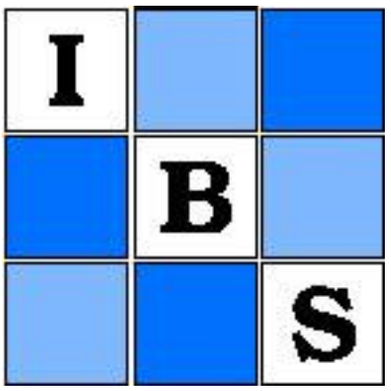

BIOMETRICS 55, 1202–1209
December 1999

Test-Based Exact Confidence Intervals for the Difference of Two Binomial Proportions

Ivan S. F. Chan* **and Zhongxin Zhang**

Clinical Biostatistics, Merck Research Laboratories, BL 2-3, West Point, Pennsylvania 19486, U.S.A.
* *email:* Ivan_Chan@Merck.Com

SUMMARY. Confidence intervals are often provided to estimate a treatment difference. When the sample size is small, as is typical in early phases of clinical trials, confidence intervals based on large sample approximations may not be reliable. In this report, we propose test-based methods of constructing exact confidence intervals for the difference in two binomial proportions. These exact confidence intervals are obtained from the unconditional distribution of two binomial responses, and they guarantee the level of coverage. We compare the performance of these confidence intervals to ones based on the observed difference alone. We show that a large improvement can be achieved by using the standardized Z test with a constrained maximum likelihood estimate of the variance.

KEY WORDS: Constrained MLE; Exact test; Two-by-two table; Unconditional distribution.

1. Introduction

In clinical trials designed to compare a new treatment with a standard treatment, the difference in proportions of responses to a specific endpoint between the two groups is often of primary interest. Confidence intervals are typically provided to estimate the treatment difference. An asymptotic confidence interval computed based on the normal approximation is quite good in large samples, but it may behave poorly when the sample size is small or when the observed rates are near the extreme. Santner and Snell (1980) proposed methods to construct confidence intervals that guarantee the coverage probability. Their method of constructing exact confidence intervals based on the tail distribution of the observed rate difference has been implemented in the commercially available software StatXact (Cytel Software, 1995).

In a vaccine clinical trial to investigate whether a new manufacturing process provides improvement over the current process, subjects were randomized to receive either the new or the current process materials. The preliminary data showed that the proportions of subjects responding to the vaccine were .944 (17/18) and .611 (11/18) for the new and current processes, respectively. The rate difference was .333, suggesting that the new process had noticeable improvement. Since the sample size was rather small, the simple tail-based method (Cytel Software, 1995) was used to provide a 95% exact confidence interval for the rate difference (new − current). It yielded an interval of (−.019, .630), indicating no significant improvement, however.

We note that the tail-based method could be overly conservative because it uses only the observed rate difference to rank the outcomes without considering the associated variability. In this paper, we propose three other statistics for constructing exact confidence intervals that take into account the variability of the outcome. We show that confidence intervals based on these three statistics provide substantial improvement over the simple tail-based method proposed by Santner and Snell (1980) and implemented in StatXact (Cytel Software, 1995). For the above example, our recommended method (standardized Z statistics with constrained maximum likelihood estimate of variance) yields a 95% confidence interval of (.049, .593), which is completely within the one produced by the simple tail-based method and indicates a statistically significant improvement of the new process.

For some background on the problem, consider a binomial trial comparing two groups of sizes N_1 and N_2. Let X and Y be the independent responses distributed as binomial random variables having parameters (N_1, P_1) and (N_1, P_2), respectively. The parameter of interest is the difference of the two binomial probabilities

$$\delta = P_1 - P_2, \tag{1}$$

where the range of δ is $(-1, 1)$. The joint probability mass function of (X, Y) is a product of two binomial probabilities. Let $P = P_1$ and substitute P_2 with $(P - \delta)$ from (1). The joint probability mass function becomes

$$\Pr\{X = i, Y = j \mid \delta, P\} = \binom{N_1}{i}\binom{N_2}{j} P^i (1-P)^{N_1 - i} (P-\delta)^j (1+\delta-P)^{N_2 - j} \tag{2}$$

for $0 \le i \le N_1$ and $0 \le j \le N_2$. For any given δ, the domain of P is

$$D(\delta) = \{P : \max(0, \delta) \le P \le \min(1, 1+\delta)\}. \tag{3}$$

Note that P is a nuisance parameter for the inference on δ. For testing the hypothesis H_0: $\delta = 0$, the marginal sum $(X + Y)$ is the sufficient statistic for P. An exact test conditional on the sufficient statistic is the well-known Fisher's exact test.

Suissa and Shuster (1985) and Haber (1986) proposed an exact test for H_0: $\delta = 0$ based on the unconditional distribution given in (2). They used the maximization method (Basu, 1977) to eliminate the effect of the nuisance parameter P, in which the exact p value is obtained by maximizing the rejection probability over the domain of P. They showed the exact unconditional test is more powerful than the Fisher's exact test. For testing null hypotheses with a nonzero difference, no simple sufficient statistics exist for P. Chan (1998) developed an exact unconditional test for such hypotheses based on a statistic proposed by Miettinen and Nurminen (1985) and Farrington and Manning (1990).

In Section 2, we extend the ideas of exact unconditional tests to constructing exact confidence intervals for the difference (δ) of proportions in two independent binomial samples. In Section 3, we discuss four well-known test statistics that are used to construct exact confidence intervals, including the simple tail-based method in StatXact (Cytel Software, 1995) as a special case. The method of dealing with a nuisance parameter with a restricted search is discussed in Section 4. We illustrate these methods with several examples in Section 5, and we examine and compare their performance in Section 6.

2. Test-Based Exact Confidence Interval

Our goal is to construct a two-sided $(1 - 2\alpha)100\%$ confidence interval (δ_L, δ_U) for the true difference δ by using the exact method (Clopper and Pearson, 1934). The proposed exact confidence interval eliminates aberrations and guarantees strict conservatism in the sense that the coverage probability is at least $1 - 2\alpha$ for every δ. The two confidence bounds are obtained by inverting the test procedure for two one-sided hypotheses, one for the lower bound and the other for the upper bound.

We find the upper bound by considering the one-sided hypothesis

$$H_0\colon \delta = \delta_0 \quad \text{versus} \quad H_1\colon \delta < \delta_0 \tag{4}$$

for a specified value δ_0. Without loss of generality, let $T = T(X, Y; \delta_0)$ be a test statistic with the rejection region defined by the left tail of the statistic. In other words, small values of T are in favor of the alternative hypothesis. Several examples of the T statistic will be discussed in the next section.

Following the ideas of Suissa and Shuster (1985), Haber (1986), and Chan (1998), one can develop an exact unconditional test based on the T statistic for the one-sided hypothesis in (4). For an observed outcome (i, j) of (X, Y), the exact p value for δ_0 is defined by

$$p(i, j \mid T, \delta_0) = \max_{P \in D} \Pr[T(X, Y; \delta_0) \le T(i, j; \delta_0) \mid \delta_0, P], \tag{5}$$

where the probability is evaluated using (2) and $D = D(\delta_0)$ as given in (3) is the space of the nuisance parameter. This test considers all possible combinations of the outcomes given the sample sizes N_1 and N_2. The p value is obtained by maximizing the tail probability over the domain of the nuisance parameters $D(\delta_0)$ to account for the worst configuration. For an α-level test, the null hypothesis in (4) will be rejected if $p(i, j \mid T, \delta_0) \le \alpha$.

Using the duality of hypothesis testing and interval estimation (Rohatgi, 1984, pp. 224–225), one can derive, using (5), the one-sided $(1 - \alpha)100\%$ confidence set for δ based on the observed outcome (i, j) as

$$C_{\alpha,T}(i, j) = \{\delta : p(i.j \mid T, \delta) > \alpha\} \tag{6}$$

and

$$\Pr\{\delta \in C_{\alpha,T}(i, j)\} \ge 1 - \alpha.$$

It is easy to see that the confidence set $C_{\alpha,T}(i, j)$ is an unique interval of the form $(-1, \delta_U)$ if the tail probability function

$$g(\delta) = p(i, j \mid T, \delta) \tag{7}$$

is monotonely decreasing in δ. In this case, the upper limit (δ_U) is the value of δ for which $g(\delta) = \alpha$. If $g(\delta)$ fluctuates with δ, then the confidence set may consist of multiple disjoint intervals. To be conservative, we define the upper limit to be

$$\delta_U = \sup_{\delta} C_{\alpha,T}(i, j)$$

so that the interval $(-1, \delta_U)$ covers the true difference (δ) with at least $(1 - \alpha)$ probability.

Extending the above method to obtain the lower limit (δ_L) of the confidence interval is straightforward. We invert the test T for the hypothesis

$$H_0\colon \delta = \delta_0 \quad \text{versus} \quad H_1\colon \delta > \delta_0. \tag{8}$$

Now large values of T favor the alternative hypothesis in (8). Following a similar procedure, we can obtain the lower limit as

$$\delta_L = \inf_{\delta} \left\{ \delta : \max_{P \in D} \Pr[T(X, Y; \delta) \ge T(i, j; \delta) \mid \delta, P] > \alpha \right\}$$

so that the interval $(\delta_L, 1)$ covers the true δ with at least $(1 - \alpha)$ probability.

The interval (δ_L, δ_U) forms a $(1-2\alpha)100\%$ exact confidence interval for the true difference (δ) since

$$\Pr\{\delta_L \le \delta \le \delta_U\} = 1 - \Pr\{\delta < \delta_L\} - \Pr\{\delta > \delta_U\} \ge 1 - 2\alpha.$$

Note that the exact confidence interval is generally conservative due to the discreteness of the distribution in (2) and the fact that the p value is obtained by taking the maximum over the domain of the nuisance parameter.

It is interesting to observe from equation (2) that the probability of observing the outcome (i, j) given $(P_1 = P, P_2 = P - \delta)$ is the same as that of the outcome $(N_1 - i, N_2 - j)$ given $(P_1 = 1 - P, P_2 = 1 - P + \delta)$. Therefore, the confidence interval for δ obtained based on the outcome (i, j) is the same as the confidence interval for $(-\delta)$ obtained based on the outcome $(N_1 - i, N_2 - j)$. In other words, if the outcome (i, j) yields a confidence interval (δ_L, δ_U) for δ, then the confidence interval for δ based on the outcome $(N_1 - i, N_2 - j)$ will be $(-\delta_U, -\delta_L)$.

3. Choices of Test Statistics

In this section, we discuss four test statistics that can be used to construct exact confidence intervals. In all cases, small values of the statistic favor the alternative hypothesis in (4).

Simple statistic (S). The tail-based exact confidence interval proposed by Santner and Snell (1980) and implemented

in StatXact (Cytel Software, 1995) uses the observed rate difference as the test statistic

$$S(i, j; \delta_0) = \hat{P}_1 - \hat{P}_2,$$

where $\hat{P}_1 = i/N_1$ and $\hat{P}_2 = j/N_2$. Note that the test statistic is independent of δ (or $\delta_0 = 0$) and the tail of any observed outcome is fixed. Therefore, the tail probability function $g(\delta)$ in (7) is monotone in δ, and the confidence set $C_{\alpha,T}(\delta)$ defined in (6) always yields a unique interval. Since the S statistic depends only on the observed difference, its distribution may have many ties, especially when sample sizes are equal in both groups. In addition, the same confidence interval will be obtained from two different sample outcomes with the same observed difference even though they may have different variance estimates. This method seems overly conservative, as one would expect a narrower confidence interval for the sample point having a smaller variance.

Binomial Z statistic (Z_1). Another choice of statistic is the well-known binomial Z statistic based on the normal approximation,

$$Z_1(i, j; \delta_0) = \frac{\hat{P}_1 - \hat{P}_2}{\{\hat{P}_1(1 - \hat{P}_1)/N_1 + \hat{P}_2(1 - \hat{P}_2)/N_2\}^{1/2}}.$$

Like the S statistic, Z_1 does not depend on δ, and thus the confidence set $C_{\alpha,T}(\delta)$ in (6) always yields a unique interval. Since Z_1 standardizes the observed difference by its variance, the distribution of Z_1 has fewer ties than that of S. As a result, Z_1 is more sensitive than S as a test statistic. One would expect that the confidence interval based on Z_1 would be generally narrower than that based on S.

Note that, when $(\hat{P}_1, \hat{P}_2)$ takes on the values of $(0, 0)$, $(1, 1)$, $(0, 1)$, or $(1, 0)$, the variance estimate is zero and Z_1 is undefined. We use a slight adjustment in these situations. For the former two cases, we define Z_1 to be zero since the observed difference is zero. For the third case, use $\hat{P}_1 = (2N_1)^{-1}$ and $\hat{P}_2 = 1 - (2N_2)^{-1}$ in calculating the variance estimate. Similarly, we use $\hat{P}_1 = 1 - (2N_1)^{-1}$ and $\hat{P}_2 = (2N_2)^{-1}$ in calculating the variance estimate for the last case. As such, the Z_1 statistic will be smallest at $(0, 1)$ and largest at $(1, 0)$.

δ-Projected Z statistic (Z_2). For testing one-sided hypotheses in the form of (4), Miettinen and Nurminen (1985) and Farrington and Manning (1990) proposed a modification to the binomial Z statistic and restricted the estimation of variance under the null hypothesis, i.e.,

$$Z_2(i, j; \delta_0) = \frac{\hat{P}_1 - \hat{P}_2 - \delta_0}{\{\tilde{P}_1(1 - \tilde{P}_1)/N_1 + \tilde{P}_2(1 - \tilde{P}_2)/N_2\}^{1/2}}.$$

Here $\tilde{P}_1$ and $\tilde{P}_2$ are the maximum likelihood estimates of P_1 and P_2, respectively, under the constraint specified in the null hypothesis (4). Explicit formulas for calculating $\tilde{P}_1$ and $\tilde{P}_2$ can be found in these two references. Miettinen and Nurminen (1985) also developed an asymptotic confidence interval based on Z_2. These two studies demonstrated that the normal approximation to Z_2 performs better than the one using the sample variance estimate in terms of controlling type I error rates of the test and maintaining coverage probabilities of confidence intervals. We call Z_2 the δ-projected statistic because $(\tilde{P}_1, \tilde{P}_2)$ is a projection of $(\hat{P}_1, \hat{P}_2)$ on the line $P_1 - P_2 = \delta_0$.

Note that the δ-projected Z statistic involves δ_0. Although rarely seen in practice, there is a theoretical possibility that the tail of an observed outcome based on Z_2 will change with a new hypothesized value of δ and thus the tail probability function $g(\delta)$ will fluctuate with δ. The consequence is twofold: (a) substantially more computational effort (compared to S or Z_1) is needed to update the tail with changing δ in the search for the confidence limits, and (b) the procedure will potentially yield multiple disjoint intervals and the resulting confidence limits will be overconservative since they represent the two extremes of these disjoint intervals. To eliminate these two problems, we adopt a modification to the procedure. First, we compute the asymptotic confidence interval $(\delta_{\mathrm{L}}^A, \delta_{\mathrm{U}}^A)$ using the approach of Miettinen and Nurminen (1985). This usually provides a good approximation to the exact interval. Then we use δ_{L}^A (or δ_{U}^A) as the final hypothesized value of δ in Z_2 to define the tail of the observed outcome for calculating the upper (or lower) limit.

Likelihood ratio statistic (LR). Denote the log likelihood function by

$$l(P, \delta \mid (i, j)) = \log\{\Pr(X = i, Y = j \mid \delta, P)\}.$$

The log-likelihood ratio statistic for the one-sided hypothesis (4) is given by

$$\mathrm{LR}(i, j; \delta_0) = \mathrm{sign}(\hat{\delta} - \delta_0)\{l(\hat{P}_1, \hat{\delta} \mid (i, j)) - l(\tilde{P}_1, \delta_0 \mid (i, j))\},$$

where $\hat{\delta} = \hat{P}_1 - \hat{P}_2$ and $\mathrm{sign}(x) = 1$, 0, or -1 if $x > 0$, $= 0$, or < 0, respectively. Just like the δ-projected Z statistic, the likelihood ratio statistic also involves δ_0. We adapt the same modification as for Z_2 to define the tail in calculating the confidence limits.

4. Restricted Search of Nuisance Parameter

The above methods search the maximum tail probability over the whole domain of the nuisance parameter (P). This search casts for the worst scenario but might result in overconservative confidence intervals if the peak occurs at a place far away from the true value of P. Berger and Boos (1994) proposed a method for a valid p value that maximizes the tail probability over a confidence interval for the nuisance parameter and showed that conservatism could be reduced if the peak was outside the confidence interval. StatXact (Cytel Software, 1995, Section 13.3) adapted their method for the confidence interval based on the simple statistic (S). We apply this method to other statistics here. Basically, one first constructs two $100(1 - \gamma)\%$ confidence intervals $A_1 = (L_1, U_1)$ and $A_2 = (L_2, U_2)$ for P_1 and P_2, respectively. Assuming the event $\varepsilon : (P_1, P_2) \in A_1 \times A_2$ is true, the range of δ is $R = (L_1 - U_2, U_1 - L_2)$ and thus the range for the nuisance parameter is restricted to $D_{\mathrm{R}}(\delta) = \{P : \max(L_1, L_2 + \delta) \leq P \leq \min(U_1, U_2 + \delta)\}$. Then one searches for the maximum tail probability over this restricted range $D_{\mathrm{R}}(\delta)$. The confidence limit will be the value of δ such that $g(\delta) = \alpha - \gamma$, where $P \in D_{\mathrm{R}}(\delta)$ and γ serves as the penalty for restricting the search of the maximum tail probability. It can be shown that the confidence interval has the right coverage since

$$\begin{aligned}
\Pr\{\delta \notin [\delta_{\mathrm{L}}, \delta_{\mathrm{U}}]\} &= \Pr\{\delta \notin [\delta_{\mathrm{L}}, \delta_{\mathrm{U}}] \mid \varepsilon\}\Pr(\varepsilon) \\
&\quad + \Pr\{\delta \notin [\delta_{\mathrm{L}}, \delta_{\mathrm{U}}] \mid \varepsilon^c\}\Pr(\varepsilon^c) \\
&\leq \Pr\{\delta \notin [\delta_{\mathrm{L}}, \delta_{\mathrm{U}}] \mid \varepsilon\} + \Pr(\varepsilon^c) \\
&\leq \Pr\{\delta < \delta_{\mathrm{L}} \mid \varepsilon\} + \Pr\{\delta > \delta_{\mathrm{U}} \mid \varepsilon\} + 2\gamma \\
&\leq (\alpha - \gamma) + (\alpha - \gamma) + 2\gamma = 2\alpha
\end{aligned}$$

Table 1

Ninety-five percent exact confidence intervals for the illustrative examples

Outcome			Search over entire domain ($\gamma = 0$)				Search over restricted domain ($\gamma = .001$)			
$\hat{P}_1 = X/N_1$	$\hat{P}_2 = X/N_2$	Observed difference $\hat{\delta} = \hat{P}_1 - \hat{P}_2$	Simple statistic (S)	Binomial Z statistic (Z_1)	δ-projected Z statistic (Z_2)	Likelihood ratio statistic (LR)	Simple statistic (S)	Binomial Z statistic (Z_1)	δ-projected Z statistic (Z_2)	Likelihood ratio statistic (LR)
17/18	11/18	.33	(−.019, .629)	(.059, .668)	(.049, .593)	(.049, .593)	(−.019, .630)	(.059, .658)	(.047, .595)	(.049, .595)
9/10	5/10	.40	(−.086, .762)	(−.020, .762)	(−.020, .741)	(−.049, .741)	(−.089, .764)	(−.020, .762)	(−.024, .743)	(−.052, .743)
12/15	7/15	.33	(−.056, .654)	(−.024, .652)	(−.024, .637)	(−.030, .637)	(−.059, .656)	(−.024, .652)	(−.027, .639)	(−.031, .639)
0/10	0/10	0	(−.456, .456)	(−.456, .456)	(−.309, .309)	(−.309, .309)	(−.424, .424)	(−.422, .422)	(−.311, .311)	(−.311, .311)
0/10	0/20	0	(−.389, .389)	(−.389, .389)	(−.188, .309)	(−.168, .309)	(−.272, .378)	(−.271, .376)	(−.190, .311)	(−.190, .311)
1/4	0/4	.25	(−.510, .830)	(−.389, .830)	(−.389, .806)	(−.389, .806)	(−.515, .832)	(−.389, .830)	(−.394, .809)	(−.394, .809)
2/4	1/4	.25	(−.510, .830)	(−.510, .821)	(−.510, .830)	(−.598, .830)	(−.515, .832)	(−.510, .821)	(−.515, .832)	(−.598, .832)

using the fact that a probability cannot exceed one and the Bonferroni inequality that $\Pr(\varepsilon^c) \leq 2\gamma$.

5. Illustrative Examples

To illustrate the methodology, we used the four test statistics described in Section 3 to calculate exact confidence intervals in several examples, including the vaccine study described in the Introduction. Throughout, we used a bisection procedure to find the confidence limits for δ. For the confidence interval with a restricted search for the nuisance parameter, we will use $\gamma = .001$, as used in Berger (1996) and StatXact (Cytel Software, 1995). The computer programs were written in FORTRAN 77 and are available from the first author.

Table 1 lists the examples and their 95% exact confidence intervals with and without the restricted search of the nuisance parameter. The first example is the vaccine study, where the simple statistic fails to show an improvement of the new manufacturing process since the lower bound of the confidence interval for the difference is less than zero. However, all other methods produce much narrower confidence intervals with lower bounds exceeding zero, thus indicating a significant improvement of the new process. The second example was taken from StatXact (Cytel Software, 1995, Section 13.9), and the third example was from an influenza vaccine study reported by Fries et al. (1993) where the outcomes were the proportions of subjects who developed clinical symptoms of influenza after viral challenge among placebo controls and vaccine recipients, respectively. These two examples represent outcomes associated with relatively higher sample variability. The δ-projected Z statistic yields the shortest confidence intervals that are completely within those based on all other statistics. It is noted that the restricted search does not reduce the length of the confidence intervals in these two examples.

The fourth and fifth examples represent extreme outcomes taken from Miettinen and Nurminen (1985). The restricted search reduces the length of confidence intervals based on the S and Z_1 statistics. But the δ-projected Z statistic (Z_2) and the likelihood ratio statistic (LR), without using restricted search, provide substantially shorter confidence intervals in these extreme outcomes, however.

The last two examples, $(1/4, 0/4)$ and $(2/4, 1/4)$, are used to explicitly illustrate the benefit of accounting for the variability of the rate difference. These two examples have the same observed rate difference of .25, and they yield the same confidence interval $(-.510, .830)$ based on the simple statistic (S) without restricted search of nuisance parameter. However, the variability with outcome $(2/4, 1/4)$ is larger than with $(1/4, 0/4)$, and hence the Z_1 statistic is smaller for $(2/4, 1/4)$ than for $(1/4, 0/4)$. Since the outcome $(1/4, 0/4)$ is in the right tail of $(2/4, 1/4)$ based on Z_1, the upper bound estimate for outcome $(2/4, 1/4)$ based on Z_1 is smaller (.821) than the one

Table 2

Average length of 95% *exact confidence intervals (number in parentheses gives the percent of confidence intervals with the shortest length among the eight methods)*

	Search over entire domain ($\gamma = 0$)				Search over restricted domain ($\gamma = .001$)			
Sample size (N_1, N_2)	Simple statistic (S)	Binomial Z statistic (Z_1)	δ-projected Z statistic (Z_2)	Likelihood ratio statistic (LR)	Simple statistic (S)	Binomial Z statistic (Z_1)	δ-projected Z statistic (Z_2)	Likelihood ratio statistic (LR)
(4, 4)	1.227 (28)	1.190 (52)	1.168 (76)	1.182 (76)	1.234 (0)	1.190 (52)	1.175 (0)	1.188 (0)
(5, 5)	1.121 (22)	1.080 (44)	1.050 (67)	1.066 (67)	1.127 (0)	1.079 (44)	1.057 (0)	1.071 (0)
(6, 6)	1.037 (14)	.990 (39)	.960 (59)	.971 (59)	1.043 (0)	.990 (39)	.967 (0)	.976 (0)
(7, 7)	.968 (13)	.919 (34)	.890 (56)	.896 (53)	.974 (0)	.918 (34)	.897 (0)	.900 (0)
(8, 8)	.911 (9)	.862 (31)	.831 (63)	.845 (44)	.916 (0)	.861 (31)	.837 (0)	.845 (0)
(9, 9)	.863 (8)	.816 (28)	.781 (62)	.793 (48)	.867 (0)	.814 (28)	.787 (0)	.794 (0)
(10, 10)	.821 (6)	.775 (26)	.740 (62)	.750 (47)	.824 (0)	.772 (26)	.745 (0)	.750 (0)
(11, 11)	.785 (6)	.742 (26)	.704 (63)	.713 (43)	.787 (0)	.738 (26)	.710 (1)	.715 (1)
(12, 12)	.753 (4)	.710 (24)	.674 (59)	.682 (44)	.754 (0)	.705 (24)	.679 (0)	.682 (0)
(13, 13)	.724 (4)	.683 (24)	.647 (60)	.654 (41)	.724 (0)	.678 (24)	.651 (0)	.654 (0)
(14, 14)	.699 (3)	.658 (32)	.623 (63)	.631 (40)	.698 (0)	.653 (32)	.628 (0)	.629 (0)
(15, 15)	.675 (3)	.636 (32)	.602 (64)	.609 (32)	.674 (0)	.630 (32)	.606 (2)	.608 (2)
(16, 16)	.654 (2)	.617 (29)	.582 (63)	.590 (34)	.652 (0)	.610 (29)	.586 (1)	.588 (2)
(17, 17)	.635 (2)	.599 (27)	.564 (63)	.570 (36)	.632 (0)	.592 (28)	.568 (2)	.569 (1)
(18, 18)	.617 (2)	.583 (28)	.547 (60)	.554 (39)	.614 (0)	.575 (29)	.551 (2)	.552 (3)
(19, 19)	.601 (2)	.568 (28)	.533 (58)	.539 (40)	.597 (0)	.559 (29)	.536 (3)	.537 (4)
(20, 20)	.586 (2)	.554 (25)	.519 (58)	.524 (36)	.581 (0)	.545 (26)	.522 (4)	.523 (3)
(4, 3)	1.245 (70)	1.245 (70)	1.231 (70)	1.256 (70)	1.246 (0)	1.239 (70)	1.238 (0)	1.259 (0)
(10, 5)	.981 (9)	.948 (9)	.896 (61)	.923 (27)	.974 (0)	.936 (12)	.899 (9)	.906 (9)
(10, 7)	.839 (52)	.842 (32)	.811 (34)	.835 (36)	.840 (0)	.836 (34)	.816 (0)	.825 (5)
(13, 11)	.708 (39)	.707 (32)	.674 (35)	.692 (29)	.708 (0)	.700 (33)	.678 (1)	.683 (6)
(15, 10)	.722 (11)	.709 (24)	.672 (53)	.690 (30)	.720 (0)	.699 (28)	.674 (9)	.678 (7)
(15, 13)	.657 (34)	.656 (31)	.624 (38)	.639 (26)	.656 (0)	.649 (31)	.627 (4)	.631 (4)
(20, 10)	.710 (4)	.695 (5)	.636 (59)	.657 (19)	.700 (0)	.674 (6)	.636 (16)	.638 (16)
(20, 15)	.602 (19)	.597 (23)	.561 (45)	.574 (24)	.599 (2)	.586 (27)	.562 (8)	.564 (9)

Table 3

Average expected error rates of coverage (%) of 95% exact confidence intervals[a]

Sample size (N_1, N_2)	Search over entire domain ($\gamma = 0$)								Search over restricted domain ($\gamma = .001$)							
	Simple statistic (S)		Binomial Z statistic (Z_1)		δ-projected Z statistic (Z_2)		Likelihood ratio statistic (LR)		Simple statistic (S)		Binomial Z statistic (Z_1)		δ-projected Z statistic (Z_2)		Likelihood ratio statistic (LR)	
	$\delta = 0$	$\delta = .4$	$\delta = 0$	$\delta = .4$	$\delta = 0$	$\delta = .4$	$\delta = 0$	$\delta = .4$	$\delta = 0$	$\delta = .4$	$\delta = 0$	$\delta = .4$	$\delta = 0$	$\delta = .4$	$\delta = 0$	$\delta = .4$
(4, 4)	.35	.82	.35	2.06	.35	2.06	.35	2.06	.35	.82	.35	2.06	.35	2.06	.35	2.06
(5, 5)	1.04	.78	1.04	1.33	1.04	1.33	1.04	1.33	1.04	.78	1.04	1.33	1.04	1.33	1.04	1.33
(6, 6)	1.98	1.83	1.98	2.99	1.98	2.99	1.98	2.09	1.98	1.83	1.98	2.99	1.98	2.99	1.98	2.09
(7, 7)	.60	1.15	1.63	2.61	1.63	2.61	1.63	1.26	.60	1.15	1.63	2.61	1.63	2.61	1.63	2.11
(8, 8)	1.02	.77	2.03	1.64	2.03	1.64	2.03	1.69	1.02	.83	2.03	1.64	2.03	1.69	2.03	1.69
(9, 9)	1.54	2.82	2.51	2.82	2.51	3.09	2.51	3.09	1.54	2.82	2.51	2.82	2.51	3.09	2.51	3.09
(10, 10)	2.13	1.94	3.06	2.89	3.06	3.99	3.06	3.99	2.13	1.94	3.06	2.89	3.06	3.99	3.06	3.99
(11, 11)	.79	2.79	2.41	2.79	3.29	3.60	3.29	3.12	.79	2.80	2.41	2.80	3.29	3.60	3.29	3.12
(12, 12)	1.09	1.83	3.54	1.83	3.54	3.89	3.54	3.89	1.09	1.83	3.54	1.83	3.54	3.89	3.54	3.89
(13, 13)	1.43	1.24	3.82	2.21	3.82	2.88	3.11	2.99	1.43	1.25	3.82	2.31	3.11	2.88	3.11	2.99
(14, 14)	1.81	2.93	3.36	2.93	3.43	4.03	3.43	4.03	1.81	2.93	3.36	2.93	3.43	4.03	3.43	4.03
(15, 15)	2.21	2.07	3.78	3.69	3.78	3.86	3.78	3.59	2.21	2.07	3.78	3.69	3.78	3.86	3.78	3.86
(16, 16)	.95	2.63	3.23	3.64	3.94	4.13	3.94	4.21	1.07	2.63	3.94	3.64	3.94	4.13	3.94	4.21
(17, 17)	1.18	1.80	3.43	3.37	4.11	3.42	4.11	4.29	1.29	1.85	3.43	3.42	4.11	3.42	4.11	4.29
(18, 18)	1.42	3.55	3.64	4.02	4.29	4.37	4.29	4.15	1.42	3.55	3.64	4.02	4.29	4.37	4.29	4.15
(19, 19)	1.68	2.53	3.44	3.45	4.07	3.61	4.07	3.91	1.68	2.53	3.44	3.45	3.61	3.32	3.61	3.91
(20, 20)	1.96	3.13	3.68	3.64	4.28	4.58	4.28	4.30	1.96	3.14	3.68	3.64	3.47	4.07	3.47	3.70
(4, 3)	.79	.82	.79	.82	.79	1.39	.79	1.39	.79	.82	.79	.82	.79	1.39	.79	1.39
(10, 5)	2.18	1.99	2.18	2.31	2.18	2.73	1.56	2.02	2.18	1.99	2.18	2.31	2.18	2.06	1.56	2.02
(10, 7)	2.49	2.53	2.49	2.84	3.13	2.61	2.53	1.36	2.49	2.53	2.49	2.84	2.53	2.61	2.53	2.61
(13, 11)	2.23	3.13	2.39	3.13	3.18	3.52	2.68	2.76	1.85	3.16	2.39	3.16	2.79	3.52	3.00	3.38
(15, 10)	2.05	3.23	2.92	3.23	3.37	3.61	3.37	3.17	2.05	2.60	2.92	3.23	3.37	3.04	3.37	3.04
(15, 13)	2.54	3.75	3.05	3.69	3.96	3.82	3.27	3.16	2.54	3.77	3.05	3.75	3.57	3.82	3.57	3.82
(20, 10)	1.30	2.64	2.50	3.05	3.14	3.72	3.61	3.18	1.36	2.65	2.88	3.06	3.14	3.32	3.09	3.32
(20, 15)	2.34	3.35	2.81	3.58	3.41	3.67	3.01	2.65	2.40	3.36	3.21	3.60	3.41	3.40	3.40	3.32

[a] For $\delta = 0$, average error rates are calculated using nine values of P_1 (.10, .20, .30, .40, .50, .60, .70, .80, .90). For $\delta = .4$, average error rates are calculated using nine values of P_1 (.5, .55, .60, .65, .70, .75, .80, .85, .90).

obtained based on S. By similar argument, the lower bound estimate for outcome $(1/4, 0/4)$ based on Z_1 is larger $(-.389)$. Therefore, the confidence intervals based on Z_1 are shorter than those based on S.

6. Comparison of Performance

In this section, we compare the performance of confidence intervals constructed using these methods in terms of interval length and coverage probabilities. For each given sample size (N_1, N_2), we obtained the 95% exact confidence intervals for all possible outcomes.

Table 2 summarizes the average length of the 95% exact confidence intervals for various sample sizes. For each sample size (N_1, N_2), the average length is calculated by taking the arithmetic mean of the length of all $(N_1+1)(N_2+1)$ confidence intervals. The table also summarizes for each method the percent of confidence intervals achieving the shortest length among the eight methods. These results show that the δ-projected Z statistic (Z_2) without the constrained search $(\gamma = 0)$ produces the shortest confidence intervals on average. In almost every case, the Z_2 method yields the highest percentage of confidence intervals achieving the shortest length among all methods. The improvement provided by Z_2 over the simple statistic (S) is greater in cases with equal sample sizes than in cases with unequal sample sizes. In addition, the reduction of length seems to increase (percentagewise) with larger sample sizes. The binomial Z statistic (Z_1) and the likelihood ratio statistic (LR) also yield shorter confidence intervals than the simple statistic (S), but the improvement is not as great as for Z_2.

It is worth noting that restricting the search of nuisance parameter $(\gamma = .001)$ slightly reduces the average length of confidence intervals based on S and Z_1. A closer look into the individual confidence intervals reveals that the restricted search may benefit the confidence intervals based on S and Z_1 in near-extreme outcomes where the tail probability is likely to peak near either extreme. However, the restricted search does not provide improvement for confidence intervals based on the δ-projected Z_2 and LR statistics. We observed that the tail probabilities based on these two statistics are smooth with peaks typically lying in the middle part of the nuisance parameter domain. As a result, the length of the confidence interval based on Z_2 or LR is often penalized with the restriction $(\gamma = .001)$.

In addition, we summarize the average expected error rates of coverage of the 95% confidence intervals for the true difference (δ). The expected error rate is calculated for a given combination of P_1 and P_2 $(= P_1 - \delta)$ using the binomial probabilities. Then the average is taken as the arithmetic mean of nine values of P_1 given $P_2 = P_1 - \delta$. Table 3 summarizes the results for $\delta = 0$ and .4. The nine values of P_1 are chosen as .10, .20, .30, .40, .50, .60, .70, .80, and .90 for $\delta = 0$ and .5, .55, .60, .65, .70, .75, .80, .85, and .90 for $\delta = .4$. For all sample sizes and combinations of P_1 and P_2 examined, the expected error rate of coverage of the 95% confidence intervals obtained by each of the four methods does not exceed the nominal 5% level. In general, the confidence intervals obtained with the simple statistic (S) are too conservative since their error rates are much lower than expected (5%). In contrast, the Z_2 statistic produces the least conservative confidence intervals among all methods. In general, the degree of conservatism is reduced when the sample size increases. The restricted search of nuisance parameter seems unable to reduce the conservatism of these exact confidence intervals.

7. Conclusion

Overall, we find the exact confidence interval based on the δ-projected Z statistic (Z_2) without restricted search for the nuisance parameter performs best in almost every case. It provides substantial improvement over the method based on the simple statistic (S), and the improvement gets more appreciable as sample sizes increase. The Z_1 statistic and the likelihood ratio statistic (LR) also yield shorter confidence intervals than the simple statistic (S); however, their confidence intervals are generally wider than those based on the δ-projected Z statistic. The search of nuisance parameter over a restricted domain does not offer benefits to the Z_2 or LR statistic as the tail probability often peaks in the middle of the domain of the nuisance parameter.

Résumé

Des intervalles de confiance sont souvent utilisés pour étudier une différence entre traitements. Quand la taille de l'échantillon est faible, comme c'est le cas en particulier dans les phases précoces d'essais cliniques, les intervalles de confiance basés sur les approximations asymptotiques peuvent ne pas être fiables. Nous proposons des méthodes basées sur les tests pour construire des intervalles de confiance exacts de la différence entre deux proportions binomiales. Ces intervalles de confiance exacts sont obtenus à partir de la distribution non conditionnelle de deux réponses binomiales, et ils garantissent le taux de couverture. Nous comparons la performance de ces intervalles de confiance et de ceux basés seulement sur la différence observée (Santner and Snell 1980, statXact 1995). Nous montrons que l'utilisation d'un test Z standardisé avec une estimation de la variance par maximum de vraisemblance contraint peut apporter une grande amélioration.

References

Basu, D. (1977). On the elimination of nuisance parameters. *Journal of the American Statistical Association* **72,** 355.

Berger, R. L. (1996). More powerful tests from confidence interval p-value. *American Statistician* **50,** 314–318.

Berger, R. L. and Boos, D. D. (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* **89,** 1012–1016.

Chan, I. S. F. (1998). Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies. *Statistics in Medicine,* **17,** 1403–1413.

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

Cytel Software. (1995). *StatXact,* Version 3. Cambridge, MA: Cytel Software.

Farrington, C. P. and Manning, G. (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine* **9,** 1447–1454.

Fries, L. F., et al. (1993). Safety and immunogenicity of a recombinant protein influenza A vaccine in adult human volunteers and protective efficacy against wild-type

H1N1 virus challenge. *The Journal of Infectious Disease* **167,** 593–601.

Haber, M. (1986). An exact unconditional test for the 2×2 comparative trials. *Psychological Bulletin* **99,** 129–132.

Miettinen, O. and Nurminen, M. (1985). Comparative analysis of two rates. *Statistics in Medicine* **4,** 213–226.

Rohatgi, V. K. (1984). *Statistical Inference.* 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.

Suissa, S. and Shuster, J. J. (1985). Exact unconditional sample sizes for the 2×2 binomial trial. *Journal of the Royal Statistical Society, Series A* **148,** 317–327.

Received February 1998. *Revised January* 1999.
Accepted January 1999.