Communications of the Korean Statistical Society
2009, Vol. 16, No. 2, 363–372

Choosing between the Exact and the Approximate Confidence Intervals: For the Difference of Two Independent Binomial Proportions

Seung-Chun Lee[1,a]

[a]Dept. of Statistics, Hanshin Univ.

Abstract

The difference of two independent binomial proportions is frequently of interest in biomedical research. The interval estimation may be an important tool for the inferential problem. Many confidence intervals have been proposed. They can be classified into the class of exact confidence intervals or the class of approximate confidence intervals. One may prefer exact confidence intervals in that they guarantee the minimum coverage probability greater than the nominal confidence level. However, someone, for example Agresti and Coull (1998) claims that "approximation is better than exact." It seems that when sample size is large, the approximate interval is more preferable to the exact interval. However, the choice is not clear when sample size is small. In this note, an exact confidence and an approximate confidence interval, which were recommended by Santner *et al.* (2007) and Lee (2006b), respectively, are compared in terms of the coverage probability and the expected length.

Keywords: Exact confidence interval, approximate confidence interval, coverage probability, expected length.

1. Introduction

The interval estimation for the difference of two independent binomial proportions is often of prime important in biology, medicine, and many other fields of scientific research. For instance, many experiments in clinical trials are designed to compare the difference in proportions of responses to a specific endpoint between a new treatment and an existing treatment. The interval estimation plays key role in these statistical problems. Variety of estimation methods has been devised. They can be classified into the approximate interval and the exact interval, but it seems that the approximate method dominates the exact method when sample size is large.

The approximate confidence interval is based on the large sample property, and hence believed to have poor characteristic in that the actual coverage probabilities are frequently and significantly less than nominal level when sample size is small. It was a common sense in my literature review that an approximate confidence interval should not be considered in the serious statistical problem with a small sample. See Santner and Yamagami (1993), Newcombe (1998), and Santner *et al.* (2007).

The term, "exact" opposes to "approximate" and is originated from the definition of the confidence set. In literatures, see Lehmann (1986) for example, a random set $S(X)$ is called a confidence set at confidence level $1 - \alpha$ if

$$P_\theta[\theta \in S(X)] \geq 1 - \alpha, \quad \text{for all } \theta \in \Theta. \tag{1.1}$$

This work was supported by Hanshin University research grant.

[1] Professor, Department of Statistics, Hanshin University, 411 Yangsan-Dong, Osan, Kyunggi-do 447-791, Korea. E-mail: seung@hs.ac.kr

A confidence interval satisfying (1.1) is called exact. Thus the exact confidence interval guarantees the minimum coverage probability. However, because of the discreteness of underlying distribution, the coverage probability oscillates unlike the continuous case. Thus, to guarantee the minimum coverage, the confidence interval should be conservative. *i.e.*, it should be unnecessarily wide. Because, wide intervals are non-informative, one may not prefer the exact confidence interval.

Another drawback of the exact confidence interval is feasibility. Most popular statistical packages provide only a few exact methods. Although StatXact provides the greatest scope for small sample inference in discrete problems, its province is limited as well. For instance, for the difference of two independent binomial proportions, Santner *et al.* (2007) evaluated various exact confidence intervals in terms of their coverage probability and expected length, and recommended the interval devised by Coe and Tamhane (1993), but it is not incorporated in StatXact. Thus, one may have difficulty in using proper exact methods. Even, the worst situation is when sample size is large. Usually exact methods require a heavy computational work, the computation is feasible only when sample size is moderately large even in these days. Thus the comparison between two methods may be meaningless when sample size is large.

The Wald interval using the maximum likelihood estimate of binomial parameter has been considered as a standard method for the interval estimations of binomial proportions. However, the erratic behavior of the coverage probability of the Wald interval has been recognized in various literatures. See for example, Blyth and Still (1983), Agresti and Coull (1998) and Brown *et al.* (2001). In particular, Brown *et al.* investigated the unsatisfactory coverage properties of the Wald interval in details. However, some approximate confidence intervals work quite well even with a small sample. For example, Agresti and Coull (1998) showed that an improved interval for the parameter of a binomial distribution could be obtained by so-called "adding two successes and two failures" to the observed counts and then using the standard method. This strategy works quiet well in various sampling designs as well as in the 1-group design. For instance, Agresti and Caffo (2000) examined the interval estimation for the difference of two binomial proportions, and concluded that the strategy performs about as well as the best available methods in this 2-group design. Price and Bonett (2004) extended the Agresti-Coull type interval for general k-group design. In addition, Lee (2006a), Lee (2006b) and Lee (2007) provided the weighted Polya posterior confidence interval for the 1-group, 2-group and k-group designs. It is believed that both the Agresti-Coull type and the weighted Polya posterior intervals are comparable to the exact confidence interval in terms of the coverage probability when sample size is small. However, it was shown in Lee (2006a) and his subsequent papers that the weighted Polya posterior intervals outperformed the corresponding Agresti-Coull type intervals.

Obviously, the choice of proper confidence interval is important for certain statistical problems. For instance, Chan and Zhang (1999) gave an interesting example in a vaccine clinical trial to investigate whether a new manufacturing process provides improvement over the current process. The preliminary data showed that the proportions of subjects responding to the vaccine were .944(17 subjects out of 18) and .611(11 subjects out of 18) for the new and current processes, respectively. Since the difference between two proportions was .333, it seemed that the new process gained noticeable improvement. However, the exact 95% exact confidence interval due to Santner and Snell (1980) yielded an interval of $(-0.019, 0.630)$ indicating no significant improvement. Interestingly enough, applying the weighted Polya posterior method to the same data, an interval of $(0.056, 0.547)$ is obtained. Thus, we can draw a different conclusion.

Both two estimates have defects. The actual coverage probability of Santner and Snell's interval is greater than .95, and hence the length of interval may be too wide. While the actual coverage probability of the Polya posterior can be less than the nominal level, and hence the length of interval may not be proper.

The choice depends on our allowance for the coverage probability and the interval length. If you think that the coverage probability should be at least greater than the nominal level for your statistical problem regardless of interval length, an exact confidence interval is your choice. However, Agresti and Coull (1998) argued that most researchers probably interpret the confidence coefficients in terms of "average performance" rather than "worst possible performance." That is, a more relevant description of performance is the long-run percentage of times that the procedure is correct when it is used repeatedly for a variety of data sets in various problems with possibly different parameter values. If you agree with Agresti and Coull, and think that the interval length is important as far as the coverage probability of an approximate interval is not significantly lower than the nominal level, than an approximate confidence interval could be your choice. I believe that most researchers are between two.

There may be a compromised loss for helping to choose a proper confidence interval. In this note, I consider a loss or utility for judging the goodness of confidence intervals to help the choice of proper interval, and compare two representative confidence intervals, one for the exact and one for the approximate, in terms of the loss or utility.

2. Confidence Intervals

2.1. Exact confidence intervals

Suppose that $X_1 \sim B(n_1, p_1)$ and $X_2 \sim B(n_2, p_2)$ are two independent binomial random variables. Many exact confidence intervals for $\Delta = p_1 - p_2$ have been constructed by inverting the hypothesis test for $H_0 : \Delta = \Delta^*$ versus $H_1 : \Delta \neq \Delta^*$, where $\Delta^* \in (-1, 1)$. That is, apply an exact size α test for each Δ^* to collect acceptance regions. Since the acceptance regions depend upon the nuisance parameter p_1, eliminate the p_1 effect on the acceptance regions somehow. The set of acceptance regions are used to construct an exact $(1 - \alpha) \times 100\%$ confidence interval.

Various tests and various methods to eliminate the p_1 effect can be applied for the testing problem. For example, Chan and Zhang (1999) used one-side exact test based on score statistics,

$$S(X) = \frac{\hat{p}_1 - \hat{p}_2 - \Delta^*}{\sqrt{\tilde{p}_1(1 - \tilde{p}_1)/n_1 + \tilde{p}_2(1 - \tilde{p}_2)/n_2}},$$

where $\hat{p}_i$ and $\tilde{p}_i$ for $i = 1, 2$ are the maximum likelihood and the restricted maximum likelihood estimates under the restriction $p_1 - p_2 = \Delta^*$ of p_i's, respectively. The nuisance parameter p_1 is eliminated toward conservatism. Agresti and Min (2001) used the same score statistics but applied two-side exact with the similar method of eliminating the nuisance parameter.

Coe and Tamhane (1993) developed different approach for the exact confidence interval. They used greedy heuristics to construct acceptance sets that contain as few points of sample values as possible for the testing problem. Coe/Tamhane method partition the Δ-space $\equiv (-1, 1)$ into a finite number of equi-spaced grid $-1 \leq \Delta_{-M} < \Delta_{-M+1} < \cdots < 0 = \Delta_0 < \Delta_1 < \cdots < \Delta_M \leq 1$, and then, for each i, partition the p_1-space $[\Delta_i, 1]$ by $0 \leq p_{i1} < \cdots < p_{iN_i}$ systematically about the midpoint $(1 + \Delta_i)/2$. Given $p_1 = p_{ij}$ and $p_2 = \Delta_i - p_{ij}$, collect most probable sample points until the sum of probabilities of the sample points is greater than $1 - \alpha$. After some refinements of the collected sample points, an exact $(1 - \alpha) \times 100\%$ confidence interval for Δ can be formed. You may refer Coe and Tamhane (1993) or Santner *et al.* (2007) for further detail of the refinements.

Santer and Yamagami (1993) used the same greedy heuristics, but the collection method for acceptance region is different from the Coe/Tamhane method. The Santner/Yamagami and Coe/Tamhane methods produce different systems of intervals with substantially different properties.

Santner *et al.* (2007) compared the four exact confidence intervals, Chan/Zhang(CZ), Agresti/Min (AM), Santner/Yamagami(SY) and Coe/Tamhane(CT), and an approximate confidence interval (AS) which is based on the score statistics, as proposed by Wilson (1927) for a single binomial proportion, and is proposed by Miettinen and Nuriminen (1985). They recommended the CT because its expected length is superior to other three exact intervals. Although the expected length of the AS is shorter than that of the CT, the AS is excluded in their recommendation because it fails to achieve the nominal coverage roughly 50% of 100 × 100 combinations of parameter values of (p_1, p_2). In other words, the coverage probability is of prime interest and conservative intervals are preferable in their comparison. In this point of view, it is hard to recommend the approximate confidence interval.

Beside the conclusion, there may be rounding errors in the calculation of the CT interval. They wrote that when $n_1 = n_2 = 15$, the coverage of CT intervals was less the nominal 90% in 36 of the 10,000 combinations. When $n_1 = n_2 = 30$, there were 56 cases failed to achieve the nominal level. These results are different from my calculations. On my calculations, the coverage probabilities of the CT intervals were all greater than the nominal level. To confirm my calculations, I asked the C++ program (available from the authors) implemented by the authors. It produced slightly different confidence intervals from mine. Because the CT interval is exact, I concluded that my calculations are correct. However, the rounding errors are not serious to affect their conclusions.

2.2. Approximate confidence intervals

Although the Wald interval is the well-known approximate confidence interval, various researchers have shown that it behaves very poorly even with moderate sample sizes. The poor performance of the Wald interval is not occurred by its length. In fact, when p_i's are near 0.5, it tends to be too wide, but the coverage probabilities are significantly lower than the nominal level. Now it is well-known that the Wald interval should not be used for a serious statistical problem especially when sample size is small.

Many approximate confidence intervals have been proposed. Among them Newcombe (1998) compared eleven methods, and recommended the AS with and without the continuity correction. The continuity correction gives slightly better performance. However, Lee (2006b) showed that the weighted Polya posterior interval(WP) outperforms the AS in the closeness of coverage probability to the nominal level and the expected length. In fact, he compared three approximate intervals including the Agresti/Caffo(AC) due to Agresti and Caffo (2000) in terms of the coverage probability and the expected length, and concluded that the WP is superior to the other two in both the closeness of coverage probability and the expected length.

The WP interval is an Wald-like confidence interval and is given by

$$(\tilde{p}_1 - \tilde{p}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{\tilde{n}_1 + 1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{\tilde{n}_2 + 1}}, \tag{2.1}$$

where $\tilde{n}_i = n_i + z^2_{\alpha/2}/2$ and $\tilde{p}_i = (X_i + z^2_{\alpha/2}/4)/\tilde{n}_i$ for $i = 1, 2$. The origin of the WP is due to Meeden (1999).

It has long been known that for small sample sizes, when sampling from a skewed distribution, the usual interval estimation of the mean covers the true values less often than their nominal level (Meeden, 1999). The situation does not much differ from the interval estimation of a binomial proportion p. Most problems occur mainly when sample size is not large enough or when p is near 0 or 1. Many authors have attempted to solve the problem, and one solution was given by Meeden. He showed that a modification of the Polya posterior, called weighted Polya posterior, gave interval estimators with improved coverage properties for the interval estimation of the mean.

Lee (2006a) applied the weighted Polya posterior to the problem of interval estimation of a binomial proportion, and showed that the weighted Polya posterior is eventually the posterior distribution with a conjugate beta prior distribution. Thus, the Agresti-Coull interval might be considered as the interval estimation based on the weighted Polya posterior. However, the Agresti-Coull interval is little bit unnecessarily wide in the Bayesian perspective not only for the 1-group design but also for those designs mentioned before. He also recognized that a better choice of the weight might be possible. In fact, the Wilson interval discussed by Wilson (1927) was shown to have slightly superior performance to the Agresti-Coull interval in the study of Brown *et al.* (2001). In this note, he adjusted the weight so that the center of the interval is same that of the Wilson interval, and shorten the interval by the Bayesian perspective. These two adjustments improved the performance in the 1-group design (Lee, 2006a), the 2-group design (Lee, 2006b), and the more general design (Lee, 2007). I believe that the WP is one of the best available approximate confidence interval.

The main aim of this paper is to show that the WP is comparable to the exact confidence interval in practice. For this purpose, I will compare the performance of the WP and the CT intervals.

3. The criterion for comparison

Let $\mathrm{CI}_\mathrm{A}(X_1, X_2)$ be a confidence interval due to method A. The usual criterions for judging A are its coverage probability and expected length which are defined as

$$C_\mathrm{A}(p_1, p_2) = \sum_{x_1=0}^{n_1} \sum_{x_2=0}^{n_2} I(\Delta \in \mathrm{CI}(x_1, x_2))\, p_{X_1,X_2}(x_1, x_2)$$

and

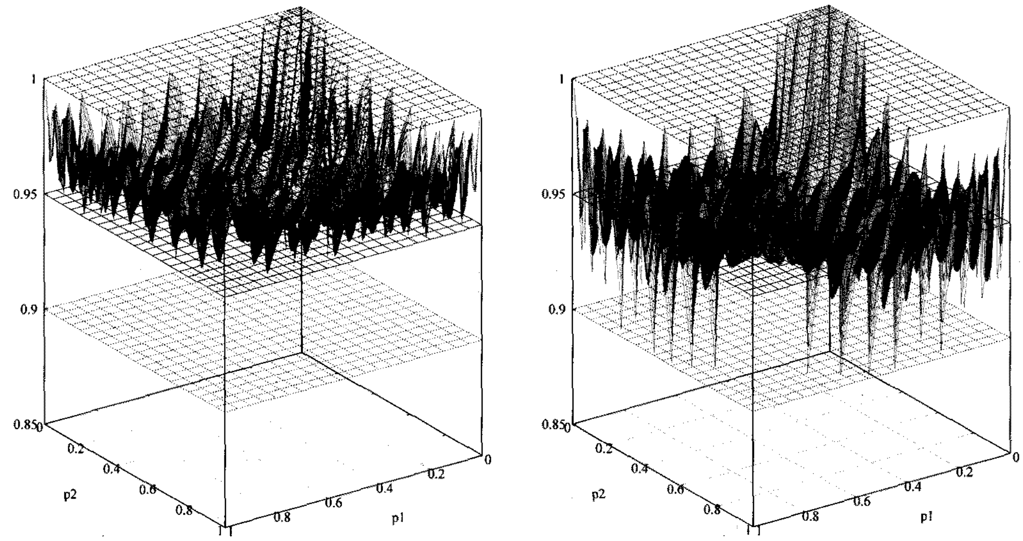
$$\mathrm{EL}_\mathrm{A}(p_1, p_2) = \sum_{x_1=0}^{n_1} \sum_{x_2=0}^{n_2} \mathrm{len}(\mathrm{CI}(x_1, x_2)) p_{X_1,X_2}(x_1, x_2),$$

respectively, where I is the usual indicator function, $\mathrm{len}(\mathrm{CI}(x_1, x_2))$ represents is the length of interval of $\mathrm{CI}(x_1, x_2)$, and $p_{X_1,X_2}(x_1, x_2)$ is the joint probability mass function of X_1 and X_2.

When comparing between exact confidence intervals, the expected length is the main factor for choosing proper exact confidence intervals. On the other hand, an approximate confidence interval is good, if the coverage probabilities are as closed to the nominal level as possible, the Mean Absolute Error(MAE) or the Mean Squared Error(MSE) of the coverage probabilities could be excellent criterions for comparing between approximate confidence intervals. When we compare between an exact and an approximate confidence intervals, the situation is slightly different.

Figure 1 shows the general pattern of the coverage probabilities of the two intervals when sample size is small. Since the CT is an exact confidence interval, the coverage probabilities of the CT are greater than the nominal 0.95, while the coverage probabilities of the WP oscillate around the nominal level. On the exact side, the coverage probability is good at the sacrifice of the expected length, and vice versa on the approximate side. It would be desirable to consider the two factors simultaneously for fair comparison. Thus, I consider the following function to help for choosing between exact and approximate confidence intervals.

$$L(\mathrm{A}, \lambda, p_1, p_2) = [(1 - \alpha) - C_\mathrm{A}(p_1, p_2)] + \lambda \times \mathrm{EL}_\mathrm{A}(p_1, p_2). \tag{3.1}$$



(a) CT interval

(b) WP interval

Figure 1: *Coverage probabilities for the difference of two independent binomial proportions with the nominal 95% CT interval and WP interval, for* $n_1 = n_2 = 10$.

$L(\mathrm{A}, \lambda, p_1, p_2)$ tends to be small value when $C_A(p_1, p_2)$ is large, but $\mathrm{EL_A}(p_1, p_2)$ is small. Thus it could be considered as an expected loss function, but is not a usual expected loss function in the sense that it can theoretically have negative values. Note, however, that a linear combination of the coverage probability and the expected length with a negative coefficient was considered as a risk for various set estimation problems. In fact, (3.1) is a variant of (2.1) of Casella *et al.* (1994). Here, λ is the weight between the coverage probability and the expected length. A large value of λ is favorable to the approximate confidence intervals, while a small value forces to choose the exact confidence interval. Since the first part of $L(\mathrm{A}, \lambda, p_1, p_2)$ can ranges from $-\alpha$ to $1 - \alpha$ and the expected length is in $[0, 2]$, it is believed that $\lambda = 1/2$ gives the fair weight for the coverage probability and the expected length.

4. Results

For the comparison of the CT and the WP, seven sample size cases are considered. These include three balanced cases $\{(n_1, n_2)|(5,5), (15,15), (30,30)\}$, and four unbalanced cases $\{(n_1, n_2)|(15,5), (30,5), (30,15), (30,25)\}$. Although the cases that n_1's are greater than or equal to n_2's are considered, the case n_1 is less than n_2 can be inferred from these results.

The confidence level considered in this comparison is 95%, since it is the most popular level of confidence. Coverage probabilities, expected lengths and expected losses of 95% CT and WP were calculated over 10,000 equally spaced (p_1, p_2) values in $[0, 1] \times [0, 1]$ for seven sample cases. Figure 2 and 3 demonstrate the distribution of them using Box-Percentile plots (Esty and Banfield, 2003).

It can be observed that the coverage probabilities of the WP are acceptable in the sense that only a few coverage probabilities, 948 out of 10,000, less than 1%, are less than or equal to 0.92 even when sample sizes are $n_1 = n_2 = 5$. In these sample sizes, the distribution of the coverage probabilities of

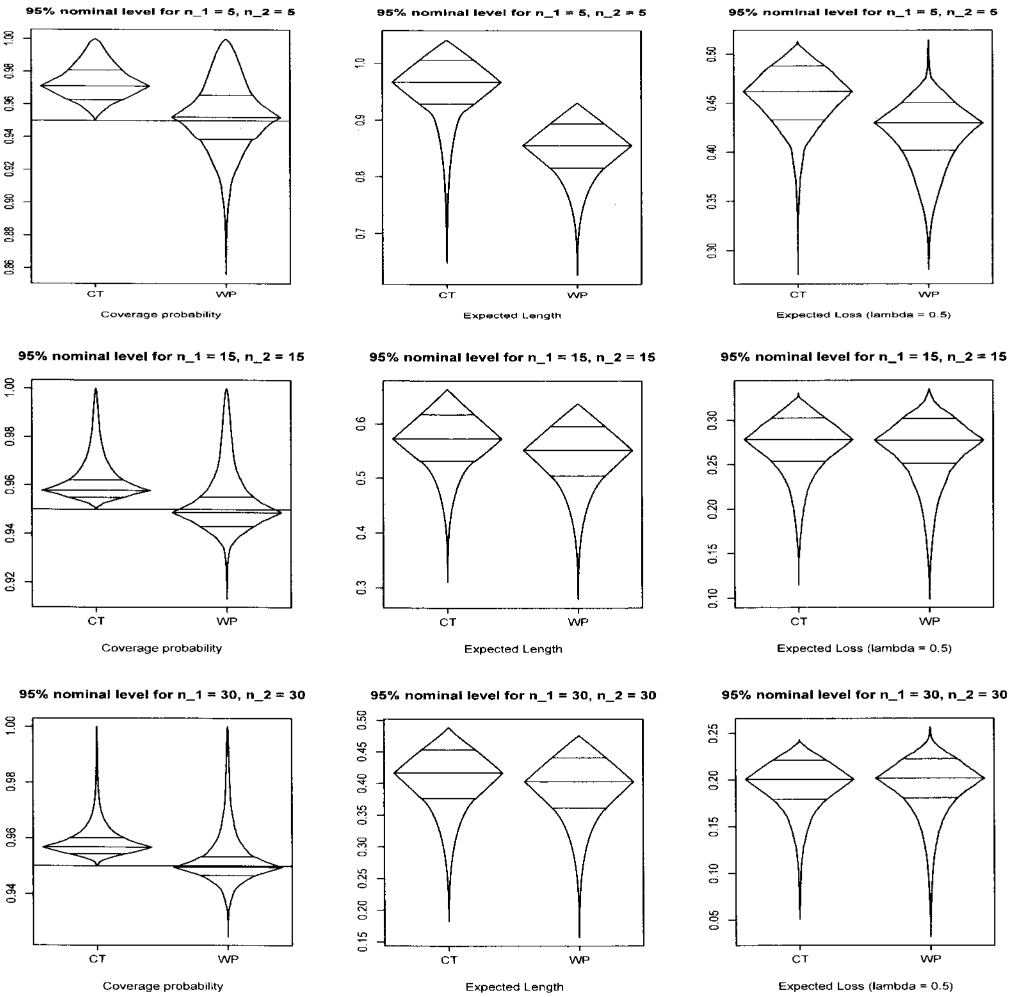


Figure 2: *Distributions of coverage probabilities, expected lengths and expected losses for CT and WP over 10,000 equally spaced* (p_1, p_2) *values in* $[0, 1] \times [0, 1]$ *(three balanced sample cases)*

CT is near symmetric at around 0.975, but skewed to the nominal level as sample sizes are increasing. That is, the CT is more conservative when sample sizes are small, and as a result, the expected length and the expected loss are big. In view of the expected loss, the WP is preferable until the λ is reduced to 0.182. With this value of weight, the means of the expected loss of CT and WP are same.

On the other hand, when sample sizes are relatively large and balanced, for instance $n_1 = n_2 = 30$, the distributions of expected losses of CT and WP are near identical. Thus, the expected loss does not give much information for choosing a confidence interval. Note however, it takes too much time to calculate the CT interval when sample size is large. In fact, I first implemented the CT interval by R 2.6.2. (2008), but the computing time was unendurable. The second implementation by C++ saved much time, but I am still wondering whether the CT interval can be practically usable when sample sizes are greater than or equal 30.

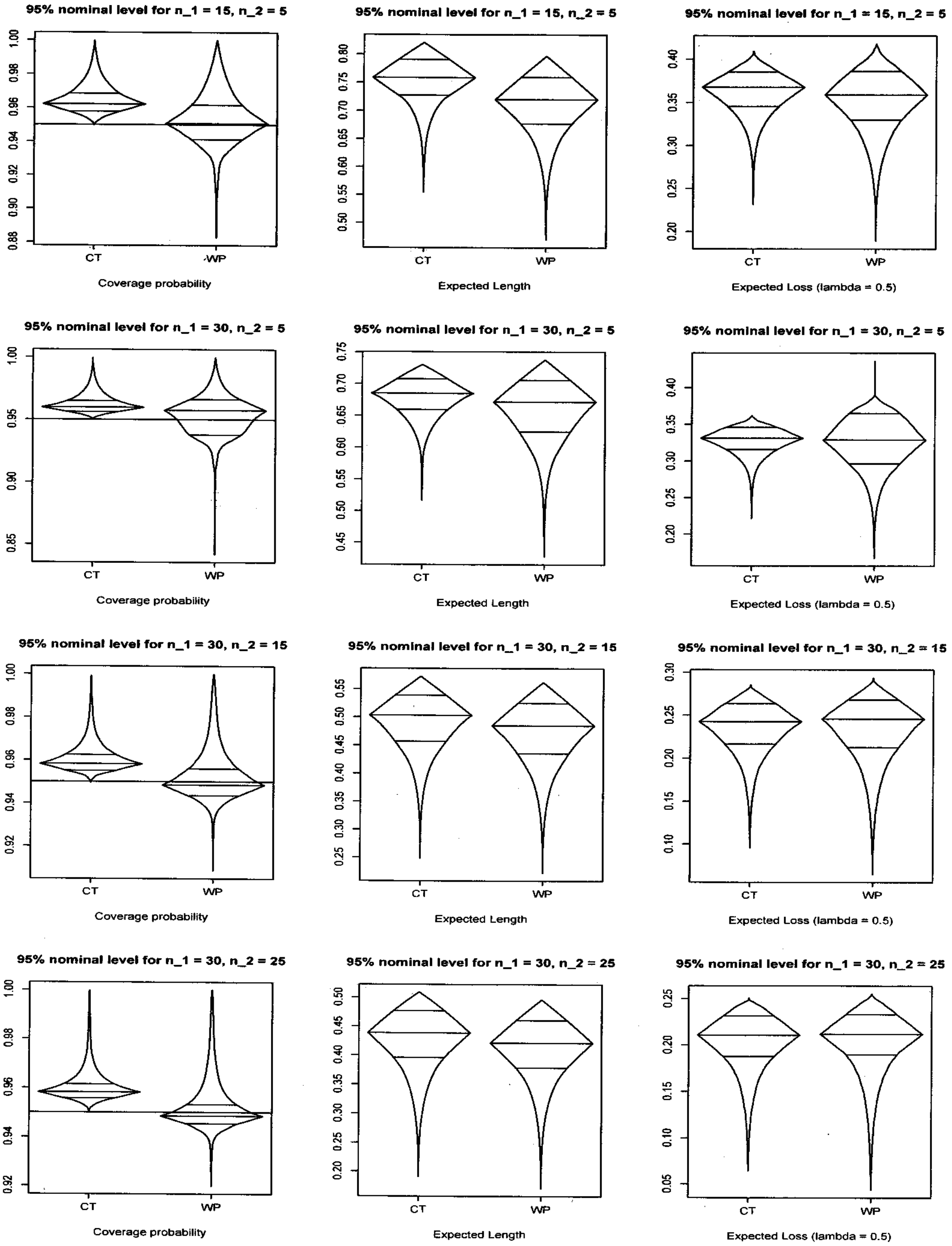


Figure 3: *Distributions of coverage probabilities, expected lengths and expected losses for CT and WP over 10,000 equally spaced* (p_1, p_2) *values in* $[0, 1] \times [0, 1]$ *(four unbalanced sample cases)*

When sample sizes are extremely unbalanced, $n_1 = 30, n_2 = 5$ and $n_1 = 30, n_2 = 15$, the CT seems to have better performance than the WP. Although the means of expected losses of the WP are slightly smaller than those of the CT, the distributions of WP have long tails. Thus, one may prefer the CT to the WP in these cases. If the unbalance is not sever, say $n_1 = 30, n_2 = 25$, they behave similarly to the balanced case $n_1 = n_2 = 30$.

5. Conclusions

Based on the comparison of the distributions of the expected losses, I recommend the WP even when sample sizes are small provided that the sample sizes are near balanced. It seems that the CT interval is useful when sample sizes are small and the unbalance is severing. However, the choice is still depend on your attitude toward the confidence level. If you agree on the interpretation of the confidence level as "average performance" rather than "worst possible performance," the recommendation would be helpful for choosing a proper confidence interval.

Acknowledgements

The author would like to thank Professor Thomas Santner and Vivek Pradhan. They offered the C++ executable file for calculating the CT interval. It was helpful to implementing the CT method. Also their paper motivated the author to prepare this manuscript.

References

Agresti, A. and Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions, *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.

Agresti, A. and Min, Y. (2001). On small-sample confidence intervals for parameters in discrete distributions, *Biometrics,* **57**, 963–971.

Blyth, C. R. and Still, H. A. (1983). Binomial confidence intervals, *Journal of the American Statistical Association,* **78**, 108–116.

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

Casella, G. T., Hwang, T. G. and Robert C. P. (1994). Loss functions for set estimation, *Statistical Decision Theory and Related Topics V* (Edited by S. S. Gupta and J. O. Berger), Springer-Verlag.

Chan, I. S. F. and Zhang, Z. (1999). Test-based exact confidence intervals for the difference of two binomial proportions, *Biometrics,* **55**, 1202–1209.

Coe, P. R. and Tamhane, A. C. (1993). Small sample confidence intervals for the difference, ratio and odds ratio of two success probabilities, *Communications in Statistics Part B-Simulation and Computation*, **22**, 925–938.

Esty, W. E. and Banfield, J. D. (2003). The box-percentile plot, *Journal of Statistical Software,* **8**, Issue 17.

Lehmann, E. L. (1986) *Testing statistical hypotheses,* John Wiley & Sons, New York.

Lee, S.-C. (2006a). Interval estimation of binomial proportions based on weighted Polya posterior, *Computational Statistics & Data Analysis,* **51**, 1012–1021.

Lee, S.-C. (2006b). The weighted Polya posterior confidence interval for the difference between two independent proportions, *The Korean Journal of Applied Statistics,* **19**, 171–181.

Lee, S.-C. (2007). Confidence intervals for a linear function of binomial proportions based on a Bayesian approach, *The Korean Journal of Applied Statistics,* **20**, 257–266.

Meeden, G. D. (1999). Interval estimators for the population mean for skewed distributions with a small sample size, *Journal of Applied Statistics,* **26**, 81–96.

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

Newcombe, R. G. (1998), Interval estimation for the difference between independent porportions: Comparison of eleven methods, *Statistics in Mediciene,* **17**, 873–890.

Price, R. M. and Bonett, D. G. (2004). An improved confidence interval for a linear function of binomial proportions, *Computational Statistics & Data Analysis,* **45**, 449–456.

R Development Core Team (2008). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Santner, T. J., Pradhan, V., Senchaudhuri, P., Mehta, C. R. and Tamhane, A. C. (2007). Small-sample comparisons of confidence intervals for the difference of two independent binomial proportions, *Computational Statistics & Data Analysis*, **51**, 5791–5799.

Santner, T. J. and Snell, M. K. (1980). Small-sample confidence intervals for $p_1 - p_2$ and p_1/p_2 in 2×2 contingency tables, *Statistics in Medicine,* **17**, 873–890.

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

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference, *Journal of the American Statistical Association,* **22**, 209–212.

Received July 2008; Accepted September 2008