Computational Statistics & Data Analysis 40 (2002) 143 – 157
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Approximate confidence intervals for one proportion and difference of two proportions

Wei Pan*

Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building, Box 303, 420 Delaware Street SE, Minneapolis, MN 55455-0378, USA

Received 1 February 2001; received in revised form 1 November 2001; accepted 1 November 2001

Abstract

Constructing a confidence interval for a binomial proportion or the difference of two proportions is a routine exercise in daily data analysis. The best-known method is the Wald interval based on the asymptotic normal approximation to the distribution of the observed sample proportion, though it is known to have a bad performance for small to medium sample sizes. Recently, Agresti and his co-workers proposed an Adding-4 method: 4 pseudo-observations are added with 2 successes and 2 failures and then the resulting (pseudo-)sample proportion is used. The method is simple and performs extremely well. Here we propose an approximate method based on a t-approximation that takes account of the uncertainty in estimating the variance of the observed (pseudo-)sample proportion. It follows the same line of using a t-test, rather than z-test, in testing the mean of a normal distribution with an unknown variance. For some circumstances our proposed method has a higher coverage probability than the Adding-4 method.

Keywords: Binomial; Satterthwaite's method; t-distribution; Wald

1. Introduction

It is a common practice to construct a confidence interval for a binomial proportion or the difference of two proportions. For instance, in clinical trials it is often needed to investigate the difference of the cure rates of two treatments. Most introductory statistics textbooks only cover the Wald method, which is based on the asymptotically normal approximation to the distribution of the observed sample proportion(s). It is tempting

* Tel.: +1-612-626-2705; fax: +1-612-626-0660.
E-mail address: weip@biostat.umn.edu (W. Pan).

PII: S0167-9473(01)00107-4

to use the Wald method due to the familiarity and simplicity. However, it has been noted in the literature (e.g. Ghosh, 1979; Vollset, 1993; Newcombe, 1998a, b) that the Wald method may perform erratically for small to medium samples. Recently, Agresti and his co-workers (Agresti and Coull, 1998; Agresti and Caffo, 2000) proposed an approximate Adding-4 method: 4 pseudo-observations are added with 2 successes and 2 failures and then the resulting (pseudo-)sample proportion is substituted into the Wald interval. The method is simple and performs extremely well.

However, in some situations, the Adding-4 method may still have a coverage percentage smaller than a specified nominal level. We suspect that there is room for improvement. Since in the Wald method, the variability of the variance estimator of the sample proportion is ignored, a proper adjustment for this variability may improve the coverage rate. This is the approach we will pursue here. The basic idea is to apply Satterthwaite's method (Satterthwaite, 1941) to approximate the distribution of the variance estimator (of the sample proportion) using a scaled chi-square distribution, leading to a t-based interval, rather than a normal-based interval. The resulting confidence interval is simple to use. In particular, a closed form solution to the approximate degrees of freedom of the corresponding t-distribution can be derived. Numerical studies show its improvement over the Wald method and Adding-4 method. In the following, we first discuss interval estimation for a binomial proportion based on one observed sample, then for the difference of two binomial proportions based on two independent samples.

2. One binomial proportion

2.1. Methods

Suppose X is from a binomial distribution $bin(n, p)$. Our goal is to construct a $(1-\alpha)\%$ confidence interval for the parameter p. The most widely used or known is based on an asymptotic normal approximation to the distribution of $\hat{p} = X/n$

$$\text{Wald}: \quad \hat{p} \pm z_{\alpha/2}\sqrt{V(\hat{p}, n)},$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution, and

$$V(p, n) = p(1-p)/n$$

is the variance of $\hat{p}$. The above so-called Wald interval is known to perform terribly (e.g. Agresti and Caffo, 2000). A much better alternative is to use the *score* interval

$$\hat{p}\left(\frac{n}{n+z_{\alpha/2}^2}\right) + \frac{1}{2}\left(\frac{z_{\alpha/2}^2}{n+z_{\alpha/2}^2}\right)$$
$$\pm z_{\alpha/2}\sqrt{\frac{1}{n+z_{\alpha/2}^2}\left[\hat{p}(1-\hat{p})\left(\frac{n}{n+z_{\alpha/2}^2}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{z_{\alpha/2}^2}{n+z_{\alpha/2}^2}\right)\right]}.$$

Agresti and Coull (1998) noticed that $z_{0.025}^2 = 1.96^2 \approx 4$, and as a simplification proposed adding 4 pseudo-observations with one-half as successes and the other half as failures to obtain a modified estimator of p, $\tilde{p} = (X+2)/(n+4)$. Then their Adding-4 confidence interval is obtained by using $\tilde{p}$ in the Wald interval:

$$\text{Adding-4:} \quad \tilde{p} \pm z_{\alpha/2}\sqrt{V(\tilde{p}, n+4)}.$$

Its performance is surprisingly good.

However, we suspect that there may be some room for improvement. As in testing the mean of a normal distribution with an unknown variance, a t-test is better than a z-test since the former takes account of the uncertainty in estimating the variance of the estimated mean. A t-test is more conservative than a z-test, and hence is more likely to maintain the Type I error within the specified nominal level. We have a similar situation here. Recall that in the Wald interval, the variance of $\hat{p}$ is replaced by its *estimate* $V(\hat{p}, n)$, and $V(\hat{p}, n)$ is treated as fixed. Following the line of Satterthwaite (1941), we propose to approximate the distribution of $V(\hat{p}, n)$ (and similarly for $V(\tilde{p}, n+4)$) by a scaled chi-square distribution $c\chi_v^2$ with degrees of freedom v. c and v are derived by matching the first two moments of $V(\hat{p}, n)$ with that of $c\chi_v^2$. Then we have

$$c = \frac{var(V(\hat{p}, n))}{2E(V(\hat{p}, n))}, \quad v = \frac{2[E(V(\hat{p}, n))]^2}{var(V(\hat{p}, n))},$$

where $var(V(\hat{p}, n))$ can be calculated based on the first four moments of X (e.g. Johnson et al., 1993, p.107) as

$$\begin{aligned}\Omega(p, n) &= var(V(\hat{p}, n)) = var(X)/n^4 + var(X^2)/n^6 - 2\,Cov(X, X^2)/n^5 \\ &= (p - p^2)/n^3 + [p + (6n-7)p^2 + 4(n-1)(n-3)p^2 \\ &\quad - 2(n-1)(2n-3)p^3]/n^5 - 2[p + (2n-3)p^2 - 2(n-1)p^3]/n^4.\end{aligned}$$

Of course, in practice we can use the plug-in estimator $\Omega(\hat{p}, n)$.

Since $\hat{p}$ is asymptotically normal, and if we assume that $\hat{p}$ and $V(\hat{p}, n)$ are approximately independent, then

$$\frac{\hat{p} - p}{\sqrt{V(\hat{p}, n)}} = \frac{\hat{p} - p}{\sqrt{\frac{V(\hat{p},n)}{cv}cv}} = \frac{\hat{p} - p}{\sqrt{\frac{V(\hat{p},n)}{cv}\frac{var(V(\hat{p},n))}{2E(V(\hat{p},n))}\frac{2E(V(\hat{p},n))^2}{var(V(\hat{p},n))}}} = \frac{(\hat{p} - p)/\sqrt{E(V(\hat{p}, n))}}{\sqrt{\frac{V(\hat{p},n)}{cv}}}$$

approximately has a t-distribution t_v with degrees of freedom v, which can be approximated by

$$v \approx \frac{2V(\hat{p}, n)^2}{\Omega(\hat{p}, n)}.$$

Let $t_{v,\alpha}$ denote the $(1-\alpha)$ quantile of t_v. Our first proposed t-interval is

$$\text{T1}: \quad \hat{p} \pm t_{v,\alpha/2}\sqrt{V(\hat{p}, n)}.$$

As to be shown later, it is more desirable to construct the t-interval using $\tilde{p}$, leading to the second t-interval

$$\text{T2}: \quad \tilde{p} \pm t_{r,\alpha/2}\sqrt{V(\tilde{p}, n+4)},$$

where the degrees of freedom r can be approximated by

$$r \approx \frac{2V(\tilde{p}, n+4)^2}{\Omega(\tilde{p}, n+4)}.$$

Note that both v and r are in the order of n, implying that the T1 and T2 methods will reduce to the Wald and Adding-4 methods, respectively, as the sample size n tends to infinite. Hence the t-based methods can be regarded as finite sample adjustments for the Wald or Adding-4 intervals.

2.2. Evaluation

A simulation study was conducted to evaluate the performance of the above various methods. We restrict our attention to $\alpha = 0.05$. The coverage probability (CP) of any interval with a given n can be calculated as

$$CP = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} I_x,$$

where I_x indicates whether the confidence interval covers p or not when $X = x$. Note that for any given n and $X = x$, any of the above methods gives a fixed confidence interval (i.e. the two endpoints of the interval are not random). Similarly, the average width of any confidence interval is

$$AW = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} W_x,$$

where W_x is the width of the confidence interval when $X = x$.

Fig. 1 gives the CPs of the four methods for four different sample sizes $n = 5$, 10, 20 and 30. First, it is verified that the Adding-4 method performs as well as the score method. Second, the bad performance of the Wald interval is obvious. Third, the T1 method improves over the Wald method but still does not work well.

The trouble of both the T1 and Wald methods is largely caused by $\hat{p} = 0$ or 1 when $X = 0$ or n, leading to $V(\hat{p}, n) = 0$ and thus a zero-width of the resulting interval. In view of the good performance of the Adding-4 method, in the T1 and Wald methods we replace $\hat{p}$ by $\tilde{p}$ if and only if $X = 0$ or n. The results are presented in Fig. 2. It is obvious that, compared with that in Fig. 1, the performance of either method has improved. However, since there is still some under-coverage when p is near $\frac{1}{2}$, and that $\tilde{p}$ can be interpreted as a weighted average of $\hat{p}$ and $\frac{1}{2}$, we decide to use the T2 method and compare it with the Adding-4 method. This will be the focus of the remaining discussion.

Fig. 3 presents the results. It is observed that the T2 method has some improvement over the Adding-4 method in terms of having CP not smaller than the specified nominal level. This is obvious for p near 0 or 1 when $n=5$. This is not surprising since $t_{v,\alpha} < z_\alpha$ for any finite v. This also implies that the T2 interval is wider and more conservative than the Adding-4 interval. Fig. 4 compares their interval widths, and the difference is not huge.

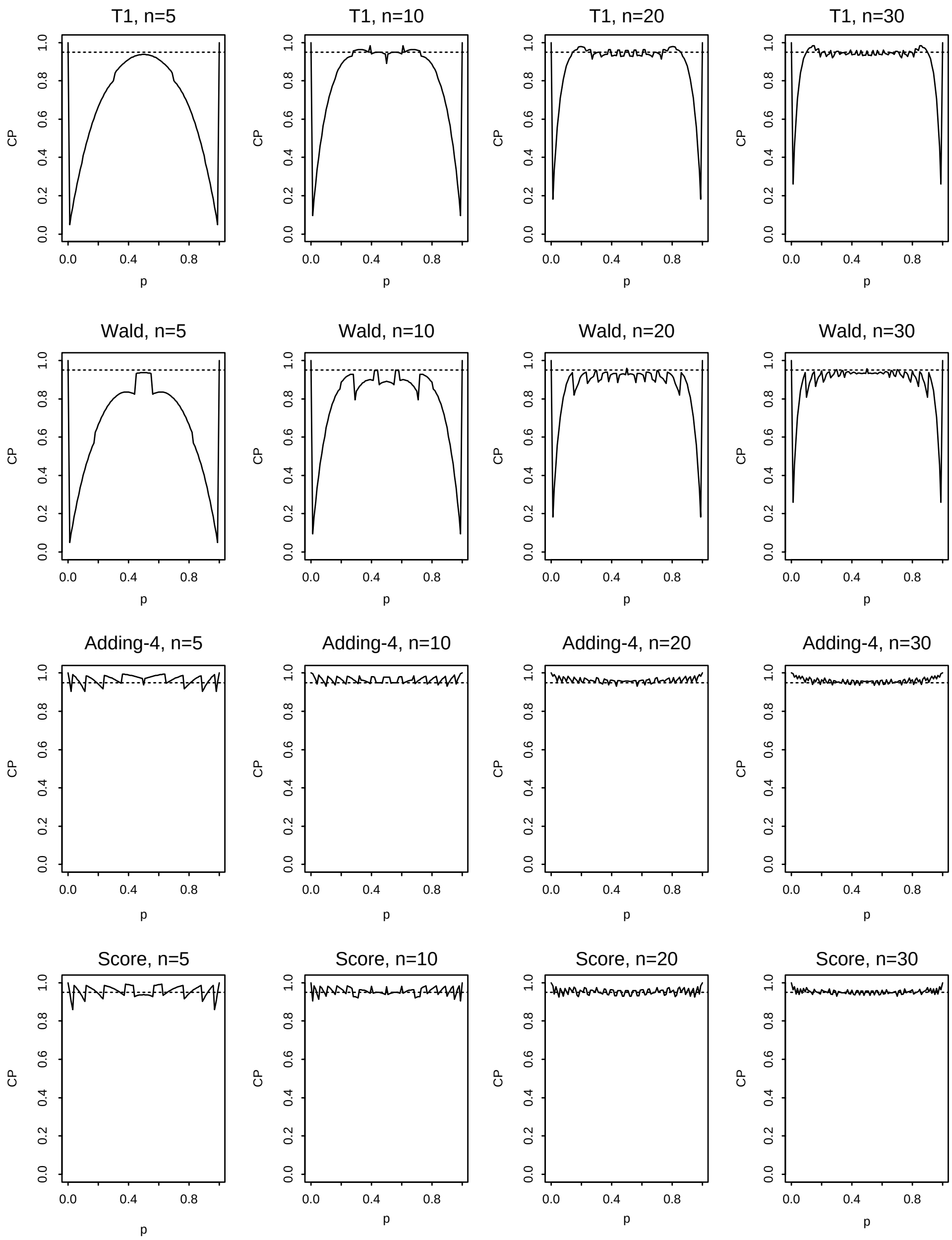

Fig. 1. Coverage probability (CP) of the four methods for a binomial proportion p with sample size n.

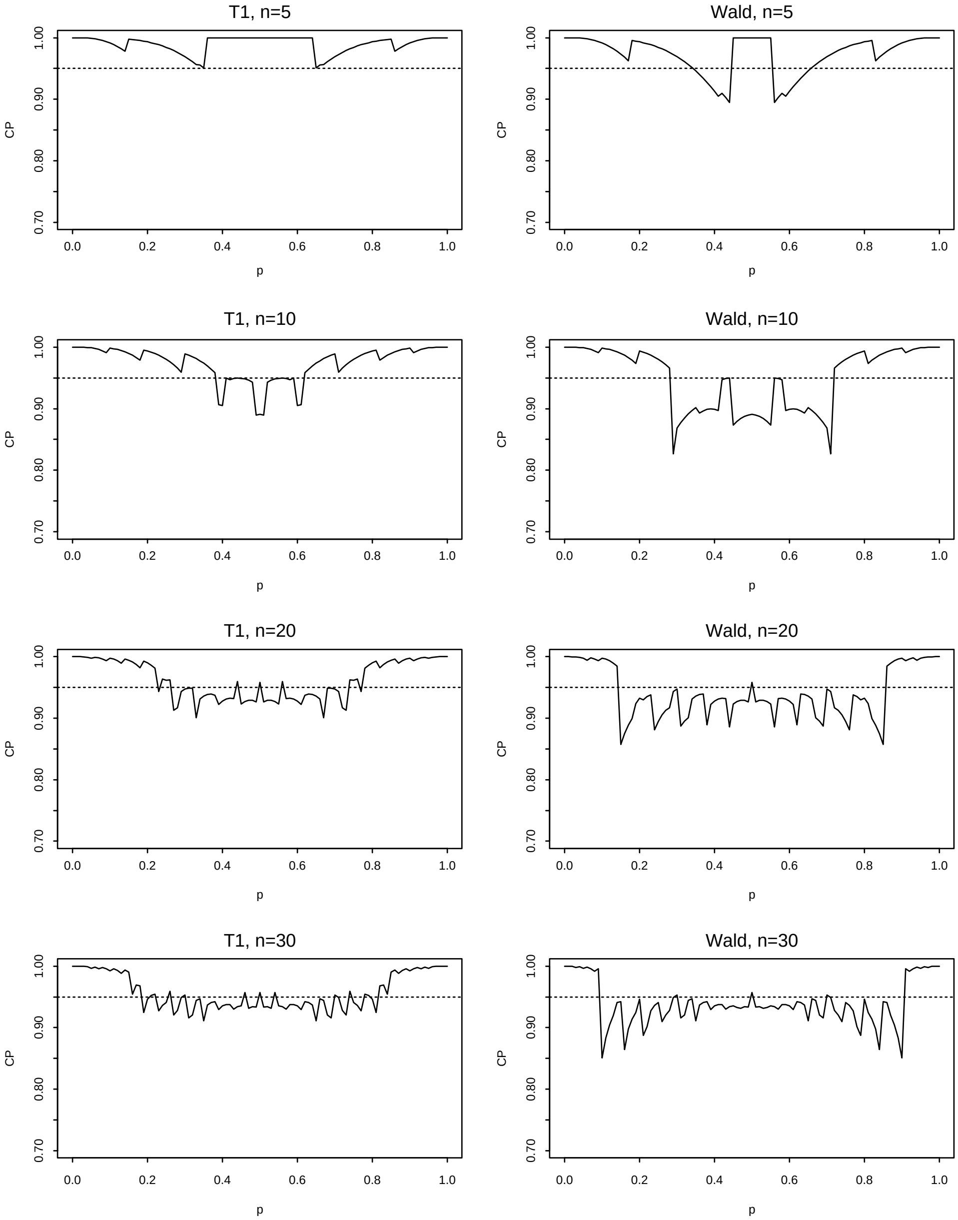

Fig. 2. Coverage probability (CP) of the modified T1 and Wald methods (adding 2 successes and 2 failures if $X = 0$ or n) for a binomial proportion p with sample size n.

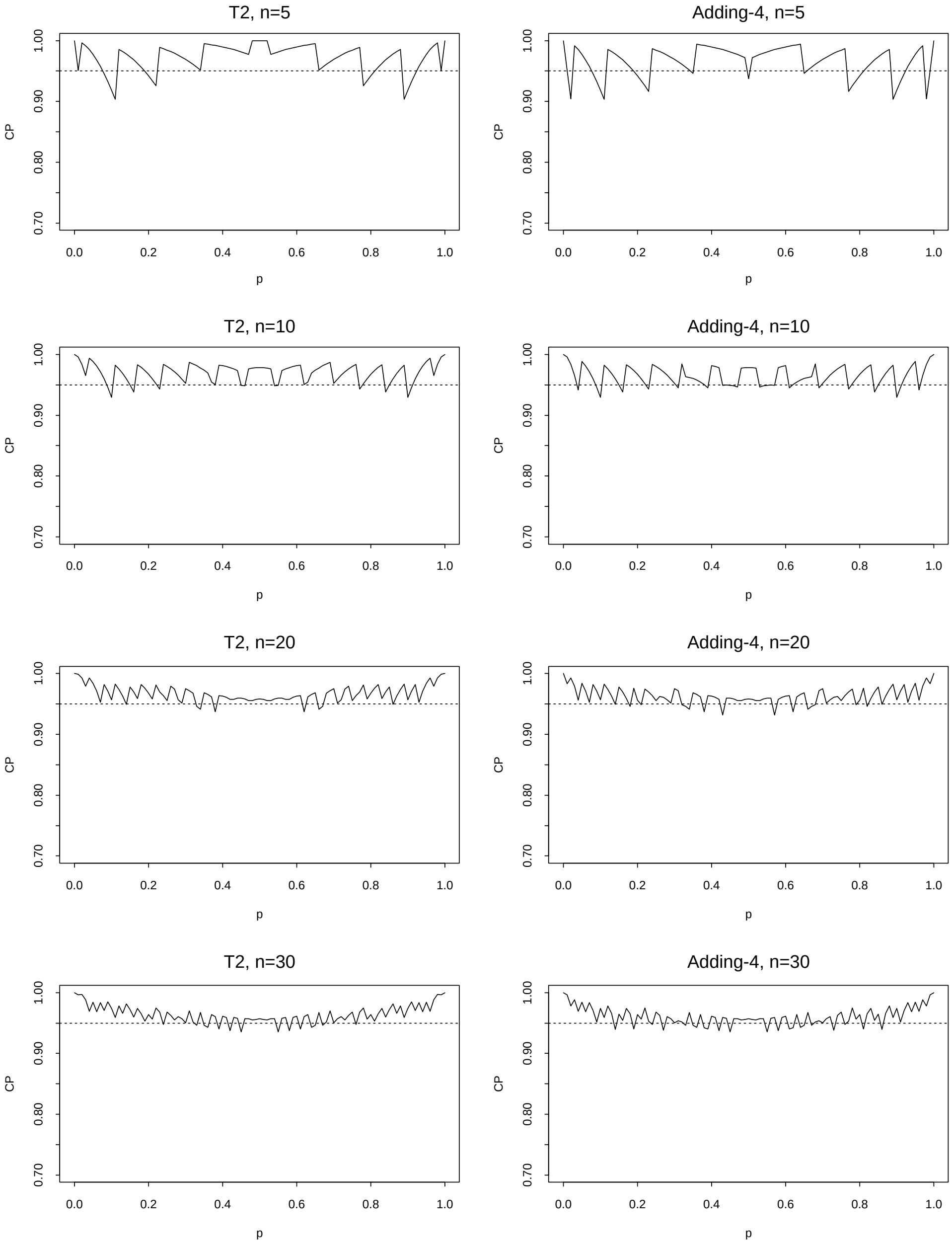

Fig. 3. Coverage probability (CP) of the T2 and Adding-4 methods for a binomial proportion p with sample size n.

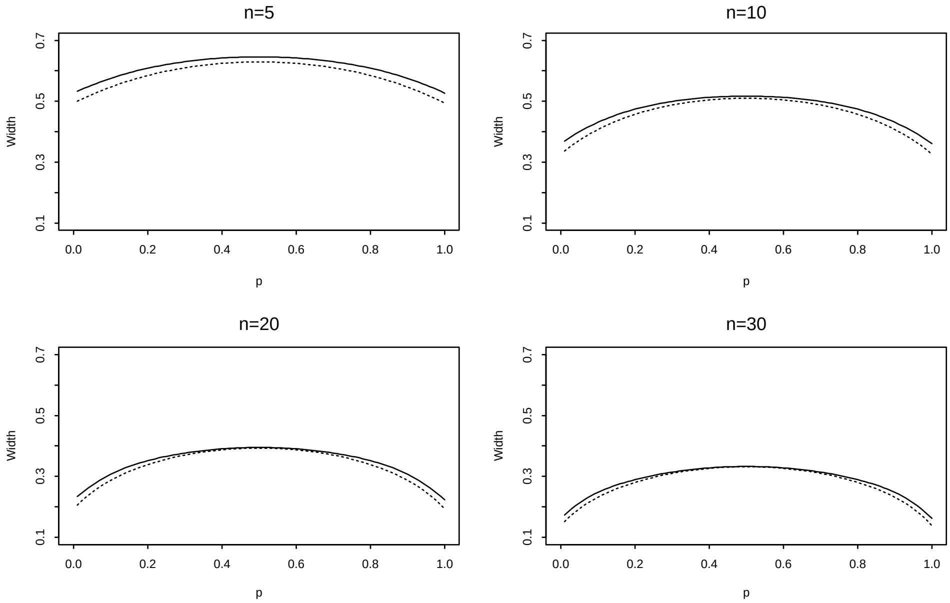

Fig. 4. Average width of the T2 (with solid lines) and Adding-4 intervals (with dotted lines) for a binomial proportion p with sample size n.

3. Difference of two proportions

3.1. Methods

Suppose that now we observe two independent binomial variables: $X_1 \sim bin(n_1, p_1)$ and $X_2 \sim bin(n_2, p_2)$. The goal is to construct a $(1-\alpha)$ level confidence interval for $p_1 - p_2$. The Wald interval is

$$\text{Wald}: \quad \hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{V(\hat{p}_1, n_1) + V(\hat{p}_2, n_2)},$$

where $\hat{p}_1 = X_1/n_1$ and $\hat{p}_2 = X_2/n_2$. Its performance is not satisfactory, as for one binomial proportion. The score interval can be also extended but it lacks a closed form here. Agresti and Caffo (2000) generalize the Adding-4 method as

$$\text{Adding-4}: \quad \tilde{p}_1 - \tilde{p}_2 \pm z_{\alpha/2}\sqrt{V(\tilde{p}_1, n_1+2) + V(\tilde{p}_2, n_2+2)},$$

where $\tilde{p}_i = (X_i + 1)/(n_i + 2)$ for $i = 1, 2$.

Our t-interval can be similarly applied here:

$$\text{T2}: \quad \tilde{p}_1 - \tilde{p}_2 \pm t_{d,\alpha/2}\sqrt{V(\tilde{p}_1, n_1+2) + V(\tilde{p}_2, n_2+2)},$$

where the degrees of freedom is

$$d \approx \frac{2[V(\tilde{p}_1, n_1+2) + V(\tilde{p}_2, n_2+2)]^2}{\Omega(\tilde{p}_1, n_1+2) + \Omega(\tilde{p}_2, n_2+2)}.$$

3.2. Evaluation

Figs. 5–7 present some numerical results for various sample sizes and $p_2 = 0.1$, 0.3 and 0.5, respectively. It can be seen that the Adding-4 method works extremely well, but in several places the T2 method improves the coverage probability over it, especially when p_1 is close to 1.

4. Other comparisons

The basic idea of our proposed t-based method is general and can be applied to other tests. Here we illustrate its use to yield a modified score method to construct a confidence interval for a binomial proportion. In addition, we compare the performance of the modified method with another modified score method, the continuity-corrected score interval, which has been found to have good performance and is a recommended method in the literature (Vollset, 1993).

The score interval presented in Section 2.1 can be regarded as a normal-based interval with the form: the point estimate plus/minus $z_{\alpha/2}\times$ SE. Hence, rather than using the standard normal-based coefficient, we can use the t-coefficient to construct a t-based interval

$$\text{TS}: \quad \hat{p}\left(\frac{n}{n+z_{\alpha/2}^2}\right)+\frac{1}{2}\left(\frac{z_{\alpha/2}^2}{n+z_{\alpha/2}^2}\right) \pm t_{r,\alpha/2}\sqrt{V_s(\hat{p},n)},$$

where

$$V_s(\hat{p},n)=\frac{1}{n+z_{\alpha/2}^2}\left[\hat{p}(1-\hat{p})\left(\frac{n}{n+z_{\alpha/2}^2}\right)+\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{z_{\alpha/2}^2}{n+z_{\alpha/2}^2}\right)\right],$$

and the degrees of freedom r is approximated by

$$r \approx \frac{2V_s(\hat{p},n)^2}{\Omega_s(\hat{p},n)}$$

with $\Omega_s(\hat{p},n)=\Omega(\hat{p},n)n^4/(n+z_{\alpha/2}^2)^4$ and $\Omega(\cdot,\cdot)$ is given in Section 2.1.

The score interval has much better performance than the Wald interval. An even better one is the so-called continuity-corrected score interval (Vollset, 1993)

$$\text{cc-Score}: \quad \frac{(x \pm \frac{1}{2})+\frac{z_{\alpha/2}^2}{2} \pm z_{\alpha/2}\sqrt{x \pm \frac{1}{2}-\frac{(x\pm 1/2)^2}{n}+\frac{z_{\alpha/2}^2}{4}}}{n+z_{\alpha/2}^2}.$$

Fig. 8 presents the coverage probabilities of the score interval and its two modified versions. It is confirmed that both modified versions have better performance than the score interval. The t-based method may still have some under-coverage while the continuity-corrected score interval almost always has a coverage probability larger than the nominal level (but may be too conservative). Fig. 9 gives the average widths of the confidence intervals. It can be seen that the t-based interval is only slightly wider than the score interval. In contrast, the continuity-corrected score interval is much wider,

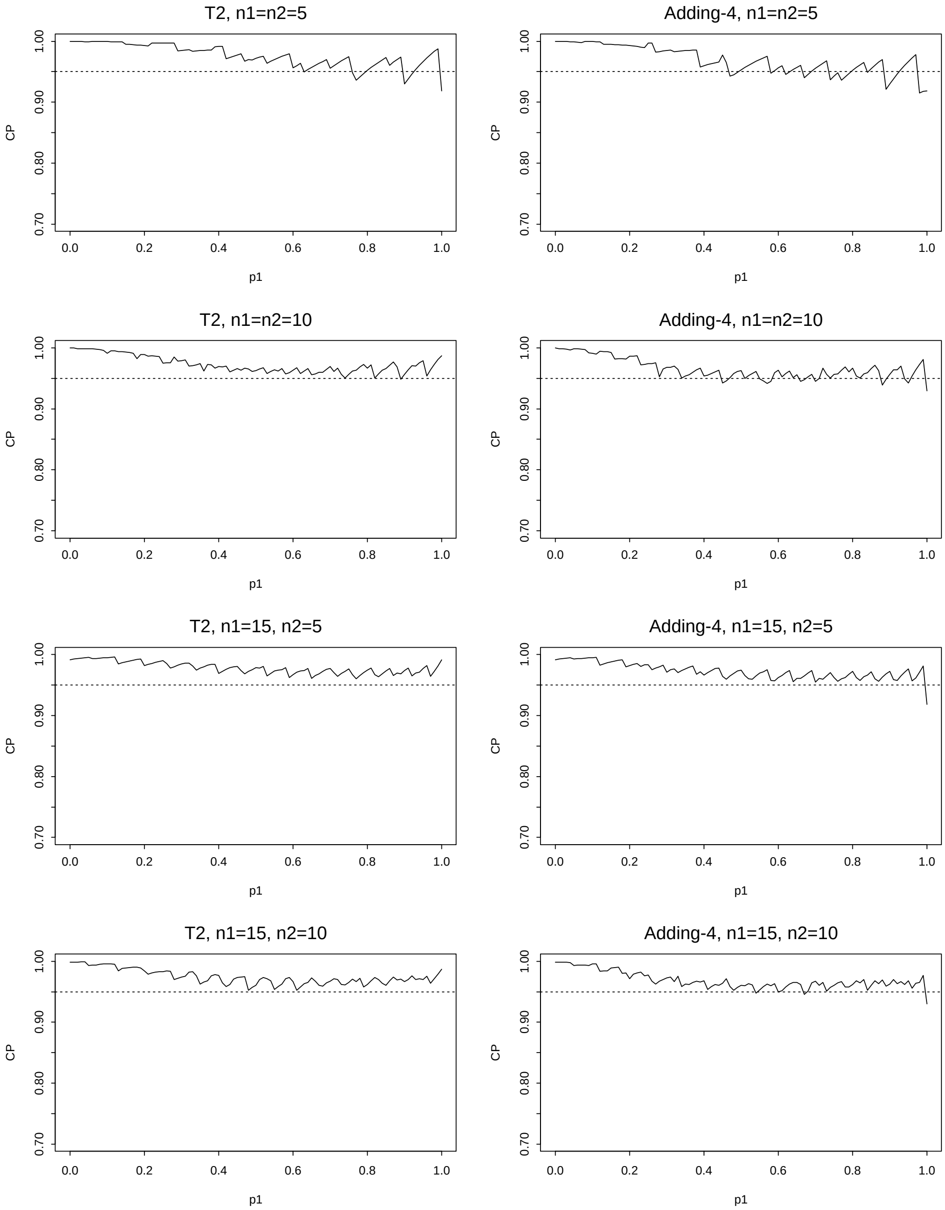

Fig. 5. Coverage probability (CP) of the T2 and Adding-4 methods for the difference of two binomial proportions, $p_1 - p_2$, with p_2=0.1 and sample sizes n_1 and n_2.

especially for small sample sizes. In addition, comparing Figs. 8 and 9 with Figs. 3 and 4, we can also see that both the T2 and Adding-4 methods are also competitive when compared with the continuity-corrected score interval. In summary, the two t-based

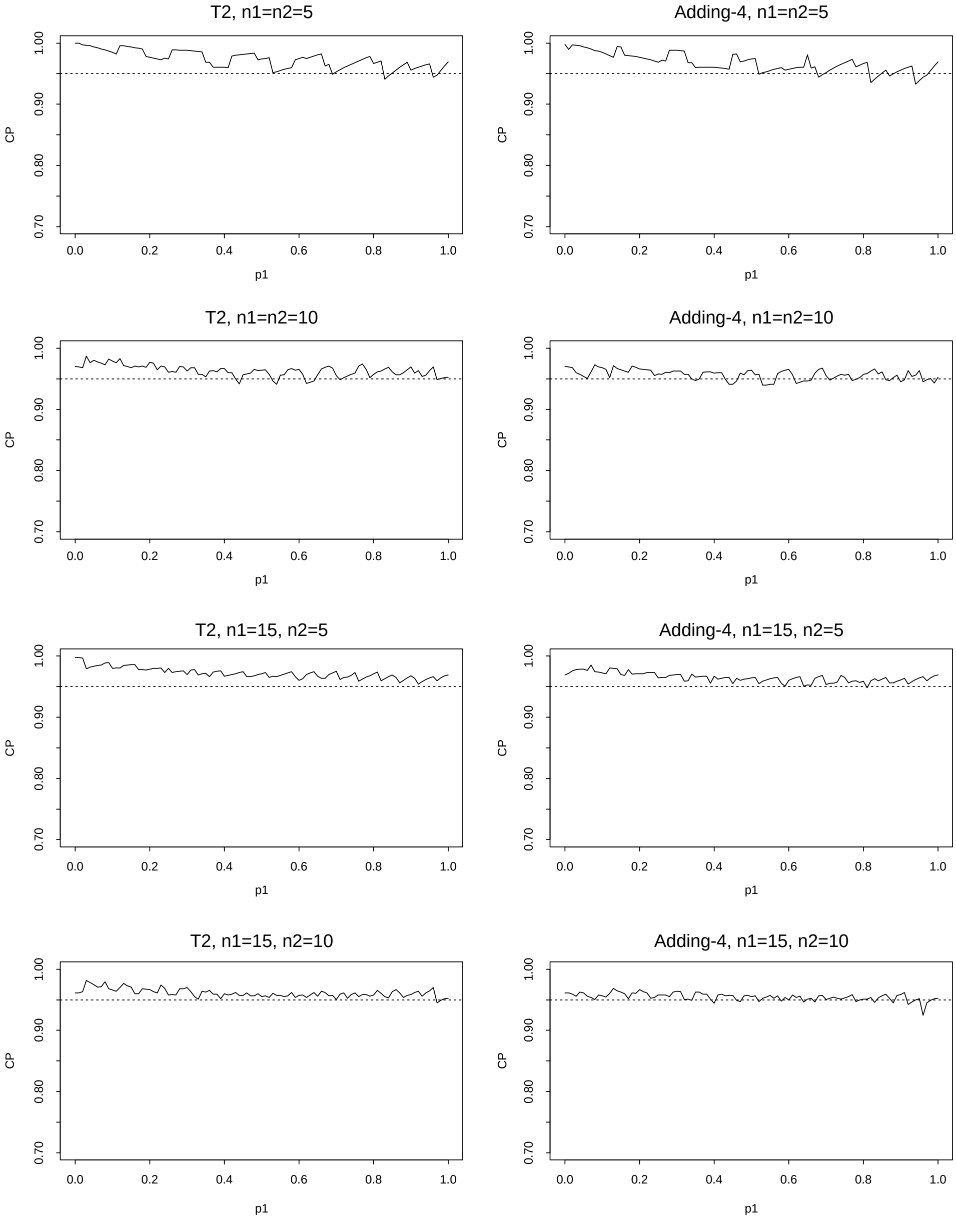

Fig. 6. Coverage probability (CP) of the T2 and Adding-4 methods for the difference of two binomial proportions, $p_1 - p_2$, with $p_2 = 0.3$ and sample sizes n_1 and n_2.

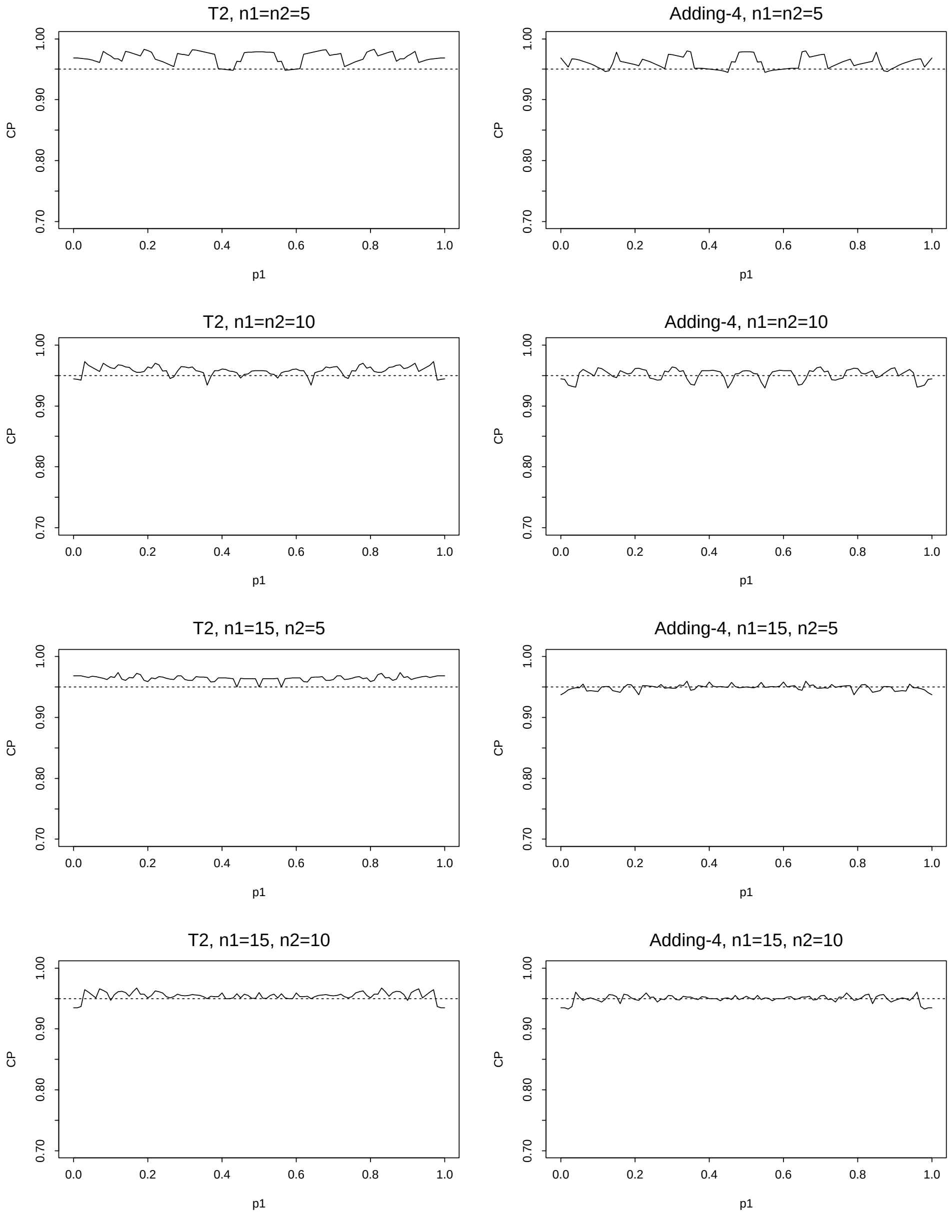

Fig. 7. Coverage probability (CP) of the T2 and Adding-4 methods for the difference of two binomial proportions, $p_1 - p_2$, with $p_2 = 0.5$ and sample sizes n_1 and n_2.

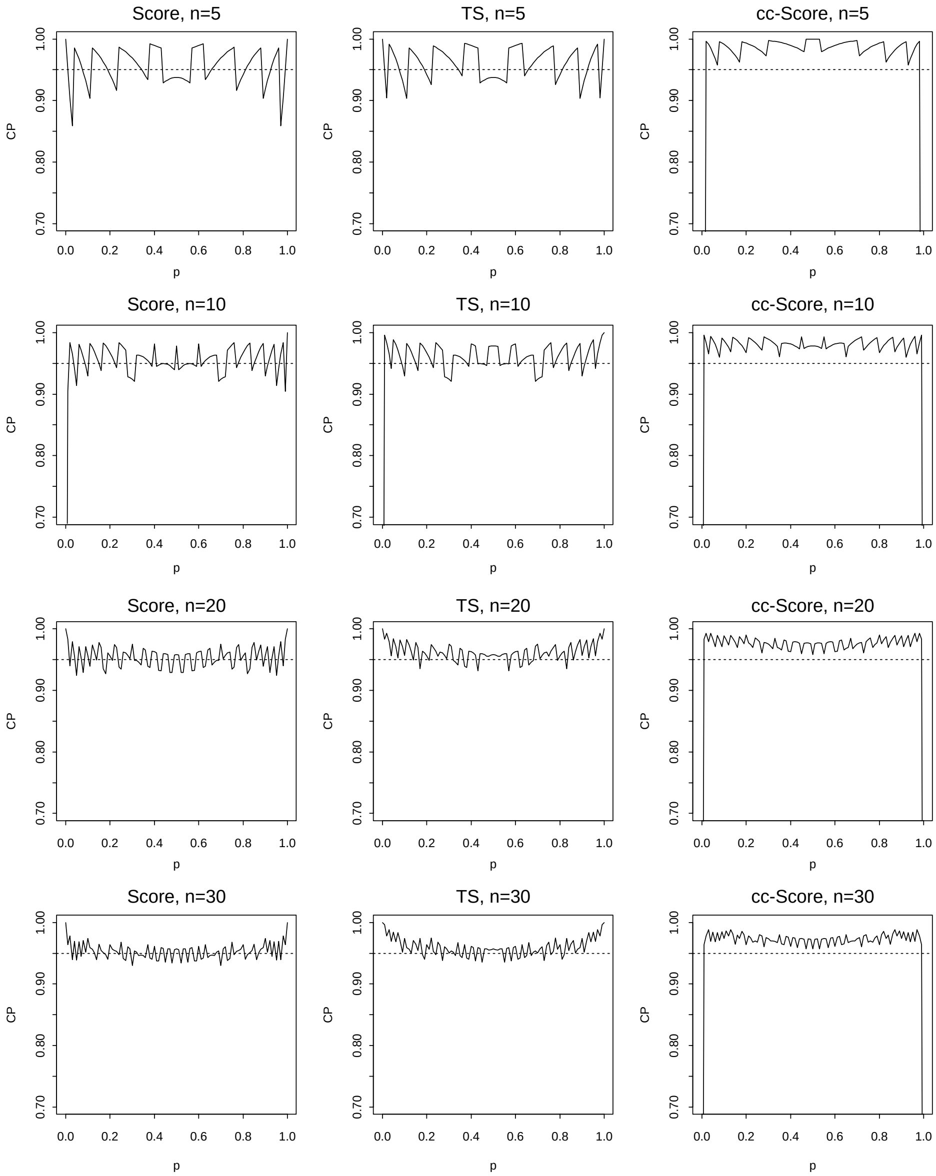

Fig. 8. Coverage probability (CP) of the score, TS and continuity-corrected score methods for a binomial proportion p with sample size n.

intervals (T2 and TS), appear to be promising methods that may strike a favorable balance between a high coverage probability and a short interval. In particular, they can be interesting alternatives to the commonly recommended continuity-corrected score interval.

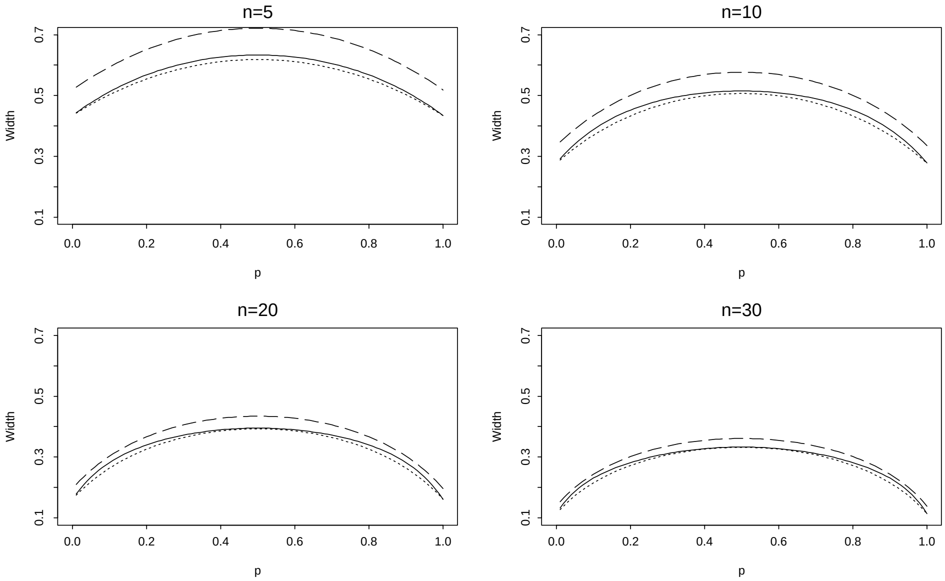

Fig. 9. Average width of the TS (with solid lines), score (with dotted lines) and continuity-corrected score (with dash lines) intervals for a binomial proportion p with sample size n.

5. Discussion

We have proposed approximate t-based confidence intervals for a single proportion and for the difference of two proportions, built on the point estimator (of a proportion or the difference of two) by adding 4 pseudo-observations and an approximate t-distribution of the standardized point estimator (i.e. the point estimator divided by its estimated standard error). The method has a similar form to that of the Adding-4 method proposed by Agresti and co-workers, except that a t quantile, rather than a standard normal quantile, is used as the coefficient in constructing the confidence interval. The idea of using a t distribution, rather than a standard normal distribution to approximate the distribution of a standardized point estimator is not new. Satterthwaite (1941) proposed the general idea almost 60 years ago, and it has been used in many other problems, but to our knowledge, not in our current context. Here the t-based method is simple to use since there is a closed form for the approximate degrees of freedom of the corresponding t-distribution. We found that in some situations our proposed method can have a higher coverage probability than the Adding-4 method, which in general is satisfactory. Of course, the price we pay for the t-based method is the resulting wider confidence intervals. Though the improvement of the t-based method over that of the Adding-4 method is not dramatic, due to the common use of confidence intervals and the minimum extra-effort needed in implementing the t-based method, we believe it is worthwhile using the t-based method. Furthermore, the idea of using the t-based method is important and general. It provides a framework to adjust for the asymptotically normal inference with finite samples in other more complex settings. For instance, Pan and Wall (2002) extended this basic idea to approximate inference in the context of using the sandwich variance estimator in generalized estimating equations. Further applications, such as to other generalized linear models, are worth future investigation.

Acknowledgements

The author would like to thank a referee and the editors for helpful comments.

References

Agresti, A., Caffo, B., 2000. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. Amer. Statist. 54, 280–288.

Agresti, A., Coull, B.A., 1998. Approximate is better than “exact” for interval estimation of binomial proportions. Amer. Statist. 52, 119–126.

Ghosh, B.K., 1979. A comparison of some approximate confidence intervals for the binomial parameter. J. Amer. Statist. Assoc. 74, 894–900.

Johnson, N.L., Kotz, S., Kemp, A.W., 1993. Univariate Discrete Distributions, 2nd Edition. Wiley, New York.

Newcombe, R., 1998a. Two-sided confidence intervals for the single proportion: comparison of seven methods. Statist. Med. 17, 857–872.

Newcombe, R., 1998b. Interval estimation for the difference between independent proportions: comparison of eleven methods. Statist. Med. 17, 873–890.

Pan, W., Wall, M.M., 2002. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. Statist. Med., in press.

Satterthwaite, F.F., 1941. Synthesis of variance. Psychometrika 6, 309–316.

Vollset, S.E., 1993. Confidence intervals for a binomial proportion. Statist. Med. 12, 809–824.