6-2-2003

A New Confidence Interval for the Difference Between Two Binomial Proportions of Paired Data

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

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

1. Introduction

Our research is motivated by a study on assessing the relative accuracy of conventional body coil magnetic resonance imaging (MRI) and transrectal ultrasound in detecting advanced stage prostate cancer (Rifkin et al., 1990). The data were obtained as part of a multi-center study conducted by the Radiology Diagnostic Oncology Group (RDOG). Whether a patient has advanced stage prostate cancer can have a direct effect on physician's treatment choice for this patient. If a patient has advanced stage prostate cancer, the patient is best managed with a combination of radiation therapy and hormonal therapy. If a patient only has localized prostate cancer, the patient has a high likelihood of surgical cure. Thus, the critical issue is whether a patient has advanced stage prostate cancer. Patients in four institutions were enrolled in the study if they had biopsy proven prostate carcinoma, and all were thought to have surgically resectable tumors. The gold standard for prostate cancer stage was established by pathology analyses on patient's specimens obtained from surgery. Each patient was imaged by both MRI and ultrasound. Images from each of the MRI and ultrasound on each patient were separated and prospectively read by a radiologist at the patient's institution. The radiologist gave an overall staging assessment on each image (advanced versus localized stage). In the current analysis we use the data from one institution, and Tables 1 and 2 summarize the data in this institution on diseased and non-diseased patients, respectively.

TABLES 1 AND 2 GO HERE

Note that our two data sets have small sample sizes; it is typical to see small sample sizes in radiology studies conducted in one institution.

When the response of a diagnostic test is binary, its accuracy is often represented by its sensitivity

and specificity. The sensitivity of a diagnostic test is defined as the probability of giving a correct diagnosis in the population of patients with advanced stage prostate cancer, and the specificity of a diagnostic test is defined as the probability of giving a correct diagnosis in the population of patients with localized stage prostate cancer. One goal of our analysis is to give confidence intervals for the differences between sensitivities of MRI and ultrasound and between specificities of MRI and ultrasound. Thus, our statistical problem is how to construct an appropriate confidence interval for the difference between paired binomial proportions when sample sizes are small.

The most commonly used interval for the difference between paired binomial proportions is the Wald confidence interval (hereafter WA) (Fleiss, 1981). Because it is based on the asymptotic theory, it has been shown that it is anti-conservative and has poor coverage probabilities. Several authors have developed "exact" confidence intervals for the difference (Armitage and Berry, 1987; Liddell, 1983). However, because of the discrete nature for a binomial distribution, Newcombe (1998) showed that these "exact" intervals tend to perform poorly.

Many authors have proposed alternative approximate intervals for the difference between paired binomial proportions. Newcombe (1998) reviewed the statistical literature on confidence intervals for the difference between paired binomial proportions. After a comprehensive simulation study on the relative advantages of existing methods, including the Wald interval, the Wald interval with continuity correction, an 'exact' Clopper-Pearson interval, a 'mid-p' interval, the three different types of profile likelihood based intervals, and the three different types of score intervals, Newcombe (1998) recommended a score interval with continuity correction (the method 10 in his paper, hereafter it is called the Newcombe's hybrid (NH) score method), which is based on the Wilson (1927) score interval for a single

proportion. However, Newcombe's procedure can still be conservative when sample sizes are small. In addition, theoretical properties (e.g. consistency) of Newcombe's hybrid score interval are unknown. Another competing interval (called the MJ interval hereafter) was studied by May and Johnson (1997) and subsequently discussed by many other authors including Lui (1998), Newcombe (1998), and Tango (1998). Basically, this interval is based on the normal approximation of the distribution of the difference between the paired sample proportions, its interval length will be zero when the number of discordant pairs is zero.

In this paper we first derive an Edgeworth expansion for the studentized difference between the two correlated sample proportions. One application of the Edgeworth expansion is to help us understand why the Wald interval for the difference between two correlated proportions has so poor coverage performance (see Section 3). This idea has been also used by Brown et al (2002) to explain poor performance of the Wald interval in the one sample binomial case. Another application of the Edgeworth expansion is to guide us to derive a new confidence interval for the difference between paired binomial proportions. The new interval corrects the skewness in the Edgeworth expansion through a monotone transformation. A third application of the Edgeworth expansion is to help us derive the asymptotic coverage accuracy of the new interval. We show that the coverage probability of the new interval converges to the nominal confidence level at the rate of $O(n^{-1/2})$. We also compare the finite-sample performance of the new interval with the existing intervals. We find that the new interval has the best average coverage accuracy among the three intervals considered here and that its average coverage probability is still close to the nominal level even for the sample size as small as 10. Finally we illustrate the application of the newly proposed method in two real studies, including the motivating example described in this section.

This paper is organized as follows. In Section 2 we review the commonly used Wald interval and the Newcombe's hybrid score interval as well as the May and Johnson's (MJ) interval. In Section 3 we give the Edgeworth expansion for the studentized difference between two correlated sample proportions. In Section 4 we derive a new confidence interval based on the Edgeworth expansion. In Section 5 we evaluate the finite-sample performance of the proposed interval and compare it to the usual Wald interval, the MJ interval, and the Newcombe's hybrid score interval in terms of the coverage probability and the interval length. In Section 6 we apply our method to our motivating example, and in Section 7 we further illustrate the application of our method in a study on the relative accuracy of two diagnostic tests in detecting hyperparathyroidism. Finally, in the Appendix we present theoretical derivations of the Edgeworth expansion and the asymptotic order of the coverage error of the new interval.

2. The Wald, Newcombe's hybrid score, and May and Johnson intervals

Let (X_{0k}, X_{1k}), $k = 1, 2, \cdots, n$, be an independent and identically distributed (i.i.d.) sample from the joint distribution of paired random variables (X_0, X_1), where X_0 and X_1 are correlated Bernoulli random variables with proportions p_0 and p_1 respectively; let $p = p_1 - p_0$. The most commonly used Wald interval for p is based on the normal approximation to the distribution of the studentized difference between the two correlated sample proportions, defined by

$$T = \frac{\sqrt{n}(\hat{p} - p)}{\sqrt{\hat{p}_0(1-\hat{p}_0) + \hat{p}_1(1-\hat{p}_1) + 2(\hat{p}_1\hat{p}_0 - \widehat{p_{11}})}}, \tag{1}$$

and the Wald interval is

$$\left[\hat{p} - z_{1-\alpha/2}n^{-1/2}\sqrt{\hat{p}_0(1-\hat{p}_0) + \hat{p}_1(1-\hat{p}_1) + 2(\hat{p}_1\hat{p}_0 - \widehat{p_{11}})},\right.$$
$$\left.\hat{p} + z_{1-\alpha/2}n^{-1/2}\sqrt{\hat{p}_0(1-\hat{p}_0) + \hat{p}_1(1-\hat{p}_1) + 2(\hat{p}_1\hat{p}_0 - \widehat{p_{11}})}\right], \tag{2}$$

where $Y_i = \sum_{k=1}^{n} X_{ik}$, $Y_{11} = \sum_{k=1}^{n} X_{0k}X_{1k}$, $\hat{p}_i = Y_i/n$, $\hat{p} = \hat{p}_1 - \hat{p}_0$, $\widehat{p_{11}} = Y_{11}/n$, and z_α is the α-th quantile of the standard normal distribution.

However, the normal approximation to the distribution of T may be a rather crude approximation, especially when sample sizes are not large; it does not take into consideration the skewness of the underlying multinomial distribution which is often the main source of error of the normal approximation. Therefore, the Wald interval can give poor coverage accuracy.

One of the best alternative intervals was proposed by Newcombe (1998) and is called the Newcombe's hybrid (NH) interval. To describe this interval we need additional notation. Let us define $Y_{00} = \sum_k (1 - X_{0k})(1 - X_{1k})$, $Y_{10} = \sum_k X_{0k}(1 - X_{1k})$, and $Y_{01} = \sum_k (1 - X_{0k})X_{1k}$. Let us denote $D = (Y_{00} + Y_{10})(Y_{01} + Y_{11})(Y_{00} + Y_{01})(Y_{10} + Y_{11})$. Let l_1 and u_1 be the roots to the following quadratic equation in x:

$$(x - \frac{Y_{00} + Y_{01}}{n})^2 = (z_{1-\alpha/2})^2 \frac{x(1-x)}{n},$$

and let l_2 and u_2 be the roots to the following quadratic equation in x:

$$(x - \frac{Y_{00} + Y_{10}}{n})^2 = (z_{1-\alpha/2})^2 \frac{x(1-x)}{n}.$$

Then, the Newcombe's hybrid score interval is defined by

$$\left[\widehat{p} - \left(\delta_1^2 - 2\widehat{\phi}\delta_1\epsilon_2 + \epsilon_2^2\right)^{1/2}, \quad \widehat{p} + \left(\epsilon_1^2 - 2\widehat{\phi}\epsilon_1\delta_2 + \delta_2^2\right)^{1/2}\right], \tag{3}$$

where

$$\begin{aligned} \delta_1 &= (Y_{00} + Y_{01})/n - l_1, & \epsilon_1 &= u_1 - (Y_{00} + Y_{01})/n, \\ \delta_2 &= (Y_{00} + Y_{10})/n - l_2, & \epsilon_2 &= u_2 - (Y_{00} + Y_{10})/n, \end{aligned}$$

$$\widehat{\phi} \;=\; \begin{cases} \left(Y_{00}Y_{11} - Y_{10}Y_{01}\right)/D, & Y_{00}Y_{11} - Y_{10}Y_{01} \le 0 \;\text{ and }\; D > 0, \\ \max(Y_{00}Y_{11} - Y_{10}Y_{01} - n/2, 0)/D, & Y_{00}Y_{11} - Y_{10}Y_{01} > 0 \;\text{ and }\; D > 0, \\ 0, & D = 0. \end{cases}$$

Let $A = (1 + z^2_{\alpha/2}/n)$, $B = -2(Y_{01} - Y_{10})/n$, $C = (Y_{01}/n - Y_{10}/n)^2 - z^2_{\alpha/2}(Y_{01} + Y_{10})/n^2$. May and Johnson (1997) proposed an alternative interval for p, and the resulting $100(1-\alpha)\%$ interval for p is given as follows:

$$\left[\max\{0, (-B - (B^2 - 4AC)^{1/2})/(2A)\}, \;\; \min\{1, (-B + (B^2 - 4AC)^{1/2})/(2A)\}\right]$$

3. Edgeworth expansion for the studentized difference

The validity of the Wald interval relies on the standard normality assumption of the studentized difference between two correlated sample proportions, T. Since the true distribution of T is skewed, the normal approximation may not be appropriate in a finite sample size. To see the impact of the skewness on the normal approximation, we develop the Edgeworth expansion for T. To state this Edgeworth expansion, we need the following notation:

$$d = p_1(1-p_1)(1-2p_1) - p_0(1-p_0)(1-2p_0) + 6(p_1 - p_0)(p_{11} - p_0p_1),$$

$$\sigma = \left(p_1(1-p_1) + p_0(1-p_0) + 2(p_0p_1 - p_{11})\right)^{1/2}, a = d/(6\sigma^2), \text{ and } \quad b = (1-2p)/2 - d/(6\sigma^2),$$

where $p_{11} = P(X_0 = 1, X_1 = 1)$. Let $Q(t) = a + bt^2$. Now we can state the Edgeworth expansion for T as follows.

Theorem 1 *If p_0 and p_1 are rational numbers, then*

$$P(T \leq t) \quad = \quad \Phi(t) + \left(n\sigma^2\right)^{-1/2} \left(Q(t) + g_n(p_0, p_1, t)\right) \phi(t) + O\left(n^{-1} loglogn\right) \tag{4}$$

where $g_n(p_0, p_1, t)$ is a discontinuous function and has a range between -0.5 and 0.5, $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative distribution and density functions of the standard normal random variable, respectively.

For a proof of this theorem, see Appendix.

From Theorem 1 we see that the accuracy of the normal approximation to the distribution of T at an argument t depends on both the main error term $Q(t)$, which is due to the skewness of a multinomial distribution, and the rounding error term $g_n(p_0, p_1, t)$, which is due to the discrete nature of a multinomial distribution. If d is close to 0 (which may happen when p is near 0, or both p_0 and p_1 are near boundary points 0 and 1), the main part of $Q(t)$ will be close to $(1-2p)t^2/2$, which is larger than the rounding error $|g_n(p_0, p_1, t)|$ when $p > (1+t^{-2})/2$ or $p < (1-t^{-2})/2$. Consequently, the distribution of T could be far from the standard normal distribution in small sample size, which result in the poor performance of the Wald interval. Therefore it is important to correct for skewness when constructing a confidence interval for the difference p.

4. A new confidence interval

In this section, we propose a new confidence interval for p by eliminating the error due to the skewness in the Edgeworth expansion given in Theorem 1. The new method uses a monotone transformation based on the Edgeworth expansion of T, given in (4), by ignoring the rounding error term $g_n(p_0, p_1, t)$. This transformation was originally introduced by Hall (1992) for removing the skewness of an asymmetric

statistic in a one-sample case. 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 $\widehat{a}$, $\widehat{b}$, $\widehat{\sigma}$, and $\widehat{d}$ are estimates of a, b, σ, and d, respectively, and a, b, σ, and d are defined in Section 3. The estimates $\widehat{a}$, $\widehat{b}$, $\widehat{\sigma}$, and $\widehat{d}$ are computed by replacing the p_i's and p_{11} in the formulas for a, b, σ, and d with the $\widehat{p}_i$'s and $\widehat{p_{11}}$. Using this transformation, we can construct the following two-sided $100(1-\alpha)\%$ confidence interval for p:

$$I_\alpha = \left[\max\left(-1, \widehat{p} - \frac{\widehat{\sigma}}{\sqrt{n}}\cdot g^{-1}(z_{1-\alpha/2})\right),\ \ \min\left(1, \widehat{p} - \frac{\widehat{\sigma}}{\sqrt{n}}\cdot g^{-1}(z_{\alpha/2})\right)\right],$$

where

$$g^{-1}(y) = n^{1/2}\left(\widehat{b}\widehat{\sigma}\right)^{-1}\left[\left(1 + 3\left(\widehat{b}\widehat{\sigma}\right)\left(n^{-1/2}y - n^{-1}\widehat{a}\widehat{\sigma}\right)\right)^{1/3} - 1\right] \text{ if } \widehat{b}\widehat{\sigma} \neq 0,$$

and $g^{-1}(y) = y - n^{-1/2}\widehat{a}\widehat{\sigma}$ if $\widehat{b}\widehat{\sigma} = 0$. The following theorem gives the asymptotic coverage probability of the proposed interval.

Theorem 2 *If p_0 and p_1 are rational numbers, then*

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

For a proof of this theorem, see Appendix.

5. A numerical study

We conducted a numerical study to assess the finite-sample performances of the newly proposed transformation based interval (denoted by TT), the Newcombe's hybrid score interval (NH), the May and Johnson interval (MJ), and the Wald interval (WA). The criteria for comparison are the coverage

probability and length of intervals. To compare the relative performance of TT, NH, MJ, and WA intervals for $p = p_1 - p_0$, we compute their coverage probabilities and the expected lengths. For fixed values of n and (p_0, p_1, p_{11}), we denote $C(n; p_0, p_1, p_{11})$ and $W(n; p_0, p_1, p_{11})$ to be the coverage probability and the expected length of a two-sided $100(1-\alpha)\%$ level confidence interval $\mathcal{L}(Y_{01}, Y_{10}, Y_{11})$ for $p = p_1 - p_0$, respectively. Then,

$$\begin{aligned} C(n; p_0, p_1, p_{11}) &= E\{I_{[p \in \mathcal{L}(Y_{01}, Y_{10}, Y_{11})]} \mid n; p_0, p_1, p_{11}\} \\ &= \sum_{y_{01}=0}^{n} \sum_{y_{10}=0}^{n-y_{01}} \sum_{y_{11}=0}^{n-y_{01}-y_{10}} Multi(n; y_{01}, y_{10}, y_{11}, p_0, p_1, p_{11}) I_{[p \in \mathcal{L}(y_{01}, y_{10}, y_{11})]} \end{aligned}$$

where $I_{[p \in \mathcal{L}(y_{01}, y_{10}, y_{11})]}$ is 1 if $p \in \mathcal{L}(y_{01}, y_{10}, y_{11})$ and zero otherwise, and $Multi(n; y_{01}, y_{10}, y_{11}, p_0, p_1, p_{11})$ is the multinomial probability when $Y_{01} = y_{01}, Y_{10} = y_{10}$, and $Y_{11} = y_{11}$, defined by

$$\begin{aligned} Multi(n; y_{01}, y_{10}, y_{11}, p_0, p_1, p_{11}) &= \frac{n!}{y_{01}! y_{10}! y_{11}! (n - y_{01} - y_{10} - y_{11})!} \\ &\quad \times p_{01}^{y_{01}} p_{10}^{y_{10}} p_{11}^{y_{11}} (1 - p_{01} - p_{10} - p_{11})^{(n - y_{01} - y_{10} - y_{11})}. \end{aligned}$$

Denote the lower and upper endpoints of $\mathcal{L}(y_{01}, y_{10}, y_{11})$ to be $lower(y_{01}, y_{10}, y_{11})$ and $upper(y_{01}, y_{10}, y_{11})$, respectively. Then, the expected interval length for $\mathcal{L}(Y_{01}, Y_{10}, Y_{11})$ is calculated using the formula,

$$\begin{aligned} W(n; p_0, p_1, p_{11}) &= \sum_{y_{01}=0}^{n} \sum_{y_{10}=0}^{n-y_{01}} \sum_{y_{11}=0}^{n-y_{01}-y_{10}} Multi(n; y_{01}, y_{10}, y_{11}, p_0, p_1, p_{11}) \\ &\quad \times [upper(y_{01}, y_{10}, y_{11}) - lower(y_{01}, y_{10}, y_{11})] \end{aligned}$$

We compared the performance of the four intervals in terms of the average of $C(n; p_0, p_1, p_{11})$'s and $W(n; p_0, p_1, p_{11})$'s over the systematically chosen values of (p_0, p_1, p_{11}), where $(p_0, p_1) = (0.05i, 0.05j)$ for $i, j = 1, 2, ..., 19$, and p_{11} is from the interval $[e_1, e_2]$ with $e_1 = max(0, p_0 + p_1 - 1)$, and $e_2 = min(p_0, p_1)$. Specifically, for a given $(p_0, p_1) \in \{(0.05i, 0.05j) : i, j = 1, 2, ..., 19.\}$, we generated a p_{11}

from $Uniform[e_1, e_2]$. With $p_{01} = p_1 - p_{11}$, $p_{10} = p_0 - p_{11}$, and $p_{00} = 1 - p_{10} - p_{01} - p_{11}$, we can compute the multinomial probability $Multi(n; p_{00}, p_{01}, p_{10}, p_{11})$ and thus the coverage probability and expected length of various confidence intervals. In the calculation of coverage probability and length of the TT interval, we replaced Y_{ij} by $Y_{ij} + 0.25$ for $i, j = 0, 1$ and n by $n + 1$. This is motivated by a similar technique used by Agresti and Coull (1998). Tables 3 and 4 display the summary performances of the four intervals.

TABLES 3-4 GO HERE

From the results on the summary measures in Tables 3-4, we conclude that the new interval has the best average coverage accuracy among the four intervals considered here and that the new interval has the average coverage probability that is very close to the nominal level for the sample size as small as 10. For all the sample sizes considered here, the Newcombe's hybrid interval have the average coverage probability that is higher than the nominal level. From Tables 3-4 we also confirm that the Wald interval has poor coverage accuracy even for the sample size as large as 50. The MJ interval improved the Wald interval, but the improvement over the Wald interval is moderate particularly for small sample sizes; this is not a surprising observation because the MJ interval is essentially an interval derived from the normal approximation of the distribution of the difference between paired sample proportions.

To obtain information about the variation and spread of the coverage probabilities and lengths of the intervals, we also produce box plots of the coverage probabilities and expected lengths of the TT, NH, MJ and WA intervals. Figures 1-4 display the box plots for the four intervals with 90% and 95% confidence levels and four sample sizes ($n = 10, 15, 30, 50$).

FIGURES 1-4 GO HERE

From the results in Figure 1-4 we see that the MJ and Wald intervals have larger spread in coverage probabilities than the new interval and the Newcombe's hybrid interval. However, the newly proposed interval has little wider spread in length than the MJ interval and the Wald interval for small sample size but the difference tends to be small as the sample size increases to moderate size (n=30,50).

6. Comparisons of diagnostic accuracy of MRI and ultrasound

As discussed in Section 1, we are interested in comparing sensitivities and specificities of MRI and ultrasound in detecting advanced stage prostate cancer. Let π_0 and π_1 be sensitivities of the MRI and ultrasound, respectively, and let ν_0 and ν_1 be specificities of the MRI and ultrasound, respectively.

Using the newly proposed method, we derived the 90% confidence intervals for both $\pi_1 - \pi_0$ and $\nu_1 - \nu_0$. The resulting intervals are $(-0.11, 0.27)$ for $\pi_1 - \pi_0$ and $(-0.09, 0.50)$ for $\nu_1 - \nu_0$. These intervals suggest that there are no statistically significant differences between sensitivities and specificities of MRI and ultrasound. Therefore, MRI and ultrasound have the similar diagnostic accuracy in detecting advance stage prostate cancer. However, it is worth noting that conventional MRI costs \$700 to \$1200 per examination and that transrectal ultrasound imaging only costs \$150 to \$400 per examination. Therefore, choosing ultrasound over MRI could save \$300 to \$1050 without compromising quality of care.

7. An additional example

In this section, we further illustrate the application of the proposed method in a study on the accuracy of Positron-emission tomography (PET) and the double-phase^{99m} Tc-sestamibi-SPECT in detecting hyperparathyroidism (Neumann et al. 1996). In the study, each of the 21 subjects was

evaluated with both PET and the double-phase^{99m} Tc-sestamibi-SPECT before undergoing surgery, the gold standard. We want o compare the specificity of PET with that of the double-phase^{99m} Tc-sestamibi-SPECT to determine if they are different. Table 5 displays results of PET and double-phase^{99m} Tc-sestamibi-SPECT for patients without evidence of hyperparathyroidism.

TABLE 5 GOES HERE

Let ν_1 and ν_0 be the specificities of the PET and double-phase^{99m} Tc-sestamibi-SPECT, respectively. Using the newly proposed method, we derived the 99% confidence intervals for $\nu_1 - \nu_0$ as $(-0.107, 0.429)$. Hence, we conclude that the specificities of the PET and double-phase^{99m} Tc-sestamibi-SPECT are not statistically different in ruling out hyperparathyroidism.

8. Discussion

Motivated by an accuracy study of diagnostic tests, we have developed a new confidence interval for the difference between paired binomial proportions. Comparing with the best existing intervals, the new interval is easier to compute, has a sound theoretical justification, and has a better average performance in finite sample sizes. For the sample size as small as 10, the average coverage probability of our new interval is still very close to the nominal level. Even though the proposed method was originally developed for comparing sensitivities and specificities of two diagnostic tests, it can be equally applied to a general situation of comparing two binomial proportions from paired designs.

APPENDIX

Proof of Theorem 1.

We first derive the Edgeworth expansion for the standardized sample difference. For each $i = 0, 1$, note that $Y_i = \sum_{k=1}^{n} X_{ik}$ where X_{ik}'s are i.i.d. Bernoulli random variables with the parameter p_i. Then the standardized sample difference is defined as follows.

$$T_n \equiv \frac{\sqrt{n}(\widehat{p} - p)}{\sqrt{p_0 q_0 + p_1 q_1 + 2(p_1 p_0 - p_{11})}} = \sum_{k=1}^{n} \frac{D_k}{\sqrt{n}\sigma}$$

where

$$D_k = (X_{1k} - X_{0k}) - (p_1 - p_0), \quad k = 1, 2, \cdots, n,$$

After deriving the Edgeworth expansion for T_n, we will then derive the Edgeworth expansion for T, the studentized difference.

Our derivation of the Edgeworth expansion for T_n is aided by a result in Kolassa (1995) on the Edgeworth expansion for the sum of independent random variables supported on the same lattice. To apply 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 lp_0 and lp_1 are integers. Let $\Delta = 1/l$ and let A be a constant such that A/Δ is an integer. Also let $k_1 = -(1 + p_1 - p_0)/\Delta - A/\Delta$, $k_2 = -(p_1 - p_0)/\Delta - A/\Delta$, and $k_3 = (1 - p_1 + p_0)/\Delta - A/\Delta$, then $\{-(1 + p_1 - p_0), -(p_1 - p_0), 1 - p_1 + p_0\} = \{A + k_1\Delta, A + k_2\Delta, A + k_3\Delta\}$ 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 $\{A + \Delta\mathbf{Z}\}$; furthermore, 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

$$\kappa_3 = \frac{1}{\sqrt{n}\sigma^3}\left[p_1 q_1(1 - 2p_1) - p_0 q_0(1 - 2p_0) + 6(p_1 - p_0)(p_{11} - p_0 p_1)\right] \equiv \frac{d}{\sqrt{n}\sigma^3}$$

and

$$\kappa_4 = \frac{1}{n\sigma^4}\left[E(D_1^4) - 3\left(E(D_1^2)\right)^2\right] = O\left(n^{-1}\right)$$

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

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

where $g_n(p_0, p_1, t) = -\Delta \cdot G\left((t - t_n)/\Delta_n\right)$, $G(t) = t - 1/2$, and $\Delta_n = \Delta/(\sqrt{n}\sigma)$ is the lattice spacing for t_n that is the largest lattice point less than t. It is easily seen that $g_n(p_0, p_1, t)$ is a discontinuous function taking values in $[-0.5, 0.5]$. Next we use the Edgeworth expansion for T_n to obtain an Edgeworth expansion for T. Note that

$$P(T \le t) = P\left(\frac{\sqrt{n}(\widehat{p} - p)}{\sqrt{\left[(\widehat{p} - p) + (\widehat{p}_0 + p)\right]\left[(1 + \widehat{p}_0 - p) - (\widehat{p} - p)\right] + (\widehat{p_0 q_0} - 2\widehat{p_{11}})}} \le t\right).$$

By carefully solving the inequality for $\widehat{p} - p$ in the right side of the above equation, we obtain that $P(T \le t) = P\left(T_n \le \widehat{t}_0\right)$, where

$$\widehat{t}_0 = \frac{\sqrt{n}}{\sigma}\left(\frac{(1-2p)t^2}{2\left(n + t^2\right)} + \frac{t\left[4\left(pq + 2\widehat{p}_{10}\right)/n + t^2\left(1 + 8\widehat{p}_{10}\right)/n^2\right]^{1/2}}{2\left(1 + t^2/n\right)}\right),$$

$$\widehat{p}_{10} = n^{-1}\sum_i X_{0i}(1 - X_{1i}), q = 1 - p.$$

Let us define t_0 to be $\widehat{t}_0$ except that $\widehat{p}_{10}$ is replaced by $p_{10} = P\left(X_0 = 1, X_1 = 0\right)$; that is,

$$t_0 = \frac{\sqrt{n}}{\sigma}\left(\frac{(1-2p)t^2}{2\left(n + t^2\right)} + \frac{t\left[4\sigma^2/n + t^2\left(1 + 8p_{10}\right)/n^2\right]^{1/2}}{2\left(1 + t^2/n\right)}\right)$$

Then,

$$P\left(T_n \le \widehat{t}_0\right) = P\left(T_n \le t_0\right) + \left(P\left(T_n \le \widehat{t}_0\right) - P\left(T_n \le t_0\right)\right) \equiv I_1 + I_2. \tag{6}$$

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

$$\begin{aligned} I_1 &= \Phi(t_0) + \left(n\sigma^2\right)^{-1/2} \cdot \frac{d}{6\sigma^2}\left(1 - t_0^2\right)\phi(t_0) \\ &\quad + \left(n\sigma^2\right)^{-1/2} g_n(p_0, p_1, t_0)\phi(t_0) + O\left(n^{-1}\right) \\ &= \Phi(t) + \left(n\sigma^2\right)^{-1/2}\left(a + bt^2\right)\phi(t) + \left(n\sigma^2\right)^{-1/2} g_n(p_0, p_1, t)\phi(t) + O\left(n^{-1}\right). \end{aligned} \tag{7}$$

Now we show that $I_2 = O\left(n^{-1}loglogn\right)$. By $\widehat{p}_{10} - p_{10} = O\left(n^{-1/2}loglogn\right)$ *a.s.*, we can find a positive constant C such that $|\widehat{t}_0 - t_0| \leq C\left(n^{-1}loglogn\right)$ *a.s.*. That is, the interval between $\widehat{t}_0$ 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

$$\begin{aligned} |I_2| &\leq P\left(T_n \leq t_0 + C\left(n^{-1}loglogn\right)\right) - P\left(T_n \leq t_0 - C\left(n^{-1}loglogn\right)\right) \\ &= O\left(n^{-1}loglogn\right). \end{aligned} \tag{8}$$

Theorem 1 then follows from (6)–(8).

Proof of Theorem 2. Let $Q_1(t) = \sigma\left(a + bt^2\right)$, $\widehat{Q}_1(t) = \widehat{\sigma}\left(\widehat{a} + \widehat{b}t^2\right)$, and

$$I_{1\alpha} = \left[\widehat{p} - \frac{\widehat{\sigma}}{\sqrt{n}}\left(z_{1-\alpha/2} - n^{-1/2}\widehat{Q}_1(z_{1-\alpha/2})\right),\ \ \widehat{p} - \frac{\widehat{\sigma}}{\sqrt{n}}\left(z_{\alpha/2} - n^{-1/2}\widehat{Q}_1(z_{\alpha/2})\right)\right].$$

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

$$P\left(T \leq z_\alpha - n^{-1/2}\widehat{Q}_1\left(z_\alpha\right)\right) = P\left(T \leq z_\alpha - n^{-1/2}Q_1\left(z_\alpha\right)\right)$$

$$+\left[P\left(T \leq z_\alpha - n^{-1/2}\widehat{Q}_1\left(z_\alpha\right)\right) - P\left(T \leq z_\alpha - n^{-1/2}Q_1\left(z_\alpha\right)\right)\right] \equiv J_1 + J_2.$$

Noting that $\Phi(x)$, $\phi(x)$ and $Q_1(x)$ are smooth functions of x, by Theorem 1 and Taylor expansion, we obtain that

$$\begin{aligned} J_1 &= \Phi\left(z_\alpha - n^{-1/2}Q_1(z_\alpha)\right) + (n\sigma^2)^{-1/2}Q\left(z_\alpha - n^{-1/2}Q_1(z_\alpha)\right)\phi\left(z_\alpha - n^{-1/2}Q_1(z_\alpha)\right) \\ &\quad + \left(n\sigma^2\right)^{-1/2} g_n\left(p_0, p_1, z_\alpha - n^{-1/2}Q_1(z_\alpha)\right)\phi\left(z_\alpha - n^{-1/2}Q_1(z_\alpha)\right) + O\left(n^{-1}loglogn\right) \\ &= \Phi\left(z_\alpha\right) + O\left(n^{-1/2}\right) = \alpha + O\left(n^{-1/2}\right). \end{aligned}$$

For the term J_2, by $\widehat{p}_i - p_i = O\left(n^{-1/2}loglogn\right)$ $a.s.$ and $\widehat{p_{11}} - p_{11} = O\left(n^{-1/2}loglogn\right)$ $a.s.$, we can get

$$\widehat{Q}_1(z_\alpha) - Q_1(z_\alpha) = o\left(1\right) a.s..$$

Hence by Theorem 1,

$$\begin{aligned} J_2 &= P\left\{z_\alpha - n^{-1/2}Q_1(z_\alpha) < T \leq z_\alpha - n^{-1/2}Q_1(z_\alpha) - n^{-1/2}\left(\widehat{Q}_1(z_\alpha) - Q_1(z_\alpha)\right)\right\} \\ &\leq P\left\{z_\alpha - n^{-1/2}Q_1(z_\alpha) < T \leq z_\alpha - n^{-1/2}Q_1(z_\alpha) + Cn^{-1/2}\right\} \\ &= Cn^{-1/2}\phi\left(z_\alpha - n^{-1/2}Q_1(z_\alpha)\right) + O\left(n^{-1/2}\right) = O\left(n^{-1/2}\right). \end{aligned}$$

Therefore,

$$\begin{aligned} P(p \in I_{1\alpha}) &= P\left(T \leq z_{1-\alpha/2} - n^{-1/2}\widehat{Q}_1\left(z_{1-\alpha/2}\right)\right) - P\left(T \leq z_{\alpha/2} - n^{-1/2}\widehat{Q}_1\left(z_{\alpha/2}\right)\right) \\ &= 1 - \alpha + O\left(n^{-1/2}\right). \end{aligned} \tag{9}$$

Now we show that

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

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

$$\left[1+3\left(\widehat{b}\widehat{\sigma}\right)\left(n^{-1/2}x-n^{-1}\widehat{a}\widehat{\sigma}\right)\right]^{1/3}-1$$

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

and hence we have

$$g^{-1}(x)=x-n^{-1/2}\widehat{Q}_1(x)+O\left(n^{-1}\right).$$

An argument similar to the proof of (9) leads to $P(p\in I_\alpha)=1-\alpha+O\left(n^{-1/2}\right)$. We then completed the proof of Theorem 2.

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.

Armitage, P. and Berry, G. (1987). *Statistical Methods in Medical Research.* 2nd edition, Blackwell, Oxford, p.123.

Berry, G. and Armitage, P. (1995). Mid-p confidence intervals: a brief review. *Statistician,* **44**, 412-423.

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

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

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions.* 2nd edition, Wiley & Sons, New York.

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

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

Liddell, F. D. K. (1983). Simplified exact analysis of case-referent studies: matched pairs; dichotomous outcome. *J. of Epidemiology and Community Health*, **37**, 82-84.

Lui, K.L. (1998). Letter to the editor: confidence intervals for differences in correlated binary proportions. *Statistics in Medicine*, **17**, 2017-2021.

May, W. L. and Johnson, W. D. (1997). Confidence intervals for differences in correlated binary proportions. *Statistics in Medicine*, **16**, 2127-2136.

Newcombe, R. G. (1998). Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine*, **17**, 2635-2650.

Neumann, D. R., Esselstyn, C. B., Go, R. T., Wong, C. O., Rice, T. W., and Obuchowski, N. A. (1997). Comparison of FDG PET and Sestamibi-SPECT in Primary Hyperparathyroidism, *Am. J. Roentgenol*, **169**, 1671-1674.

Rifkin, M. D., Zerhouni, E. A., Gatsonis, C. A., et al. (1990). Comparison of magnetic resonance imaging and ultrasonography in staging early prostate cancer. *New England Journal of Medicine*, Vol. 323 621-626.

Tango, T. (1998). Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine*, **17**, 891-908.

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

Table 1. The diagnostic results of MRI and ultrasound on 16 patients with true localized prostate cancer

	Ultrasound		
MRI	Localized stage cancer	Advanced stage cancer	Total
Localized stage cancer	4	6	10
Advanced stage cancer	3	3	6
Total	7	9	16

Table 2. The diagnostic results of MRI and ultrasound on 15 patients with true advanced stage prostate cancer

	Ultrasound		
MRI	Localized stage cancer	Advanced stage cancer	Total
Localized stage cancer	1	2	3
Advanced stage cancer	1	11	12
Total	2	13	15

Table 3. The result of summary measures of nominal 90% confidence interval for $p_1 - p_0$, averaging with respect to $(p_0, p_1, p_{11})'s$, where (p_0, p_1) varies over the points given by $(0.05i, 0.05j)$ for $i, j = 1, 2, ..., 19$, and $p_{11} \sim U[max(0, p_0 + p_1 - 1), min(p_0, p_1)]$.

Characteristic	n	TT	NH	MJ	NA
Ave. Cov.	10	0.9112(0.0351)	0.9186(0.0301)	0.8708(0.0673)	0.8346(0.0721)
	15	0.9046(0.0289)	0.9153(0.0288)	0.8862(0.0457)	0.8624(0.0483)
	30	0.8980(0.0199)	0.9107(0.0213)	0.8925(0.0189)	0.8762(0.0226)
	50	0.8997(0.0103)	0.9058(0.0132)	0.8983(0.0081)	0.8874(0.0097)
	100	0.9003(0.0085)	0.9023(0.0090)	0.8977(0.0059)	0.8935(0.0077)
Length	10	0.6166(0.1525)	0.5781(0.0755)	0.4158(0.1423)	0.5363(0.1576)
	15	0.4896(0.1165)	0.4763(0.0789)	0.3665(0.1196)	0.4537(0.1284)
	30	0.3406(0.0841)	0.3402(0.0676)	0.2845(0.0876)	0.3312(0.0828)
	50	0.2684(0.0619)	0.2682(0.0544)	0.2335(0.0709)	0.2644(0.0681)
	100	0.1913(0.0470)	0.1914(0.0441)	0.1708(0.0501)	0.1880(0.0483)

Note:

$C(n; p_0, p_1, p_{11})$ = coverage probability for $p = p1 - p0$.

Ave. Cov.= mean of coverage probabilities $C(n; p_0, p_1, p_{11})$'s.

Length = mean of expected confidence interval lengths.

Values in the parentheses are the corresponding standard deviations.

Table 4. The result of summary measures of nominal 95% confidence interval for $p_1 - p_0$, averaging with respect to $(p_0, p_1, p_{11})'s$, where (p_0, p_1) varies over the points given by $(0.05i, 0.05j)$ for $i, j = 1, 2, ..., 19$, and $p_{11} \sim U[max(0, p_0 + p_1 - 1), min(p_0, p_1)]$.

Characteristic	n	TT	NH	MJ	NA
Ave. Cov.	10	0.9501(0.0295)	0.9578(0.0195)	0.9014(0.0804)	0.8712(0.0794)
	15	0.9460(0.0266)	0.9565(0.0175)	0.9262(0.0470)	0.9052(0.0483)
	30	0.9469(0.0144)	0.9558(0.0121)	0.9391(0.0203)	0.9294(0.0219)
	50	0.9487(0.0083)	0.9529(0.0097)	0.9447(0.0114)	0.9378(0.0108)
	100	0.9492(0.0085)	0.9523(0.0080)	0.9470(0.0067)	0.9441(0.0063)
Length	10	0.8189(0.2738)	0.6844(0.0853)	0.4625(0.1690)	0.6199(0.2010)
	15	0.6144(0.1600)	0.5719(0.0861)	0.4217(0.1391)	0.5418(0.1485)
	30	0.4161(0.1006)	0.4092(0.0744)	0.3347(0.0999)	0.4023(0.1026)
	50	0.3276(0.0789)	0.3243(0.0659)	0.2721(0.0823)	0.3125(0.0781)
	100	0.2240(0.0561)	0.2236(0.0510)	0.2002(0.0587)	0.2210(0.0555)

Table 5. Comparison of Paired Test Results for PET and ^{99m}Tc-sestamibi-SPECT

	^{99m}Tc-sestamibi-SPECT			
		Positive	Negative	Total
PET	Positive	4	4	8
	Negative	1	12	13
	Total	5	16	21

Figure 1: Box plots of coverage probabilities of the various two-sided 90% intervals

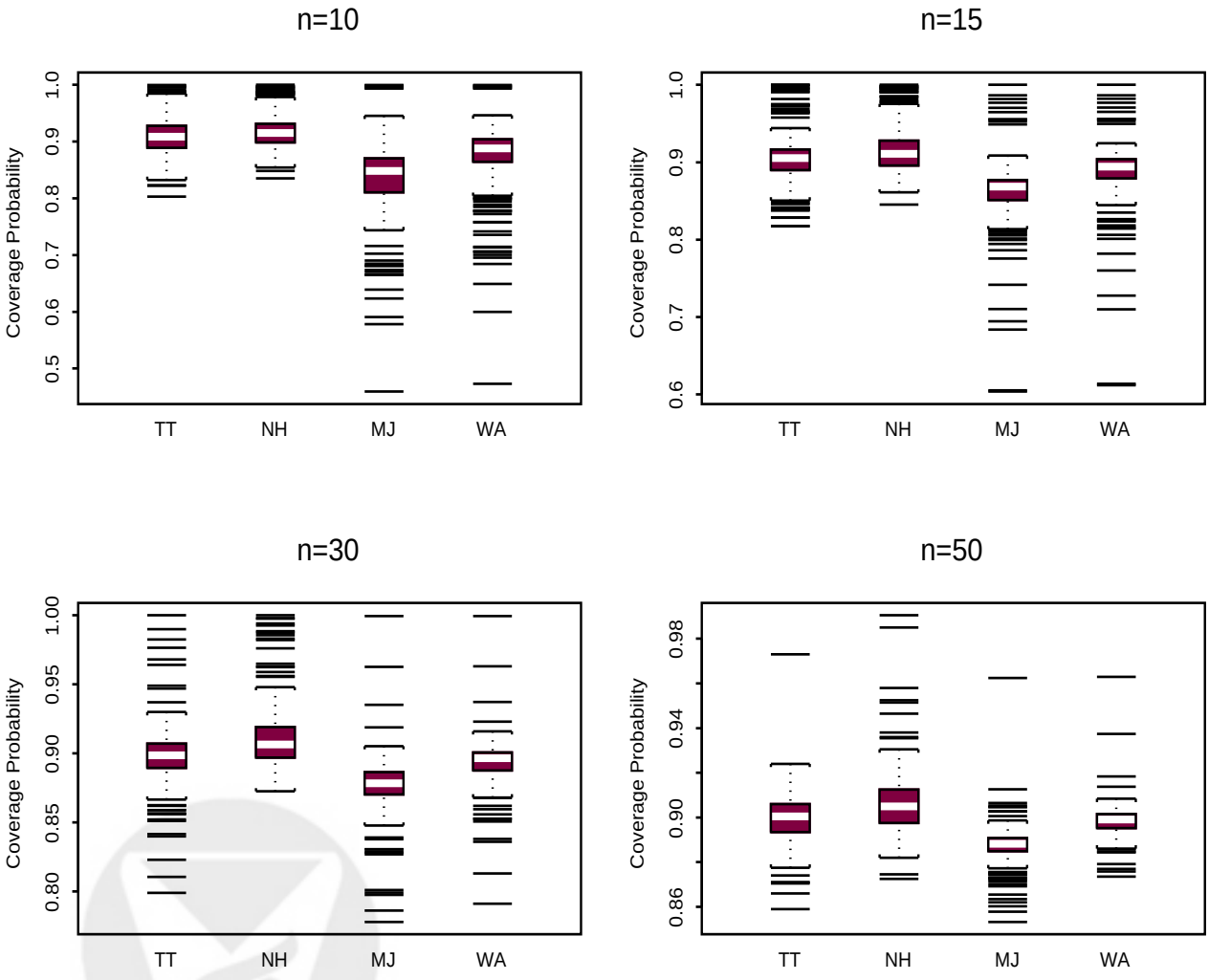

Figure 2: Box plots of coverage probabilities of the various two-sided 95% intervals

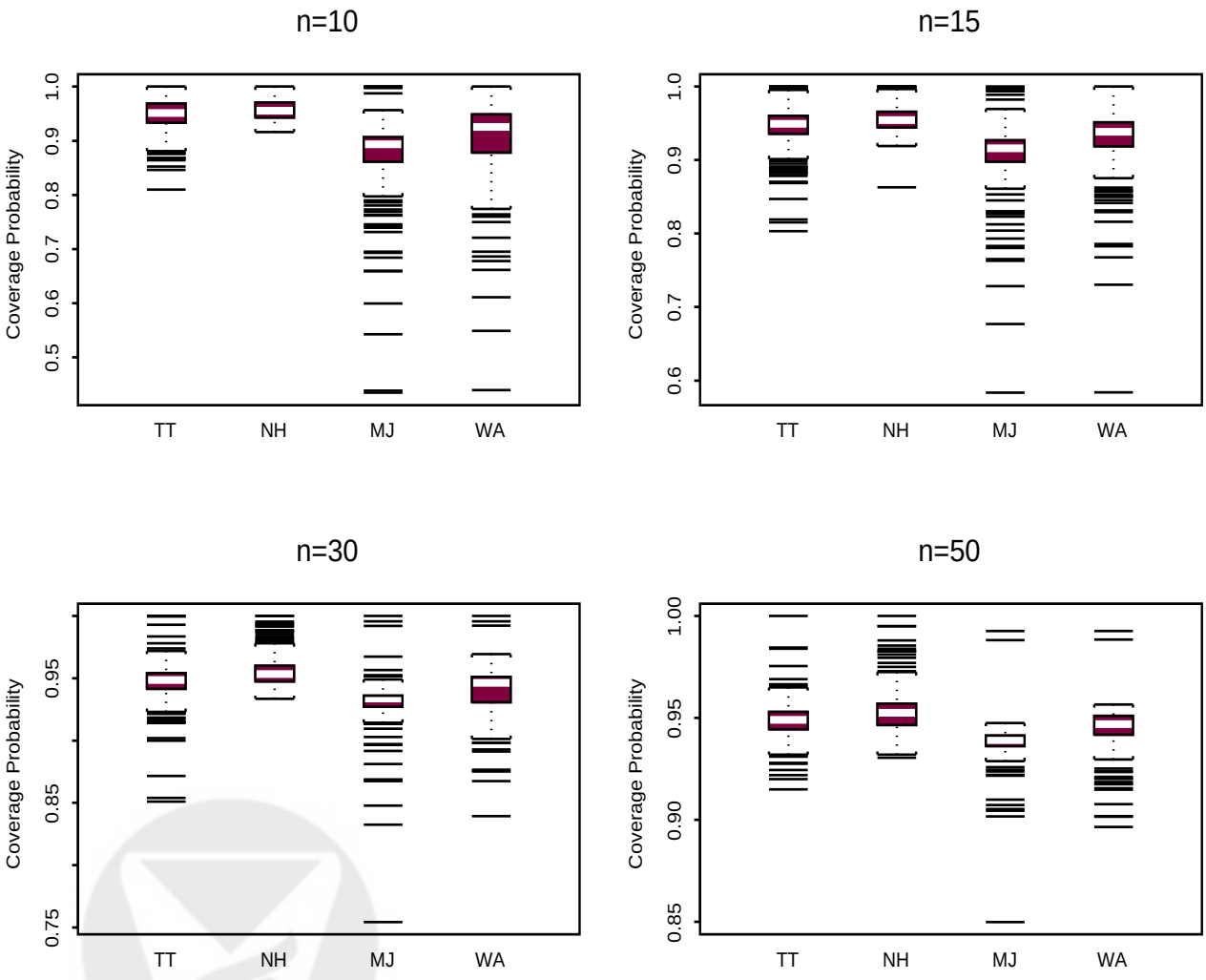

Figure 3: Box plots of expected interval lengths of the various two-sided 90% intervals

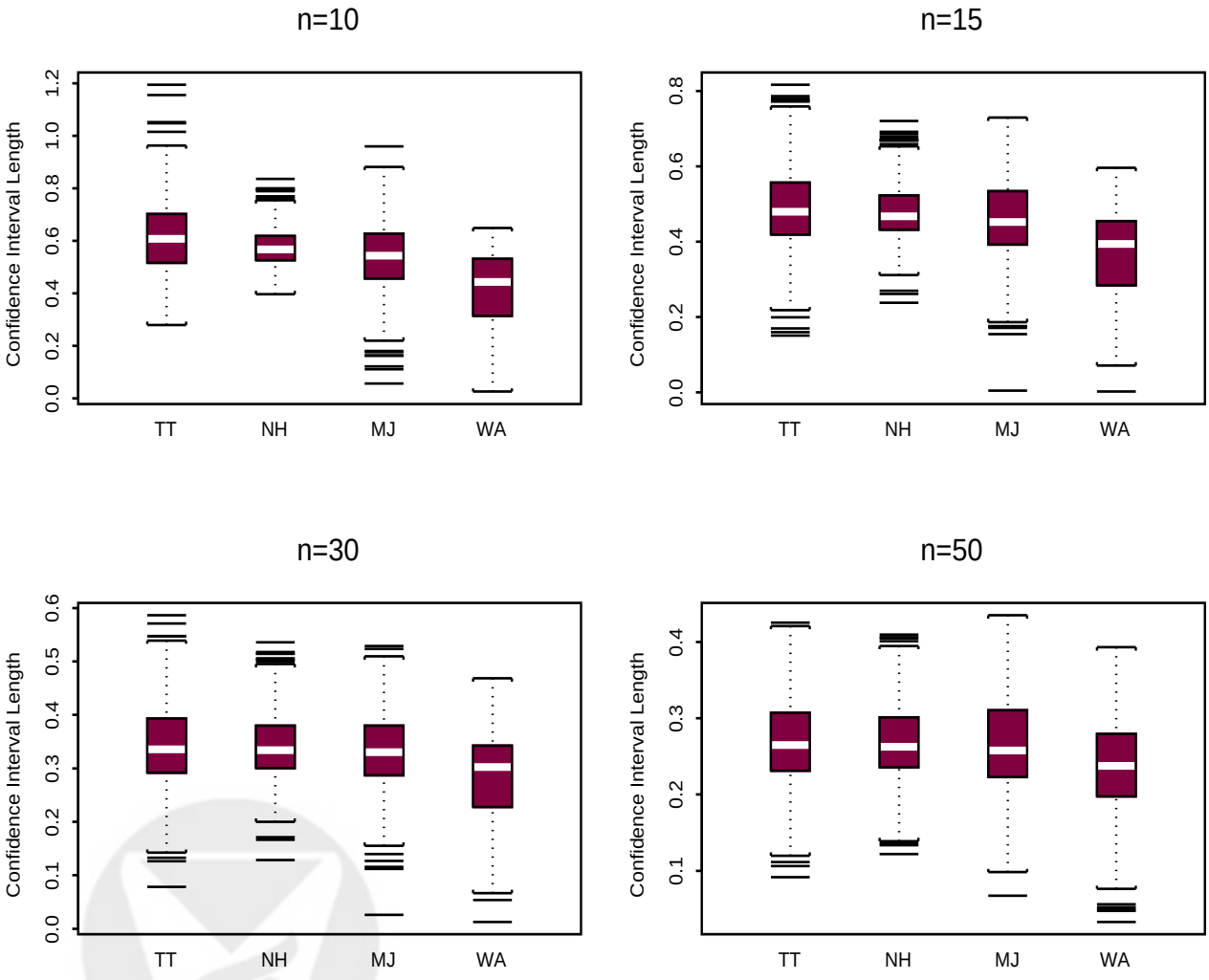

Figure 4: Box plots of expected interval lengths of the various two-sided 95% intervals