5-19-2003

New Intervals for the Difference Between Two Independent Binomial Proportions

Xiao-Hua Zhou
University of Washington, azhou@u.washington.edu

Min Tsao
University of Victoria, tsao@math.uvic.ca

Gengsheng Qin
Georgia State University, gqin@mathstat.gsu.edu

1. INTRODUCTION

Comparisons of two independent binomial proportions are one of most commonly encountered problems in medical studies. However, the most commonly used Wald interval can have poor coverage accuracy. This point has been nicely illustrated by Brown et al. (2001) for the single binomial proportion. Brown et al. (2001) and Brown et al. (2002) have also discussed other types of intervals for the single binomial proportion, including Bayesian credible intervals. In this paper we propose two new methods for constructing confidence intervals for the difference between two binomial proportions based on the Edgeworth expansion of the studentized difference.

Let X_0 and X_1 be two independent random variables with the binomial $\text{Bin}(n_0, p_0)$ and $\text{Bin}(n_1, p_1)$ distributions, respectively; let $p = p_1 - p_0$. Most commonly used confidence interval for p is so called the Wald interval (WA). Let $\hat{p}_i = X_i/n_i$ and $\hat{p} = \hat{p}_1 - \hat{p}_0$. Then, the $100(1-\alpha)\%$ Wald interval is defined by

$$\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_0} + \frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_0} + \frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}\right], \tag{1}$$

where z_α is the α quantile of the standard normal distribution. Even though this interval is very simple to use and has been almost universely adapted in biostatistics textbooks, it has been shown that this interval can behave poorly (Agresti and Caffo, 2000).

Many authors have proposed more complicated alternative intervals that can improve on the Wald interval. For example, Thomas and Gart (1977), Santner and Snell (1980), Santner and Yamagami (1993) and Coe and Tamhane (1993) developed methods for constructing "exact" intervals for p. The coverage probabilities of such confidence intervals are guaranteed to be no less than the desired nominal level, but the computation of these "exact" intervals is complicated and the resulting intervals tend to have wide interval lengths.

To search for computationally simpler intervals, Anbar (1983) and Mee (1984) derived two different asymptotic confidence intervals for p. Newcombe (1998) conducted a comprehensive study on relative advantages of existing asymptotic methods for constructing confidence intervals for p. He recommended a method (hereafter called the Newcombe's hybrid score method) which is based on the score test for a single proportion (Wilson, 1927) and performs substantially better than the Wald interval, while being computationally simpler than the "exact" intervals. Newscombe's

hybrid score interval with the nominal level of $100(1-\alpha)\%$ is defined by

$$\left[\widehat{p} - \left((\widehat{p}_1 - l_1)^2 + (u_0 - \widehat{p}_0)^2\right)^{1/2}, \widehat{p} + \left((u_1 - \widehat{p}_1)^2 + (\widehat{p}_0 - l_0)^2\right)^{1/2}\right],$$

where l_1 and u_1 are the roots of $|p_1 - \widehat{p}_1| = z_{1-\alpha/2}[p_1(1-p_1)/n_1]^{1/2}$, and l_0 and u_0 are the roots of $|p_0 - \widehat{p}_0| = z_{1-\alpha/2}[p_0(1-p_0)/n_0]^{1/2}$. However, the Newcombe's hybrid score method still has two potential drawbacks: (1) its theoretical properties are unknown, and (2) its computation may be too complex for most biostatistics textbooks.

Most recently Agresti and Caffo (2000) proposed an even simpler method than the Newcombe's hybrid score method. This method is a simple adjustment to the Wald interval by adding two successes and two failures, and they showed by a simulation study that their procedure works quite well for two-sample comparisons of binomial proportions when the nominal level is 95%. Let us call their procedure the AC method, and the AC interval is defined by

$$\left[\widetilde{p} - z_{1-\alpha/2}\sqrt{\widetilde{p}_1\widetilde{q}_1/n_1 + \widetilde{p}_0\widetilde{q}_0/n_0}, \quad \widetilde{p} + z_{1-\alpha/2}\sqrt{\widetilde{p}_1\widetilde{q}_1/n_1 + \widetilde{p}_0\widetilde{q}_0/n_0}\right],$$

where $\widetilde{p}_i = (X_i + 1)/(n_i + 2)$, $\widetilde{q}_i = 1 - \widetilde{p}_i$ for $i = 0, 1$, and $\widetilde{p} = \widetilde{p}_1 - \widetilde{p}_0$. One major advantage of the AC method over the other methods lies with its computation and presentation. However, the AC method also has two potential drawbacks. First, it is unknown whether theoretical support exists for their simulation conclusion that their interval has good coverage accuracy. Second, since their proposed method of adding 2 successes and 2 failures was developed specifically for the 95% nominal, it is unclear whether their proposed method will still have good coverage accuracy when the pre-set nominal level is different from 95%.

In this paper we obtain an Edgeworth expansion for the studentized difference between two binomial proportions. Based on the Edgeworth expansion, we propose two new easy to compute confidence intervals for the difference of two binomial proportions. The first interval directly corrects skewness in the Edgeworth expansion and can be thought of as an extension of Hall's (1982) method for the single proportion. The second one corrects the skewness in the Edgeworth expansion through a monotone transformation.

The Edgeworth expansion is also used to study the coverage accuracy of the proposed intervals. We first show that both the intervals have their coverage probabilities converging to the nominal confidence level at the rate of $O(n^{-1/2})$, where n is the size of the combined samples. We then

compare the finite-sample performance of the proposed intervals with the best existing intervals in simulation studies. Simulation results suggest that in finite samples the new interval based on the indirect method has the very similar performance to the best existing intervals in terms of coverage accuracy and average interval length and that the another new interval based on the direct method has the best average coverage accuracy but could have poor coverage accuracy when two true binomial proportions are close to the boundary points. This paper is organized as follows. In Section 2 we give the Edgeworth expansion for the studentized difference. In Section 3 we describe the two new methods based on this expansion. In Section 4 we evaluate the finite-sample performance of the proposed methods and compare them to the usual normal approximation based method, the AC method, and Newcombe's hybrid score method in terms of the coverage probability and the average length of the confidence interval. Theoretical derivations of the Edgeworth expansion and the asymptotic order of the error of the new methods are included in the Appendix. In Section 5 we contrast our methods with the existing methods in three real clinical studies.

2. EDGEWORTH EXAPNSION FOR THE STUDENTIZED DIFFERENCE

Let X_0 and X_1 be two independent binomial random variables with distributions $Bin(n_0, p_0)$ and $Bin(n_1, p_1)$, respectively. Let $q_i = 1 - p_i$ for $i = 0, 1$. The most commonly used interval for $p = p_1 - p_0$ is based on the standard normal approximation to the distribution of the studentized difference in the two sample proportions,

$$T \equiv \frac{\widehat{p} - p}{\sqrt{\widehat{p}_1\widehat{q}_1/n_1 + \widehat{p}_0\widehat{q}_0/n_0}}, \tag{2}$$

where $\widehat{p}_i = X_i/n_i$, $\widehat{q}_i = 1 - \widehat{p}_i$ for $i = 0, 1$ and $\widehat{p} = \widehat{p}_1 - \widehat{p}_0$.

The normal approximation is a rather crude approximation, especially when sample sizes are not large; it does not take into consideration the skewness of the underlying distribution which is often the main source of error of the normal approximation. To see the impact of the skewness, we develop the Edgeworth expansion for T. To state this Edgeworth expansion we need the following notation. Let $R_n(p_0, p_1, t)$ be a periodic function and has a range of $[-0.5, 0.5]$. Define δ, σ, a, and b to be

$$\delta = \left(\frac{n}{n_1}\right)^2 p_1q_1(1 - 2p_1) - \left(\frac{n}{n_0}\right)^2 p_0q_0(1 - 2p_0),$$

$$\sigma = \left(\frac{n}{n_1}p_1q_1 + \frac{n}{n_0}p_0q_0\right)^{1/2}, a = \frac{\delta}{6\sigma^2}, \text{ and } b = \frac{n(1-2p_1)}{2n_1} - \frac{\delta}{6\sigma^2},$$

respectively. Define $Q(t) = \sigma^{-1}(a + bt^2)$, and $n = n_0 + n_1$. Now we can state the Edgeworth expansion for T as follows.

Theorem 1 *Assume that p_0 and p_1 are rational numbers, $\min(n_0, n_1) \longrightarrow \infty$, and $n_1 = O(n_0)$. Then,*

$$\begin{aligned} P(T \leq t) &= \Phi(t) + n^{-1/2}Q(t)\phi(t) \\ &\quad + \left(n\sigma^2\right)^{-1/2} R_n(p_0, p_1, t)\phi(t) + O\left(n^{-1}loglogn\right), \end{aligned} \tag{3}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cdf and the pdf of the standard normal distribution respectively.

In the Edgeworth expansion (3), $Q(t)$ represents the error due to the skewness of the binomial distributions, and $R_n(p_0, p_1, t)$ represents the rounding error. The proof of Theorem 1 is given in the Appendix. It is worthwhile to note that the reminder term in our Edgeworth expansion is at rate of $n^{-1}\log\log n$, which is larger than the rate for the one-sample binomial case.

From Theorem 1 we see that if δ is close to 0 (which may happen when p is near 0, or both p_0 and p_1 are near boundary point 0 and 1), then the main part of $\sigma Q(t)$ is $n(1-2p_1)t^2/(2n_1)$ which is larger than the rounding error $|R_n(p_0, p_1, t)|$ if $p_1 > (1 + c_0)/2$ or $p_1 < (1 - c_0)/2$ where $c_0 = 1/((1 + n_0/n_1)t^2)$.

3. TWO NEW CONFIDENCE INTERVALS

We propose two intervals by eliminating the error due to the skewness in the Edgeworth expansion of T given in Theorem 1. The first approach directly eliminates this error from the Edgeworth expansion, as suggested in Hall (1982). The resulting two-sided $100(1-\alpha)\%$ skewness-corrected confidence interval for p is defined as follows:

$$I_{1\alpha} = \left[\widehat{p} - \left(\frac{\widehat{p}_1\widehat{q}_1}{n_1} + \frac{\widehat{p}_0\widehat{q}_0}{n_0}\right)^{1/2}\left(z_{1-\alpha/2} - n^{-1/2}\widehat{Q}(z_{1-\alpha/2})\right),\right.$$

$$\left.\widehat{p} - \left(\frac{\widehat{p}_1\widehat{q}_1}{n_1} + \frac{\widehat{p}_0\widehat{q}_0}{n_0}\right)^{1/2}\left(z_{\alpha/2} - n^{-1/2}\widehat{Q}(z_{\alpha/2})\right)\right],$$

where $\widehat{Q}(t) = \widehat{\sigma}^{-1}\left(\widehat{a} + \widehat{b}t^2\right)$. Here $\widehat{a}$, $\widehat{b}$, $\widehat{\sigma}$, and $\widehat{\delta}$ are estimates of a, b, σ, and δ, respectively. They are computed by replacing the p_i's in the formulas for a, b, σ, and δ with the $\widehat{p}_i$'s.

Another method for removing the skewness is to use a monotone transformation of T, derived from the Edgeworth expansion. This method was originally introduced by Hall (1992) for removing the skewness of a statistic in an one-sample setting. The monotone transformation is defined by (see Hall,1992)

$$g(T) = n^{-1/2}\widehat{a}\widehat{\sigma} + T + n^{-1/2}\left(\widehat{b}\widehat{\sigma}\right)T^2 + n^{-1}\cdot\frac{1}{3}\left(\widehat{b}\widehat{\sigma}\right)^2 T^3,$$

where $\widehat{\sigma} = \{(n/n_1)\cdot\widehat{p}_1\widehat{q}_1 + (n/n_0)\cdot\widehat{p}_0\widehat{q}_0\}^{1/2}$. Using this transformation, we can construct another two-sided $100(1-\alpha)\%$ confidence interval for p,

$$I_{2\alpha} = \left[\widehat{p} - \left(\frac{\widehat{p}_1\widehat{q}_1}{n_1} + \frac{\widehat{p}_0\widehat{q}_0}{n_0}\right)^{1/2} g^{-1}(z_{1-\alpha/2}), \widehat{p} - \left(\frac{\widehat{p}_1\widehat{q}_1}{n_1} + \frac{\widehat{p}_0\widehat{q}_0}{n_0}\right)^{1/2} g^{-1}(z_{\alpha/2})\right],$$

where

$$g^{-1}(T) = n^{1/2}\left(\widehat{b}\widehat{\sigma}\right)^{-1}\left\{\left(1 + 3\left(\widehat{b}\widehat{\sigma}\right)\left(n^{-1/2}T - n^{-1}\widehat{a}\widehat{\sigma}\right)\right)^{1/3} - 1\right\}.$$

The following theorem gives the asymptotic coverage probabilities of the two proposed intervals. The proof for this result is given in the Appendix.

Theorem 2

$$P(p \in I_{k\alpha}) \;=\; 1 - \alpha + O(n^{-1/2}), \quad k = 1, 2.$$

4. A NUMERICAL STUDY

In this section, we conduct a numerical study to assess the finite-sample performance of the two newly proposed intervals, the direct Edgeworth expansion method, denoted by EE, and the transformation method, denoted by TT. In the numerical study we also compare their performance with the two of the better existing methods on the basis of coverage probability and expected length, Newscombe's hybrid score method (NH) and the AC method, as well as the commonly used Wald interval (WA). To compare the relative performance of EE, TT, NH, AC, and WA intervals for $p = p_1 - p_0$, we compute their coverage probabilities and the average lengths. For fixed values of (p_0, p_1) and (n_0, n_1), we let $C_{n_0,n_1}(p_0, p_1)$ and $W_{n_0,n_1}(p_0, p_1)$ denote the coverage probability and the expected length of a two-sided $(1-\alpha)\%$ level confidence interval $\mathcal{L}(X_0, X_1)$ for $p = p_1 - p_0$, given n_0, n_1, p_0, and p_1, respectively. Then,

$$C_{n_0,n_1}(p_0, p_1) = E\{I_{[p_1-p_0\in\mathcal{L}(x_0,x_1)]} \mid n_0, n_1, p_0, p_1\} = \sum_{x_0=0}^{n_0}\sum_{x_1=0}^{n_1} bin(x_0; n_0, p_0)bin(x_1, n_1, p_1)I_{[p\in\mathcal{L}(x_0,x_1)]}, \tag{4}$$

where $I_{[p\in\mathcal{L}(x_0,x_1)]}$ is 1 if $p \in \mathcal{L}(x_0, x_1)$ and zero otherwise, and $bin(x_k; n_k, p_k)$ is the binomial probability when $X_k = x_k$. Denote the lower and upper endpoints of $\mathcal{L}(x_0, x_1)$ to be $lower(x_0, x_1)$ and $upper(x_0, x_1)$, respectively. Then, the expected interval length for $\mathcal{L}(x_0, x_1)$ is calculated using the formula,

$$W_{n_0,n_1}(p_0, p_1) = \sum_{x_0=0}^{n_0} \sum_{x_1=0}^{n_1} \{upper(x_0, x_1) - lower(x_0, x_1)\} bin(x_0; n_0, p_0) bin(x_1; n_1, p_1).$$

We first compare the performance of the five intervals for fixed values of $p = p_1 - p_0$ as p_1 varies on (0,1). In Figures 1-3, we plot the coverage probability $C_{n_0,n_1}(p_0, p_1)$ for the five intervals, p_1 varying over the points given by $0.05 + 0.02j$ for $j = 0, 1, \cdots, 45$ as p fixed at 0, p_1 varying over the points given by $0.5 + 0.01j$ for $j = 0, 1, \cdots, 45$ as p fixed at 0.4, and p_1 varying over the points given by $0.85 + 0.002j$ for $j = 0, 1, \cdots, 50$ as p fixed at 0.8, for $(n_1, n_0) = (15, 15), (30, 30)$, and $(30, 15)$, respectively.

FIGURES 1-3 GO HERE

Tables 1-3 summarize the average coverage probability of three nominal levels confidence intervals for fixed values of $p = p_1 - p_0$, averaging with respect to p_1's. Table 4 presents the average length of the confidence intervals for fixed $p = p_1 - p_0$, averaging with respect to p_1's.

TABLES 1-4 GO HERE

We then compare the performance of the five intervals in three averaging performance measures of $C_{n_0,n_1}(p_0, p_1)$ and $W_{n_0,n_1}(p_0, p_1)$ over the 10000 randomly chosen values of p_0 and p_1 from the unit square [0,1]x[0,1]. The first two measures are the average coverage probability and average expected length, which are defined by

$$\int_0^1 \int_0^1 C_{n_0,n_1}(p_0, p_1) dp_0 dp_1, \text{ and } \int_0^1 \int_0^1 W_{n_0,n_1}(p_0, p_1) dp_0 dp_1,$$

respectively; the last one is the proportion of the chosen values of p for which the coverage probability of the nominal 90% interval falls below 0.88, which is defined by

$$\frac{\#\text{ of 10,000 pairs } (p_0, p_1) : C_{n_0,n_1}(p_0, p_1) < 0.88}{10,000}.$$

Since averaging performance measures do not provide information on effects of particular values of p_0 and p_1 on the coverage probability and expected interval length, we also plot $C_{n_0,n_1}(p_0, p_1)$

as functions of p_0 and p_1 for the EE, TT, NH, and AC intervals when (n_0, n_1)=(15, 15) and (30, 30), respectively. The statistic T is undefined when (X_0, X_1) is $(0, 0)$, $(0, n_1)$, $(n_0, 0)$ or (n_0, n_1). In our study, we replace X_k by $X_k + 0.5$ and n_k by $n_k + 1$ for $k = 1, 0$. This is motivated by a similar technique used by Agresti and Coull (1998).

Table 5 displays the summary performances of the five intervals.

TABLE 5 GOES HERE

Figures 4-5 display the coverage probabilities of the four intervals as functions of p_0 and p_1 over a grid of points given by (p_0, p_1)=$(0.02i, 0.02j)$ for $i, j = 0, 1, \ldots, 50$ when $(n_0, n_1) = (15, 15)$ and $(30, 30)$, respectively.

FIGURES 4-5 GO HERE

From the results on the summary measures in Tables 1-5, we conclude that the two new intervals and the two best existing intervals all have good coverage accuracy and are superior to the Wald interval. Among the four good intervals, the direct Edgeworth expansion method has the best average coverage accuracy, closely followed by the Newscombe's hybrid score method and the transformation method, and then by the AC method. However, when looking at effects of particular values of p_0 and p_1 on the coverage accuracy in Figures 1-5, we see that the direct Edgeworth expansion method can have the poor coverage accuracy when p_0 and p_1 are near 0 or 1. The transformation method still has very similar coverage accuracy to those of the existing methods.

5. REAL EXAMPLES

In this section, we contrast our methods with the existing methods in three real datasets.

5.1 A study on prostate cancer

Tempany et al (1994) conducted a study on the accuracy of conventional magnetic resonance imaging (MRI) in detecting advanced stage prostate cancer (Tempany et al, 1994). This study was a multi-center trial. We are interested in assessing whether the sensitivity of the conventional MRI is the same between two hospitals. Sensitivity of a test is defined as the probability of giving a positive result in a patient with the advanced stage prostate cancer. We summarize the data in Table 2.

TABLE 6 GOES HERE

Let p_1 be the sensitivity of the MRI among the patients in hospital 1 and p_0 be the sensitivity of the MRI among the patients in hospital 2. Using the methods described in this paper, we derived 95% confidence intervals for $p_1 - p_0$. The resulting intervals are $[-0.361, 0.074]$ using the direct Edgeworth expansion method, $[-0.361, 0.074]$ using the transformation method, $[-0.364, 0.076]$ using the Wald method, $[-0.347, 0.074]$ using the Newscombe's hybrid score method, and $[-0.353, 0.077]$ using the Agresti and Caffo method. Although there is some difference among these four intervals, they point to the same conclusion that there is no statistical difference between two proportions. It is worth to point out that although the Wald interval in this example has the similar length as the other methods, in general it has a shorter length than the two new methods.

5.2 A study on sudden infant death syndrome (SIDS) children

Fisher and Van Belle (1993) reported a study by Peterson et al (1980) on the effect of the genetic component on sudden infant death syndrome (SIDS). In the study, two groups of twins with at least one SIDS child were examined to see whether both twins died during the study period. In the one group, all twins are identical ones, and in the another group all twins are fraternal ones. We summarize the data in Table 7.

TABLE 7 GOES HERE

Let p_1 be the probability that both twins died for an identical twin and p_0 be the probability that both twins died for an fraternal twin. Using the methods described in this paper, we derived 95% confidence intervals for $p_1 - p_0$. The resulting intervals are $[0.005, 0.516]$ using the direct Edgeworth expansion method, $[-0.024, 0.544]$ using the transformation method, $[-0.081, 0.426]$ using the Wald method, $[-0.011, 0.483]$ using the Newscombe's hybrid score method, and $[-0.058, 0.452]$ using the Agresti and Caffo method. The direct Edgeworth expansion method gives an opposite conclusion than the other methods. Since the observed proportions are 0.1 and 0.03, respectively, we may assume that p_0 is close to 0.0. From the simulation results, we know that in this case, the transformation method produces a better confidence interval than the direct Edgeworth method. Therefore, we would use $[-0.024, 0.544]$ as our 95% confidence interval for $p_1 - p_0$.

5.3 A vaccine example

To illustrate the conservativeness of an "exact" confidence interval for $p_1 - p_0$, we used the data from a vaccine trial to compute the one commonly used "exact" interval that was proposed by Santner and Snell (1980) and implemented by Cytel software in its 3 verion of StatXact. This example also illustrates that the Wald interval produces a slightly different interval. We summarize the data in Table 8.

TABLE 8 GOES HERE

The 95% confidence interval for $p_1 - p_0$ is $[0.046, 0.467]$ using the direct Edgeworth expansion method, $[0.051, 0.497]$ using the transformation method, $[0.125, 0.542]$ using the Wald method, and $[-0.019, 0.629]$ using the "exact" interval method. From these intervals, we see that the exact interval has the longest length and that the Wald interval has the smallest length. The result from the exact method is different from the other methods. Although the Wald method leads to the same conclusion of no statistical difference as the two new methods, it produces a lower endpoint that is much larger than the onses given by the two new methods.

6. DISCUSSION

Agresti and Caffo (2000) have shown by simulation that the standard Wald interval for the difference in two binomial proportions has poor coverage accuracy. In this paper, we first derived Edgeworth expansion for Studentized t statistics. We then derived two new confidence intervals for the difference in the two binomial proportions. The newly proposed methods share the same good property of being computational simple as the two of the better existing intervals. However, unlike the two of the existing intervals, we have shown that the proposed intervals also have a sound theoretical property that their coverage probabilities converge to the nominal level at the rate of $O(n^{-1/2})$. Our simulation study suggests one of the two proposed method, the transformation method, has similar coverage accuracy and length with the two best existing intervals. The other one has the best average coverage accuracy over 10,000 values of (p_0, p_1) from [0,1]x[0,1], but has the worst coverage accuracy when p_0 and p_1 are are close to the boundary points. Among the two newly proposed methods, we recommend the direct Edgeworth corrected interval (EE) if p_0 and p_1 are not close to the boundary points; otherwise we recommend the transformation interval (TT).

Although our two new intervals have much better coverage accuracy than the Wald interval, they do not have much improvement over the best existing intervals. However, it is worth noting that our methods for the problem of two-sample interval estimation are based on general transformation and skewness correction techniques whereas the others are specifically targeted at this problem. Thus, our successful application of these two general techniques to the problem of two-sample interval estimation adds further credibility to these general techniques. This result naturally leads to a future research topic that is whether it is possible to use the transformation and skewness correction methods for other problems where the Wald interval performs poorly, such as for the odds ratio.

ACKNOWLEDGMENTS

We would like to thank one referee and associate editor for their helpful comments that results in an improved version of the manuscript.

APPENDIX

Proof of Theorem 1:

To derive the Edgeworth expansion for the studentized sample difference T, as stated in Theorem 1, we first derive the Edgeworth expansion for the standardized sample difference, T_n, to be defined below. Note that for each $i = 0, 1$, we can write $X_i = \sum_{k=1}^{n_i} X_{ik}$ where X_{ik}'s are i.i.d. Bernoulli random variables with parameter p_i. Then the standardized sample difference is defined as follows.

$$T_n \equiv \frac{\widehat{p} - p}{\sqrt{p_1 q_1/n_1 + p_0 q_0/n_0}} = \sum_{k=1}^{n} \frac{D_k}{\sqrt{n}\sigma}$$

where

$$D_k = \begin{cases} -(1 + n_1/n_0) \cdot (X_{0k} - p_0), & k = 1, 2, \cdots, n_0, \\ (1 + n_0/n_1) \cdot (X_{1k} - p_1), & k = n_0 + 1, n_0 + 2, \cdots, n. \end{cases}$$

Our derivation of the Edgeworth expansion for T_n is different from that in Hall (1982) for one sample binomial proportion because T_n is no longer a sum of i.i.d. discrete random variables but is a sum of independent discrete random variables with different distributions. To derive the

Edgeworth expansion for T_n we will use a result by Kolassa (1995, page 170) on the Edgeworth expansion for the sum of independent but nonidentically distributed random variables supported on the same lattice. Kolassa's result was originally developed for the Edgeworth expansion of the rank sum test statistics.

To apply the Kolassa's result to our setting, we need to show that the D_k's are independent random variables supported on the same lattice. Since p_0 and p_1 are rational, we can take a positive integer l large enough such that $l(1+n_1/n_0)$, $l(1+n_0/n_1)$, $l(1+n_1/n_0)p_0$ and $l(1+n_0/n_1)p_1$ are integers. Let $\Delta = 1/l$ and let A be a constant such that A/Δ is an integer. Also let

$$\begin{aligned} k_1 &= (1+n_1/n_0)p_0/\Delta - A/\Delta, \qquad k_2 = (1+n_1/n_0)p_0/\Delta - ((1+n_1/n_0)/\Delta + A/\Delta), \\ k_3 &= -(1+n_0/n_1)p_1/\Delta - A/\Delta, \quad \text{and} \quad k_4 = -(1+n_0/n_1)p_1/\Delta + ((1+n_0/n_1)/\Delta - A/\Delta), \end{aligned}$$

then $\{(1+n_1/n_0)p_0, -(1+n_1/n_0)(1-p_0), -(1+n_0/n_1)p_1, (1+n_0/n_1)(1-p_1)\} = \{A+k_1\Delta, A+k_2\Delta, A+k_3\Delta, A+k_4\Delta\}$ fall in the lattice $\{A+\Delta\mathbf{Z}\} = \{..., A-2\Delta, A-\Delta, A, A+\Delta, A+2\Delta, ...\}$. Thus the D_k's are all constrained to the same lattice $\{A+\Delta\mathbf{Z}\}$. Further, they are independent with mean zero and finite variances. Also, it is not difficult to show that T_n has mean zero and variance 1, and its third and fourth cumulants are

$$\kappa_3 = \frac{1}{\sqrt{n}\sigma^3}\left[\left(\frac{n}{n_1}\right)^2 p_1q_1(1-2p_1) - \left(\frac{n}{n_0}\right)^2 p_0q_0(1-2p_0)\right] \equiv \frac{\delta}{\sqrt{n}\sigma^3}$$

and

$$\kappa_4 = \frac{1}{n\sigma^4}\left[\left(\frac{n}{n_0}\right)^3\left(E(X_{01}-p_0)^4 - 3p_0^2q_0^2\right) + \left(\frac{n}{n_1}\right)^3\left(E(X_{11}-p_1)^4 - 3p_1^2q_1^2\right)\right]$$

respectively. By the theorem in Kolassa (1995, page 170), we obtain that T_n has the following Edgeworth expansion:

$$\begin{aligned} P\left(T_n \le t\right) &= \Phi(t) + \left(n\sigma^2\right)^{-1/2}\cdot\frac{\delta}{6\sigma^2}\left(1-t^2\right)\phi(t) \\ &\quad + \left(n\sigma^2\right)^{-1/2} R_{n_0}(p_0,p_1,t)\phi(t) + O\left(n^{-1}\right) \end{aligned} \tag{5}$$

where $R_{n_0}(p_0,p_1,t)$ is a function taking values in $[-0.5, 0.5]$ and represents the rounding error, whose exact form can be found in Kolassa (1995, page 170). Next we use the Edgeworth expansion for T_n to obtain an Edgeworth expansion for T. Note that

$$P(T \le t) = P\left(\frac{\widehat{p}-p}{\sqrt{\left((\widehat{p}-p)+(\widehat{p}_0+p)\right)\left(1-\left((\widehat{p}-p)+(\widehat{p}_0+p)\right)\right)/n_1 + \widehat{p}_0\widehat{q}_0/n_0}} \le t\right).$$

By solving the inequality for $\widehat{p}-p$ in the right side of the above equation, we obtain that

$$P(T \leq t) \;=\; P\left(T_n \leq \widetilde{t}_0\right), \tag{6}$$

where

$$\begin{aligned}\widetilde{t}_0 \;=\; & \left(\frac{1}{p_1q_1/n_1+p_0q_0/n_0}\right)^{1/2}\left(\frac{\left(1-2\widehat{p}_0-2p\right)t^2}{2\left(n_1+t^2\right)}\right.\\ & \left.+\frac{(n/n_1)^{1/2}t\left[4\left(p(q-2\widehat{p}_0)+(1+n_1/n_0)\widehat{p}_0\widehat{q}_0\right)/n+t^2\left(1+4n_1\widehat{p}_0\widehat{q}_0/n_0\right)/(n_1n)\right]^{1/2}}{2\left(1+t^2/n_1\right)}\right).\end{aligned}$$

Let us define t_0 to be the $\widetilde{t}_0$ except that $\widehat{p}_0$ and $\widehat{q}_0$ are replaced by p_0 and q_0 respectively, i.e.,

$$\begin{aligned}t_0 \;=\; & \left(\frac{1}{p_1q_1/n_1+p_0q_0/n_0}\right)^{1/2}\left(\frac{\left(1-2p_1\right)t^2}{2\left(n_1+t^2\right)}\right.\\ & \left.+\frac{t\left[4\left(p_1q_1/n_1+p_0q_0/n_0\right)+t^2\left(1+4p_0q_0n_1/n_0\right)/n_1^2\right]^{1/2}}{2\left(1+t^2/n_1\right)}\right).\end{aligned}$$

Then,

$$\begin{aligned}P\left(T_n \leq \widetilde{t}_0\right) \;&=\; P\left(T_n \leq t_0\right)+\left(P\left(T_n \leq \widetilde{t}_0\right)-P\left(T_n \leq t_0\right)\right)\\ &\equiv\; I_1+I_2.\end{aligned} \tag{7}$$

The Edgeworth expansion (5) may be used to obtain an expansion for I_1. We have, after some algebra, that

$$\begin{aligned}I_1 \;&=\; \Phi(t_0)+\left(n\sigma^2\right)^{-1/2}\cdot\frac{\delta}{6\sigma^2}\left(1-t_0^2\right)\phi(t_0)\\ &\quad+\left(n\sigma^2\right)^{-1/2}R_{n_0}(p_0,p_1,t_0)\phi(t_0)+O\left(n^{-1}\right)\\ &=\; \Phi(t)+\left(n\sigma^2\right)^{-1/2}\left(a+bt^2\right)\phi(t)\\ &\quad+\left(n\sigma^2\right)^{-1/2}R_n(p_0,p_1,t)\phi(t)+O\left(n^{-1}\right).\end{aligned} \tag{8}$$

Now we show that $I_2=O\left(n^{-1}loglogn\right)$. By $\widehat{p}_0-p_0=O\left(n^{-1/2}loglogn\right)$ $a.s.$, we can find a positive constant C such that $|\widetilde{t}_0-t_0| \leq C\left(n^{-1}loglogn\right)$ $a.s.$. That is, the interval formed by $\widetilde{t}_0$ and t_0 is contained by $\left(t_0-C\left(n^{-1}loglogn\right), t_0+C\left(n^{-1}loglogn\right)\right]$ $a.s.$. It follows from (5) that

$$\begin{aligned}|I_2| \;&\leq\; P\left(T_n \leq t_0+C\left(n^{-1}loglogn\right)\right)-P\left(T_n \leq t_0-C\left(n^{-1}loglogn\right)\right)\\ &=\; O\left(n^{-1}loglogn\right)\end{aligned} \tag{9}$$

Theorem 1 then follows from (7)–(9).

Note the remainder term in the Edgeworth expansion (3) has the same rate as that of $P\left(T_n \leq \widetilde{t}_0\right) - P\left(T_n \leq t_0\right)$ in (9) whereas that for the Edgeworth expansion for a single studentized sample proportion has a rate of $O\left(n^{-1}\right)$ (Hall, 1982). This is because in the one-sample case we do not need to consider this difference. In the two-sample case, however, $\widetilde{t}_0$ is involved in $\widehat{p}_0$ which has a convergence rate of $O\left(n^{-1/2}loglogn\right)$ with probability one, the remainder term has a rate of $O\left(n^{-1}loglogn\right)$.

Proof of Theorem 2:

First we show that

$$P(p \in I_{1\alpha}) \;=\; 1 - \alpha + O\left(n^{-1/2}\right).$$

For any $0 < \alpha < 1$, we have

$$\begin{aligned} & P\left(T \leq z_\alpha - n^{-1/2}\widehat{Q}\left(z_\alpha\right)\right) \\ & \quad = P\left(T \leq z_\alpha - n^{-1/2}Q\left(z_\alpha\right)\right) \\ & \qquad + \left[P\left(T \leq z_\alpha - n^{-1/2}\widehat{Q}\left(z_\alpha\right)\right) - P\left(T \leq z_\alpha - n^{-1/2}Q\left(z_\alpha\right)\right)\right] \\ & \quad \equiv J_1 + J_2. \end{aligned}$$

Noting that $\Phi(x)$, $\phi(x)$ and $q_1(x)$ are smooth functions of x, by Theorem 1 and Taylor expansion, we obtain that

$$\begin{aligned} J_1 \;&=\; \Phi\left(z_\alpha - n^{-1/2}Q(z_\alpha)\right) + n^{-1/2}Q\left(z_\alpha - n^{-1/2}Q(z_\alpha)\right)\phi\left(z_\alpha - n^{-1/2}Q(z_\alpha)\right) \\ &\quad + \left(n\sigma^2\right)^{-1/2} g_n\left(p_0, p_1, z_\alpha - n^{-1/2}Q(z_\alpha)\right)\phi\left(z_\alpha - n^{-1/2}Q(z_\alpha)\right) + O\left(n^{-1}loglogn\right) \\ &=\; \Phi\left(z_\alpha\right) + O\left(n^{-1/2}\right) = \alpha + O\left(n^{-1/2}\right). \end{aligned}$$

For the term J_2, by $\widehat{p}_i - p_i = O\left(n^{-1/2}loglogn\right), a.s.$, we can get

$$\widehat{Q}(z_\alpha) - Q(z_\alpha) = o\left(1\right), a.s.$$

Hence by Theorem 1,

$$\begin{aligned} J_2 \;&=\; P\left\{z_\alpha - n^{-1/2}Q(z_\alpha) < T \leq z_\alpha - n^{-1/2}Q(z_\alpha) - n^{-1/2}\left(\widehat{Q}(z_\alpha) - Q(z_\alpha)\right)\right\} \\ &\leq\; P\left\{z_\alpha - n^{-1/2}Q(z_\alpha) < T \leq z_\alpha - n^{-1/2}Q(z_\alpha) + Cn^{-1/2}\right\} \\ &=\; Cn^{-1/2}\phi\left(z_\alpha - n^{-1/2}Q(z_\alpha)\right) + O\left(n^{-1/2}\right) = O\left(n^{-1/2}\right). \end{aligned}$$

Therefore,

$$\begin{aligned} P(p \in I_{1\alpha}) &= P\left(T \le z_{1-\alpha/2} - n^{-1/2}\widehat{Q}\left(z_{1-\alpha/2}\right)\right) - P\left(T \le z_{\alpha/2} - n^{-1/2}\widehat{Q}\left(z_{\alpha/2}\right)\right) \\ &= 1 - \alpha + O\left(n^{-1/2}\right). \end{aligned} \tag{10}$$

Now we show that

$$P(p \in I_{2\alpha}) = 1 - \alpha + O\left(n^{-1/2}\right).$$

Using a Taylor expansion on the function $(1+y)^{1/3}$, we get

$$\begin{aligned} &\left[1 + 3\left(\widehat{b}\widehat{\sigma}\right)\left(n^{-1/2}x - n^{-1}\widehat{a}\widehat{\sigma}\right)\right]^{1/3} - 1 \\ &\quad = n^{-1/2}\left(\widehat{b}\widehat{\sigma}\right)x - n^{-1}\left(\widehat{b}\widehat{\sigma}\right)\left[\left(\widehat{a}\widehat{\sigma}\right) + \left(\widehat{b}\widehat{\sigma}\right)x^2\right] + O_p\left(n^{-3/2}\right), \end{aligned}$$

hence we have

$$g^{-1}(x) = x - n^{-1/2}\widehat{Q}(x) + O\left(n^{-1}\right).$$

An argument similar to the proof of (10) leads to $P(p \in I_{2\alpha}) = 1 - \alpha + O\left(n^{-1/2}\right)$. The proof of Theorem 2 is thus completed.

REFERENCES

Agresti, A. and Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportion. *The American Statistician*, **52**, 119-126.

Agresti, A. and Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician*, **54**, 280-288.

Anbar, D. (1983). On estimating the difference between two probabilities, with special reference to clinical trials. *Biometrics*, **39**, 257-262.

Brown, L. D., Cai, T. T. and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, **16**, 101-133.

Brown, L. D., Cai, T. T. and DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *Ann. of Statist.*, **30**, 160-201.

Coe, P. R. and Tamhane, A. C. (1993). Small sample confidence intervals for the difference, ratio and odds ration of two success probabilities. *Commun. in Statist.–Simula.*, **22**, 925-938.

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

Fisher, L. D. and Van Belle, G. (1993). *Biostatistics: A methodology for the health sciences* . New York, U.S.A.: Wiley & Sons.

Hall, P. (1982). Improving the normal approximation when constructing one-side confidence intervals for binomial or Poisson parameters. *Biometrika,* **69**, 647-652.

Hall, P. (1992). On the removal of skewness by transformation. *J. Roy. Statist. Soc.*, **B 54**, 221-228.

Hall, P. (1992). The Bootstrap and Edgeworth Expansion *Springer*, New York.

Kolassa, J. E. (1995). Edgeworth approximations for rank sum test statistics. *Statist. & Probab. Lett.*, **24**, 169-171.

Mee, R. W. (1984). Confidence bounds for the difference between two probabilities (letter). *Biometrics*, **40**, 1175-1176.

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*, **17**, 873-890.

Peterson, D. R., Chinn, N. M., and Fisher, L. D. (1980). The sudden infant death syndrome: repetitions in families. *Journal of Pediatrics*, **97**, 265-267.

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 continence tables. *J. Amer. Statist. Assoc.*, **75**, 386-394.

Santner, T. J. and Yamagami, S. (1993). Invariant small sample confidence intervals for the difference of two success probabilities. *Commun. in Statist.–Simula.*, **22**, 33-59.

Thomas, D. G. and Gart, J. J. (1977). A table of exact confidence limits for differences and ratios of two proportions and their odd ratios. *J. Amer. Statist. Assoc.*, **72**, 73-76.

Tempany, C. M., Zhou, X. H., Zerhouni, E. A., et al (1994). Staging of prostate cancer with MRI: the results of Radiology Diagnostic Oncology Group project: comparison of different techniques, including the endorectal coil. *Radiology*, **192**, 47-54.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *J. Amer. Statist. Assoc.*, **22**, 209-212.

Table 1. Average coverage probability of nominal 90% confidence intervals for fixed $p = p_1 - p_0$, averaging with respect to p_1's.

p	(n_1, n_0)	EE	TT	NH	AC	WA
0.0	(15,15)	89.64	92.51	91.12	92.64	85.97
	(30,30)	90.02	91.02	90.51	91.61	88.89
	(30,15)	90.05	91.76	90.52	91.97	86.08
0.4	(15,15)	90.28	91.34	89.08	89.39	88.49
	(30,30)	89.82	90.19	89.69	90.60	89.65
	(30,15)	89.54	89.76	89.82	90.02	87.74
0.8	(15,15)	90.53	88.61	92.78	92.78	79.25
	(30,30)	90.40	88.56	91.44	91.44	83.03
	(30,15)	91.92	89.59	90.74	90.60	83.24

Note:

When $p = 0$, p_1 varies over the points given by $0.05 + 0.02j$ for $j = 0, 1, \cdots, 45$.

When $p = 0.4$, p_1 varies over the points given by $0.5 + 0.01j$ for $j = 0, 1, \cdots, 45$.

When $p = 0.8$, p_1 varies over the points given by $0.85 + 0.002j$ for $j = 0, 1, \cdots, 50$.

Figure 1: Coverage probability of the various confidence intervals for $p = p_1 - p_0$

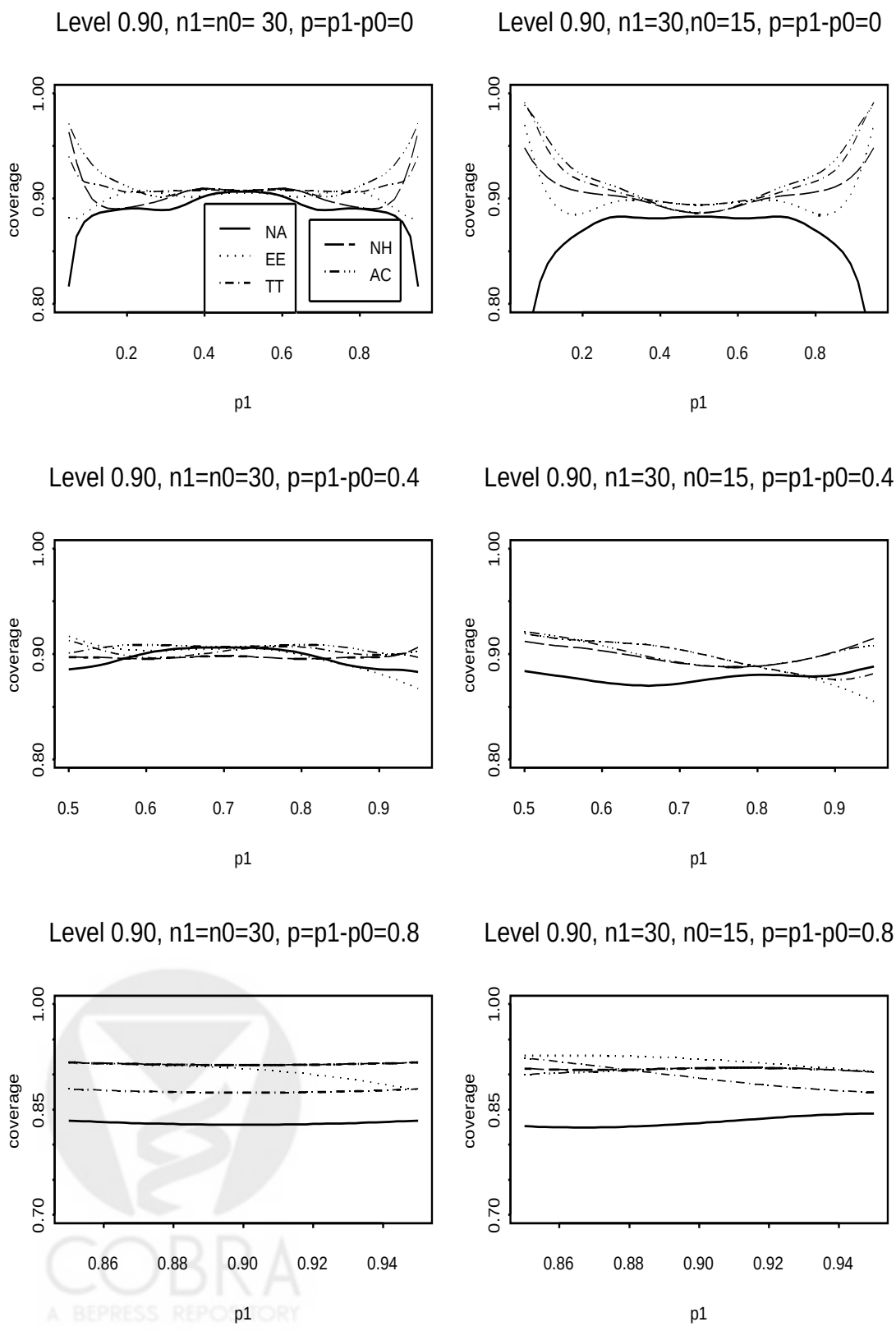

Figure 2: Coverage probability of the various confidence intervals for $p = p_1 - p_0$

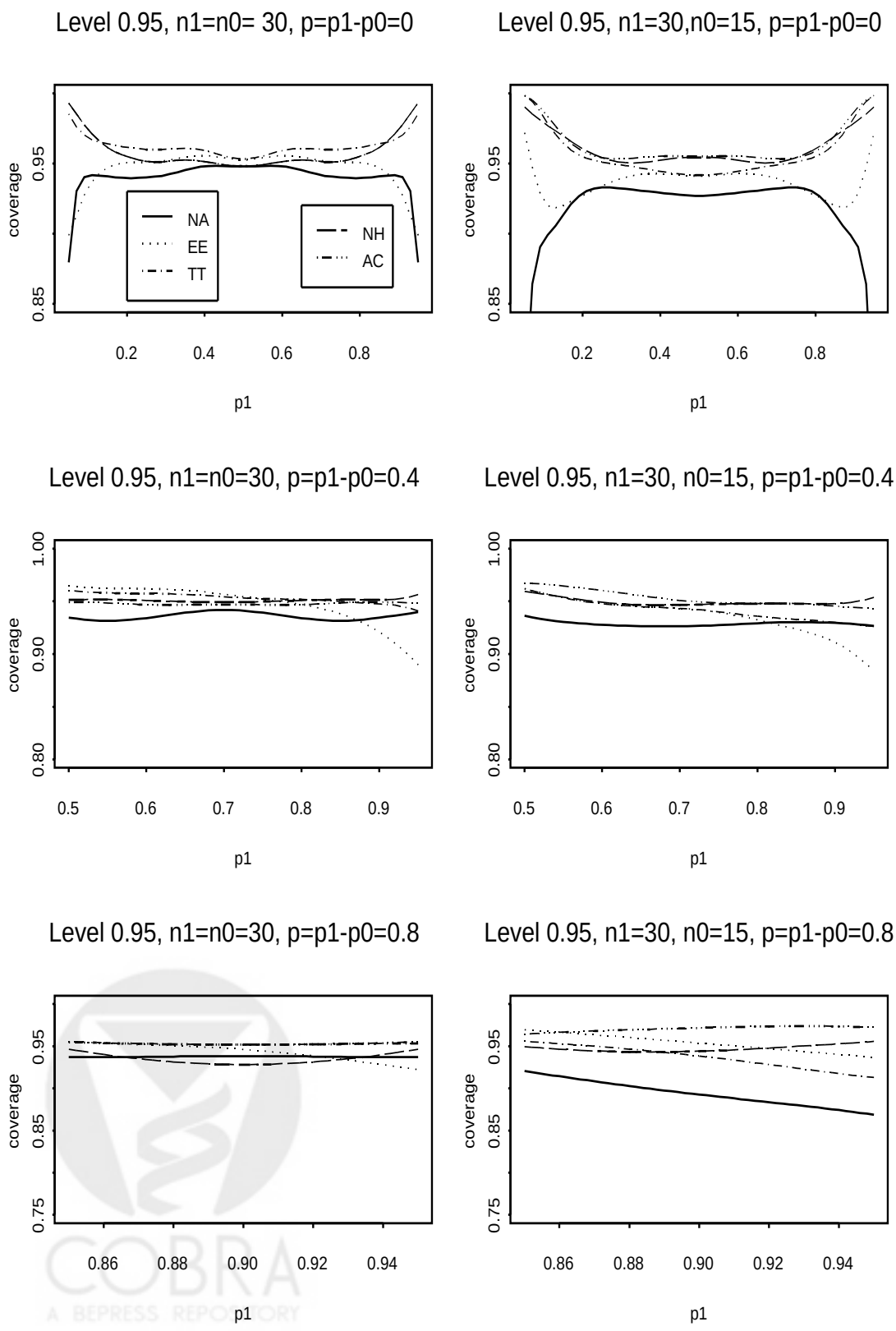

Figure 3: Coverage probability of the various confidence intervals for $p = p_1 - p_0$

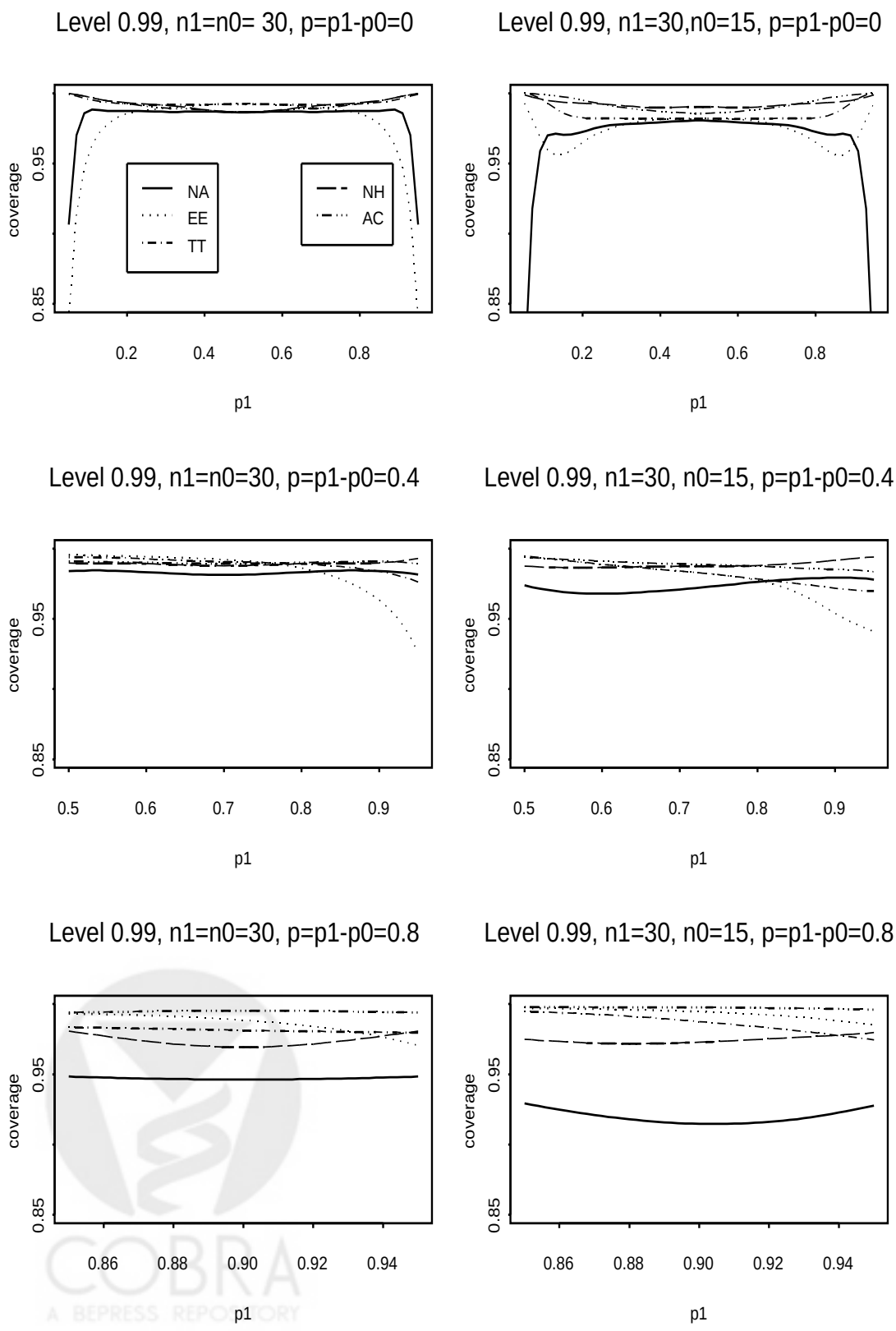

Figure 4: Coverage probability of the various two-sided 90% intervals when n_1=15 and n_0=15

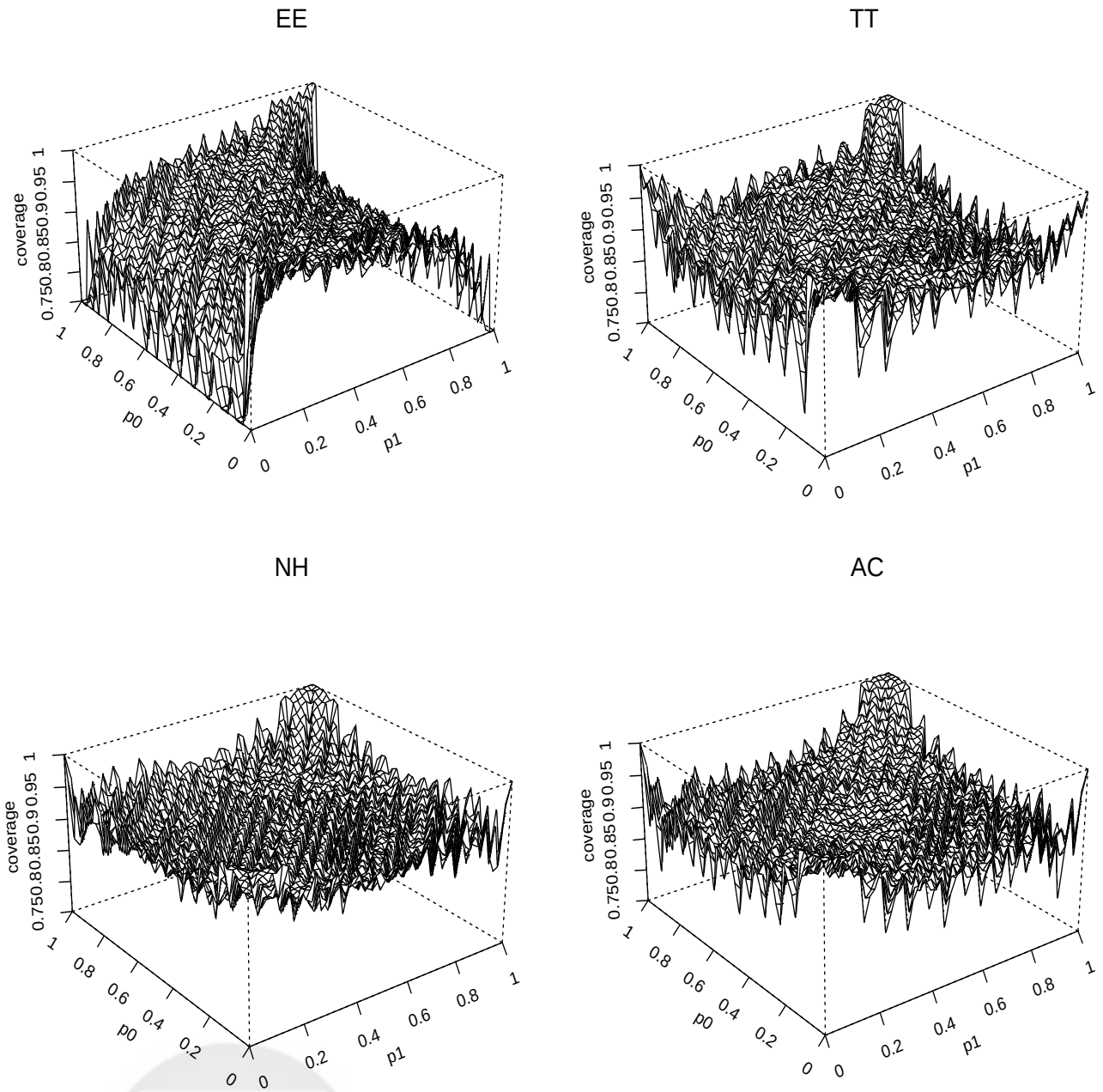

Figure 5: Coverage probability of the various two-sided 90% intervals when n_1=30 and n_0=30

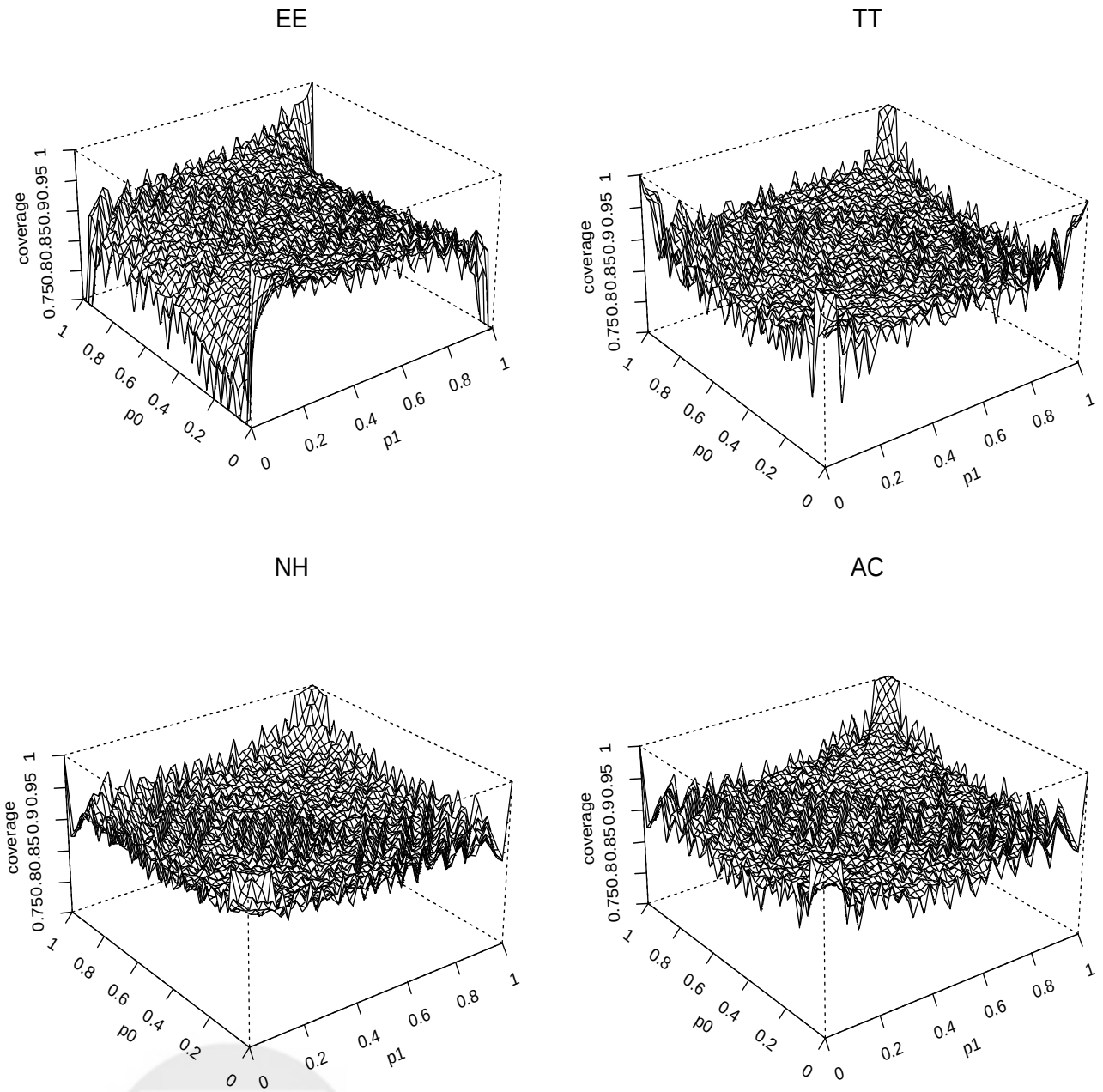

Table 2. Average coverage probability of nominal 95% confidence intervals for fixed $p = p_1 - p_0$, averaging with respect to p_1's.

p	(n_1, n_0)	EE	TT	NH	AC	WA
0.0	(15,15)	95.73	97.21	95.66	97.02	91.02
	(30,30)	94.61	96.22	95.79	95.79	94.00
	(30,15)	93.65	95.69	96.07	96.33	91.49
0.4	(15,15)	94.46	95.41	95.22	95.22	90.93
	(30,30)	94.72	95.31	95.09	94.77	93.58
	(30,15)	93.53	94.12	94.92	95.28	92.88
0.8	(15,15)	94.01	92.64	93.02	97.23	80.97
	(30,30)	94.32	95.28	93.46	95.30	93.76
	(30,15)	95.35	93.66	94.71	94.70	89.36

Table 3. Average coverage probability of nominal 99% confidence intervals for fixed $p = p_1 - p_0$, averaging with respect to p_1's.

p	(n_1, n_0)	EE	TT	NH	AC	WA
0.0	(15,15)	97.08	99.37	99.14	99.14	94.73
	(30,30)	97.52	99.30	99.19	99.13	98.27
	(30,15)	97.40	98.51	99.21	99.20	96.53
0.4	(15,15)	97.59	98.80	98.54	98.96	96.88
	(30,30)	98.37	98.92	98.92	99.01	98.30
	(30,15)	97.76	98.21	98.83	98.87	97.34
0.8	(15,15)	97.85	99.43	97.48	99.75	95.77
	(30,30)	98.65	98.14	97.36	99.48	94.71
	(30,15)	99.34	98.65	97.42	99.74	91.97

Table 4. Average length of the confidence intervals for fixed $p = p_1 - p_0$, averaging with respect to p_1's.

nominal level	(n_1, n_0)	EE	TT	NH	AC	WA
90%	(15,15)	0.4683	0.4786	0.4634	0.4693	0.4609
	(30,30)	0.3360	0.3396	0.3344	0.3368	0.3340
	(30,15)	0.4068	0.4117	0.4029	0.4081	0.4014
95%	(15,15)	0.5580	0.5763	0.5479	0.5593	0.5492
	(30,30)	0.4004	0.4066	0.3975	0.4013	0.3980
	(30,15)	0.4848	0.4934	0.4771	0.4863	0.4783
99%	(15,15)	0.7333	0.7844	0.7010	0.7350	0.7218
	(30,30)	0.5262	0.5411	0.5170	0.5274	0.5231
	(30,15)	0.6371	0.6588	0.6139	0.6391	0.6285

Table 5. Summary of performance of nominal 90% confidence interval for $p = p_1 - p_0$, averaging with respect to uniform distributions for (p_0, p_1): $p_0 \sim U[0,1]$, $p_1 \sim U[0,1]$

Characteristic	(n_1, n_0)	EE	TT	NH	AC	WA
Ave. Cov.	(15,15)	0.895	0.910	0.905	0.912	0.863
	(30,30)	0.897	0.905	0.903	0.907	0.882
	(60,60)	0.898	0.903	0.902	0.904	0.891
	(30,15)	0.898	0.904	0.905	0.911	0.866
	(60,30)	0.898	0.902	0.903	0.906	0.884
Length	(15,15)	0.469	0.479	0.464	0.470	0.462
	(30,30)	0.335	0.339	0.334	0.336	0.333
	(60,60)	0.239	0.240	0.239	0.240	0.239
	(30,15)	0.407	0.411	0.403	0.408	0.401
	(60,30)	0.291	0.293	0.290	0.292	0.290
Cov. Prob.< .88	(15,15)	0.234	0.072	0.086	0.029	0.674
	(30,30)	0.173	0.038	0.025	0.013	0.269
	(60,60)	0.098	0.009	0.006	0.003	0.076
	(30,15)	0.202	0.081	0.006	0.008	0.713
	(60,30)	0.093	0.017	0.001	0.003	0.186

Note:

$k = 10000$ observations for (p_1, p_0).

Ave. Cov.= mean of coverage probabilities $C(n_0, p_0; n_1, p_1)$'s.

Length = mean of expected confidence interval lengths.

Cov. Prob. = proportion of cases with $C(n0, p0; n1, p1) < 0.88$.

Table 6. Test results of the conventional MRI among patients with advanced stage prostate cancer

Hospital	Positive	Negative	Total
Hospital 1	18	17	35
Hospital 2	27	14	41

Table 7. Data on SIDS children

Twin type	One death	Two deaths	Total
Identical	8	2	10
Fraternal	35	1	36