**, 33–45.

Jennison C. and Turnbull, B. W. (1990). "Interim analyses: The repeated confidence intervals approach," *Journal of the Royal Statistical Society, Series B*,

Lai, T. L. (1984). "Incorporating scientific, ethical and economic considerations into the design of clinical trials in the pharmaceutical industry: A sequential approach," *Communications in Statistics, Series A*, **13**, 2355–2368.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, Second Edition, New York: John Wiley.

Santner, T. J. and Snell, M. K. (1980). "Small sample confidence intervals for $p_1 - p_2$ and p_1/p_2 in 2×2 contingency tables," *Journal of the American Statistical Association*, **75**, 386–394.

Santner, T. J. and Yamagami, S. (1993)."Invariant small sample confidence intervals for the difference of two success probabilities," *Communications in Statistics (Simulation and Computation)*, **22**, 33- -59.

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

Thomas, D. G. and Gart, J. J. (1977). "A table of exact confidence limits for differences and ratios of two proportions and their odds ratios," *Journal of the American Statistical Association*, **72**, 73–76.

Received January 1992; Revised May 1993.

COMMUN. STATIST.—SIMULA., 22(4), 939-954 (1993)

COMBINING MONTE CARLO AND COX TESTS OF NON-NESTED HYPOTHESES

Nicholas Schork

Department of Medicine &
Department of Epidemiology
R6592 Kresge I
University of Michigan
Ann Arbor, Michigan, 48109-0500

Key words and phrases: hypothesis testing; likelihood ratio test; power; separate families of hypotheses; simulation

ABSTRACT

The problem of deriving reliable tests for separate families of hypotheses is discussed. Two competing methodologies for testing hypotheses from separate distributional families, the classical asymptotic approach of Cox [1961,1962] and more modern methods using Monte Carlo or parametric bootstrap simulation, are contrasted. It is shown that the two methods can be combined to form a test with excellent statistical properties. Variants of simulation-based tests are discussed. In addition, simple computational strategies using parallel computers are described that can be used to reduce the combined test's heavy simulation load.

I. INTRODUCTION

Many hypothesis tests are *parameter oriented*: a basic distributional model for relevant data is given and the question of primary concern is whether or not a certain parameter, or parameter vector, assumed in or describing the given distributional model, takes on a certain theoretically derived or hypothesized value. However, there exist situations where what is in question is *not* necessarily whether or not a certain parameter takes on a certain value, but rather whether or not the data conform to one of two or more parametric distributional *models* (e.g., normal, log-normal, Weibull, etc.). One approach used in such situations involves the use of goodness-of-fit techniques to determine whether or not one or more of the distributional models does not