Exact Confidence Intervals for the Difference of Proportions

Jimmy A. Doi\(^1\) and Roger L. Berger\(^2\)
\(^1\)Calif. Polytechnic State Univ., San Luis Obispo and
\(^2\)NC State University, \(^3\)National Science Foundation

Department of Statistics, San Luis Obispo, CA 93407 (jdoi@calpoly.edu)

Key Words: Exact test, Two-by-two table, Unconditional distribution, Confidence interval

Abstract: Exact confidence intervals for the difference of two independent binomial proportions are popularly used in the analysis of data stemming from clinical trials. Exact methods are best suited for studies based on small sample sizes since large sample approximation methods can perform poorly. We apply the confidence region p-value method (Berger and Boos 1994) in proposing exact confidence intervals which have been found to perform better than the standard exact confidence intervals in many situations. The exact confidence intervals are based on the unconditional distribution of the binomial responses. We examine the two methods by providing coverage probability and expected length comparisons.

1 Introduction

Consider a clinical trial where the goal is to compare the efficacy of a new treatment (drug 1) versus that of a standard (drug 2). Let us denote \(\pi_1\) and \(\pi_2\) as the true response rates of the new and standard treatments respectively. Often, researchers compare these two rates through their difference which we will denote as \(\delta = \pi_1 - \pi_2\).

Suppose \(X\) and \(Y\) are independent binomial random variables. The sample size for \(X\) is \(n_1\) and its success probability is \(\pi_1\). The sample size for \(Y\) is \(n_2\) and its success probability is \(\pi_2\). Assume higher probabilities indicate stronger efficacy of the drug. Let us denote the binomial probability mass function of \(X\) by

\[
\text{bin}(x, n_1, \pi_1) = \binom{n_1}{x} \pi_1^x (1 - \pi_1)^{n_1 - x},
\]

where \(x = 0, 1, \ldots, n_1\). Denote \(\text{bin}(y, n_2, \pi_2)\) as the analogous representation of the probability mass function of \(Y\).

For clinical trials where a new treatment is compared to a standard control, researchers are often interested in testing whether the treatment is significantly better than, or superior to, the control. For superiority trials, we define the appropriate hypotheses to be:

\[
H_0 : \pi_1 - \pi_2 \leq \delta_0 \\
H_1 : \pi_1 - \pi_2 > \delta_0,
\]

where \(\delta_0 \geq 0\) is a clinically significant value determined by researchers.

Instead of establishing superiority, some clinical trials are designed to show that the efficacy of a new treatment is no worse than that of a control. For non-inferiority trials, the appropriate hypotheses are:

\[
H_0 : \pi_1 - \pi_2 \geq \delta_0 \\
H_1 : \pi_1 - \pi_2 < \delta_0
\]

where \(\delta_0 < 0\) is a clinically significant value determined by researchers.

In the context of hypothesis testing, the test statistic we will use in ordering the sample space is the so-called \(\delta\) projected \(Z\) statistic, as discussed by Chan (1999). Given \((x, y) \in \Omega = \{0, 1, \ldots, n_1\} \times \{0, 1, \ldots, n_2\}\) and \(\delta_0 \in (-1, 1)\), the \(\delta\) projected \(Z\) statistic is defined as

\[
Z(x, y; \delta_0) = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{\frac{\bar{x}(1 - \bar{x})}{n_1} + \frac{\bar{y}(1 - \bar{y})}{n_2}},
\]

where \(\bar{x}\) and \(\bar{y}\) are the restricted maximum likelihood estimates for \(\pi_1\) and \(\pi_2\) respectively. The restricted maximum likelihood estimation is based upon the restriction \(\delta_0 = \pi_1 - \pi_2\) and its specific form was shown by Miettinen and Nurminen (1985) and by Farrington and Manning (1990).

2 p-value and the Nuisance Parameter Problem

Assuming the difference \(\pi_1 - \pi_2\) is at the null boundary (i.e. \(\delta = \delta_0\)), the probability of observing a particular sample point \((X, Y) = (x, y)\) is given by

\[
f_{\pi_1, \delta_0}(x, y) = \text{bin}(x, n_1, \pi_1) \text{bin}(y, n_2, \pi_1 - \delta_0).
\]

The expression in (2.1) is our basis in defining a p-value and the presence of a nuisance parameter, \(\pi_1\), poses a problem.

We can address this problem in one of two ways. First, we can attempt a conditioning approach. In the context of testing for superiority or non-inferiority, when \(\delta_0 \neq 0\) we can apply a conditional approach by using Fisher’s Exact Test. However, in the general case when \(\delta_0 \neq 0\), a simple sufficient statistic has yet to be found. Since a non-trivial conditional approach is unavailable, an alternative
is to use an unconditional approach by employing what is known as the maximization method.

In the framework of hypothesis testing, let us assume that larger values of the chosen test statistic, say Z, give stronger evidence against the null hypothesis. As given by Casella and Berger (2002), in the presence of a generic nuisance parameter \( \theta \), we define a p-value, as

\[
p = \sup_{\theta \in \Theta_0} P_\theta(Z \geq z),
\]

(2.2)

where \( z \) is the observed value of the test statistic \( Z \) and \( \Theta_0 \) denotes the null space. For the hypotheses we will address, we define \( \Theta_0 = \{ (\pi_1, \pi_2) : \pi_1 - \pi_2 \leq \delta_0 \} \).

By using \( Z(x, y; \delta_0) \) as our test statistic of choice we see that, as in the case when testing for superiority and non-inferiority, larger values of the test statistic yield stronger evidence against the null hypothesis. Applying (2.2), an unconditional exact test can be based upon the following p-value:

\[
p_u(x, y) = \sup_{(\pi_1, \pi_2) \in \Theta^*_0} \left( \sum_{Z(a,b,\delta_0) \geq Z(x,y,\delta_0)} f_{\pi_1, \delta_0}(a,b) \right),
\]

(2.3)

where \((a,b) \in \Omega\). Note that the p-value is found by determining the supremum of the sum over the entire two dimensional null space, which can be an extremely time consuming and computationally intensive search. However, it can be shown that the supremum of the argument in (2.3) is achieved on the null boundary. That is, the supremum occurs over the set \( \Theta^*_0 = \{ (\pi_1, \pi_2) : \pi_1 - \pi_2 = \delta_0 \} \). Hence, we can simplify the definition of the p-value in (2.3) as

\[
p_u(x, y) = \sup_{(\pi_1, \pi_2) \in \Theta^*_0} \left( \sum_{Z(a,b,\delta_0) \geq Z(x,y,\delta_0)} f_{\pi_1, \delta_0}(a,b) \right).
\]

(2.4)

Given \( \pi_1 - \pi_2 = \delta_0 \), for a fixed value of \( \delta_0 \in (-1, 1) \), it is easy to show that the nuisance parameter \( \pi_1 \) is restricted to the interval \( I = [\max(0, \delta_0), \min(1, 1 + \delta_0)] \). Thus, an equivalent definition for the p-value is

\[
p_u(x, y) = \sup_{\pi_1 \in I} \left( \sum_{Z(a,b,\delta_0) \geq Z(x,y,\delta_0)} f_{\pi_1, \delta_0}(a,b) \right).
\]

(2.5)

\( p_u(x, y) \), so labeled since the maximization is performed in an unrestricted fashion over the entire nuisance parameter range, is the basis for the standard unconditional exact test. We will commonly refer to \( p_u(x, y) \) as simply \( p_u \). \( p_u \) is a valid p-value (i.e. for every \((\pi_1, \pi_2) \in \Theta_0\), \( P_{(\pi_1, \pi_2)}[p_u \leq \alpha] \leq \alpha \)). We will denote the \( \alpha \) level test determined by \( p_u \leq \alpha \) by \( T_{p_u} \). As a level \( \alpha \) test, \( T_{p_u} \) cannot be liberal however it opens the possibility of the test to be quite conservative. The p-value maximization algorithm has surely been simplified by reducing the search across one dimension instead of two. However, this method offers a conservative approach since it accounts for the ‘worst case scenario’ with respect to the nuisance parameter. Thus, the p-value \( p_u \) can, in many situations, be unnecessarily high.

### 3 Confidence Region p-value

Berger and Boos (1994) proposed a method that alleviates the conservativeness of the standard exact unconditional test based upon \( p_u \). As opposed to the unrestricted maximization over the entire nuisance parameter space, their method involves a restricted maximization which yields a less conservative p-value.

Again, let \( \theta \) denote the nuisance parameter of interest (possibly vector valued). Given data \( x \), suppose \( C_\beta(x) \) is a \((1 - \beta)\) confidence region for \( \theta \). Denote \( T(x) \) as the statistic used to order the sample space and assume large values of \( T \) lend stronger evidence against the null hypothesis of interest. Define \( p(\theta|x) = P(\mathcal{T}(X) \geq T(x)) \). The Berger and Boos confidence region p-value is given by

\[
p_r(x, y) = \sup_{\theta \in C_\beta(x)} p(\theta|x) + \beta.
\]

(3.1)

\( p_r(x, y) \), so labeled since the maximization is based on a restricted search of the nuisance parameter space, will be used to compare against \( p_u \). We will commonly refer to \( p_r(x, y) \) as simply \( p_r \). In defining \( p_r \), although the choice of \( \beta \) is left to the discretion of the researcher, if \( \beta \) is chosen to be too small (i.e. \( 1 - \beta \approx 1 \)), then the resulting ‘restricted’ search would nearly encompass the entire nuisance parameter space. Berger and Boos suggest to use values of \( \beta \) such as 0.001 and 0.0001. We chose \( \beta = 0.001 \) for all our computations involving \( p_r \). In their work, Berger and Boos showed that, as with \( p_u, p_r \) is also a valid p-value. We will denote the \( \alpha \) level test determined by \( p_r \leq \alpha \) by \( T_{p_r} \).

In our definition of \( p_r \), we will use the argument of the supremum in (2.5) to serve the role of \( p(\theta|x) \) as found in (3.1). To construct \( C_\beta(x) \) in (3.1), a \((1 - \beta)\) confidence region for \((\pi_1, \pi_2) \in \Theta_0 \) is generated by the cross product of two \((1 - \beta)^{1/2} \) Clopper Pearson confidence intervals (Clopper and Pearson, 1934), one for \( \pi_1 \) and the other for \( \pi_2 \). Given a particular observation \((X, Y) = (x, y)\), we will use \((l_1, u_1)\) and \((l_2, u_2)\) to denote the Clopper Pearson confidence intervals for \( \pi_1 \) and \( \pi_2 \) respectively. The \((1 - \beta)\) confidence region for \((\pi_1, \pi_2) \in \Theta_0 \) is \([l_1, u_1] \times [l_2, u_2] \cap \Theta_0 \).

Next, we will examine confidence intervals by inverting the confidence region p-value exact unconditional test \( T_{p_u} \) and compare its performance with that of the confidence intervals generated by inverting the standard exact unconditional test \( T_{p_u} \). Based on these two methods, we will provide coverage probability, average length, and expected length comparisons.
4 Comparison of the Exact Confidence Intervals

We have examined the performances of the exact methods under various settings of the sample sizes \( n_1 \) and \( n_2 \). Here, we examined a total of 15 sample size combinations of \( (n_1, n_2) \), where \( n_1 : n_2 \in \{1:1,2:1,3:1\} \) and \( n_1 + n_2 \in \{20, 40, 60, 80, 100\} \). In the interest of space, we include figures for only some of these sample size combinations, however we will discuss the overall performances of the exact methods across all sample size settings.

We will first examine coverage probability comparisons of the two methods which will be followed by comparing average length values and expected length plots.

4.1 Comparison of Coverage Probability

The coverage probability plots we will discuss are based on a fixed \( \pi_2 \) approach. To keep the number of coverage probability plots for a given \( (n_1, n_2) \) to a manageable size, we considered \( \pi_2 \in D \) where

\[
D = \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50\}
\]

Here, there is no need to consider any \( \pi_2 \) values beyond 0.50 due to symmetry.

As we will see, the coverage probability plots are extremely jagged. This is explained by the fact that, at a given value of \( \pi_2 \), as \( \pi_1 \) varies across \([0, 1]\), so does the value of \( \delta \). As \( \delta \) changes, the set \( A = \{(x, y) \in \Omega : \delta \in CI(x, y)\} \) may change as sample points are added or dropped depending on whether their corresponding confidence intervals capture or no longer capture \( \delta \). Obviously, such changes occur at the end points of confidence intervals. Because the coverage probability is defined as a summation over the set \( A \), as the cardinality of \( A \) changes, this leads to a spike or vertical jump in the plot which explains its overall jagged nature.

4.1.1 \( n_1 : n_2 = 1:1 \)

For \( n_1 : n_2 = 1:1 \), we find that the coverage probabilities based on \( p_r \) are generally not as conservative as the corresponding graphs based on \( p_u \). For most of the coverage plots in this sample size ratio, the dotted \( p_r \) line generally lies on or above the solid \( p_u \) line. It is worth noting for a given plot where \( p_r \) is more conservative, the degree to which the coverage probability plot based on \( p_r \) lies above that of \( p_u \) is not the case that \( p_r \) is uniformly less conservative than \( p_u \) for this sample size ratio, however, as the sum \( n_1 + n_2 \) increases, the number of cases where \( p_r \) is less conservative also increases. In the smallest sample size ratio, \( n_1 = 10, n_2 = 10 \) at \( \pi_2 = 0.25 \), the solid line is based on \( p_u \), the dotted line is based on \( p_r \).

4.1.2 \( n_1 : n_2 = 2:1 \)

For \( n_1 : n_2 = 2:1 \), we find that the coverage probabilities based on \( p_r \) are generally not as conservative as the corresponding graphs based on \( p_u \). For most of the coverage plots in this sample size ratio, the dotted \( p_r \) line generally lies on or above the solid \( p_u \) line. It is worth noting for a given plot where \( p_r \) is more conservative, the degree to which the coverage probability plot based on \( p_r \) lies above that of \( p_u \) is not the case that \( p_r \) is uniformly less conservative than \( p_u \) for this sample size ratio, however, as the sum \( n_1 + n_2 \) increases, the number of cases where \( p_r \) is less conservative also increases. In the smallest sample size ratio, \( n_1 = 66, n_2 = 34 \) at \( \pi_2 = 0.30 \), the solid line is based on \( p_u \), the dotted line is based on \( p_r \).
ple size setting \((n_1, n_2) = (13, 7)\), \(p_r\) coverage is found to be at least as conservative as \(p_u\). When the \(p_r\) coverage is more conservative in a given plot, this occurs across a relatively small range of \(\pi_1\) values. For the intermediate sample size settings, \(n_1 + n_2 \in \{40, 60, 80\}\), \(p_r\) is found to be less conservative for the vast majority of cases. In the largest sample size setting \((n_1, n_2) = (66, 34)\), virtually all plots show \(p_r\) coverage being less conservative. This can be seen in Figure 2 where \(\pi_2 = 0.30\).

Except for the smallest sample size setting \(n_1 = 13\) and \(n_2 = 7\), in most of the remaining cases \(p_r\) is noticeably less conservative than \(p_u\). For a given case where \(p_r\) is less conservative, we again note the degree to which the coverage probability plot based on \(p_r\) lies below that of \(p_u\). Among the various plots we have examined in this sample size ratio, the degree to which \(p_r\) is more conservative is overshadowed by the degree to which \(p_r\) is less conservative.

### 4.1.3 \(n_1 : n_2 = 3:1\)

For \(n_1 : n_2 = 3:1\), except for the small sample size setting \(n_1 = 15\) and \(n_2 = 5\), in most of the remaining cases \(p_r\) is less conservative than \(p_u\). In fact, the separation between the two lines is quite prominent as early as \(n_1 + n_2 = 40\). For \(n_1 + n_2 > 40\), \(p_r\) is found to be considerably less conservative than \(p_u\) as evidenced in Figure 3 by the increasing rift between the two coverage graphs. Among the various plots for the 3:1 setting, bearing in mind the relatively increasing degree to which the coverage probability plot based on \(p_r\) lies below that of \(p_u\), the degree to which \(p_r\) is more conservative is small compared to the degree to which \(p_r\) is less conservative.

### 4.1.4 All Sample Size Combinations

As an overall summary across all 15 combinations of \((n_1, n_2)\) we find that, for the most part, the degree to which \(p_r\) is more conservative than \(p_u\) is relatively small compared to the degree to which \(p_r\) is less conservative than \(p_u\). Among the plots within each of the 15 combinations of \((n_1, n_2)\), it is not often the case where \(p_u\) is less conservative than \(p_r\). When \(p_r\) is more conservative, the difference between the coverage plots is small. However, when \(p_r\) is less conservative, this occurs for a majority of cases and the gap between the solid and dotted lines is noticeably large across a significant range of the plots. The \(p_r\) coverage becomes increasingly less conservative as \(n_1 + n_2\) grows and as \(n_1 : n_2\) becomes more and more unbalanced.

### 4.2 Comparison of Average Length

We begin our length comparisons by considering average interval lengths. Again, we used sample size settings \((n_1, n_2)\) where \(n_1 : n_2 \in \{1:1, 2:1, 3:1\}\) and \(n_1 + n_2 \in \{20, 40, 60, 80, 100\}\). The results are in Table 1.

![Cov. Prob.](image)

Figure 3: Coverage probability plot for \(n_1 = 75, n_2 = 25\) at \(\pi_2 = 0.30\). The solid line is based on \(p_u\), the dotted line is based on \(p_r\).

Given a particular \(n_1\) and \(n_2\), the sample space \(\Omega\) has cardinality \(N = (n_1 + 1)(n_2 + 1)\). Given \((x, y) \in \Omega\), we denote \(CI_{p_r}(x, y)\) as the corresponding confidence interval generated by \(p_r\) and \(CI_{p_u}(x, y)\) as the corresponding confidence interval generated by \(p_u\). For each \((x, y) \in \Omega\), we partition the sample space by placing each sample point in one of two mutually disjoint classes depending on whether or not the length of \(CI_{p_u}(x, y)\) is less than the length of \(CI_{p_r}(x, y)\). Using \(\|CI(x, y)\|\) to denote the length of the confidence interval \(CI(x, y)\), we will denote \(\|CI_{p_r}\| < \|CI_{p_u}\|\) as the category where the length of \(CI_{p_r}\) for a given \((x, y)\) is less than the corresponding length of \(CI_{p_u}\). Also, we will denote \(\|CI_{p_r}\| \not< \|CI_{p_u}\|\) as the category where the length of \(CI_{p_r}\) for a given \((x, y)\) is not less than the corresponding length of \(CI_{p_u}\). For notational convenience to be used in the table, we will associate the category “\(\|CI_{p_r}\| \not< \|CI_{p_u}\|\)” with

\[
\{(x, y) : (x, y) \in \Omega, \quad \|CI_{p_r}(x, y)\| \not< \|CI_{p_u}(x, y)\|\}
\]

and associate the category “\(\|CI_{p_r}\| \not< \|CI_{p_u}\|\)” with

\[
\{(x, y) : (x, y) \in \Omega, \quad \|CI_{p_r}(x, y)\| < \|CI_{p_u}(x, y)\|\}.
\]
For a fixed $n_1$ and $n_2$, given the sample points captured in a particular category $\mathcal{G}$, we define the average absolute difference of interval lengths ($AADIL$) as

$$AADIL = \frac{1}{n_y} \sum_{(x,y) \in \mathcal{G}} ||CI_{p_1}(x,y)|| - ||CI_{p_2}(x,y)||$$

where $\mathcal{G}$ is the corresponding set in (4.1) or (4.2) and $n_y$ is the number of sample points captured in category $\mathcal{G}$. Based on the categorization of the table rows, $AADIL$ gives the average difference when the $CI_{p_1}$ length is shorter than the $CI_{p_2}$ length and when it is not. Although the $CI_{p_1}$ interval lengths may not be uniformly shorter than corresponding $CI_{p_2}$ lengths, we can use $AADIL$ as a means to determine whether, on average, the $CI_{p_1}$ lengths are significantly shorter as compared to when they are not.

### 4.2.1 Results

For $n_1 : n_2 = 1:1$, we find that for the majority of cases, the $CI_{p_1}$ lengths are not less than corresponding $CI_{p_2}$ lengths as indicated by the proportion $n/N$. Note, however, that the proportion of cases when $CI_{p_1}$ lengths are shorter as $n_1 + n_2$ increases. Comparing the $AADIL$ values in a given column, we find that, except for where $(n_1, n_2) = (10, 10)$, the value corresponding to $||CI_{p_1}|| < ||CI_{p_2}||$ is greater than that corresponding to $||CI_{p_1}|| \neq ||CI_{p_2}||$. That is, when the $CI_{p_1}$ lengths are shorter, they are shorter (on average) to a greater degree than when they are not. Notice that the disparity between the $AADIL$ values increases as $n_1 + n_2$ increases. In the $(n_1, n_2) = (20, 20)$ case, the $AADIL$ values are relatively similar however in the $(n_1, n_2) = (50, 50)$ case the $AADIL$ for the $||CI_{p_1}|| < ||CI_{p_2}||$ category is twice that of the $||CI_{p_1}|| \neq ||CI_{p_2}||$ case. The advantage of larger $AADIL$ values for the $||CI_{p_1}|| < ||CI_{p_2}||$ category is, admittedly, less impressive when considering the fact that, for each $(n_1, n_2)$ setting in this table, only a minority of sample points are captured in $||CI_{p_1}|| < ||CI_{p_2}||$.

For $n_1 : n_2 = 2:1$, when $(n_1, n_2) = (13, 7)$ over 80% of sample points are captured in $||CI_{p_1}|| \neq ||CI_{p_2}||$ and the corresponding $AADIL$ values are relatively similar. However, for $(n_1, n_2) = (27, 13)$, although a majority of points are again captured in $||CI_{p_1}|| \neq ||CI_{p_2}||$, over 40% are in $||CI_{p_1}|| < ||CI_{p_2}||$. Here, the $AADIL$ for the $||CI_{p_1}|| < ||CI_{p_2}||$ category is about 2.5 times that of the $||CI_{p_1}|| \neq ||CI_{p_2}||$ case. In the remaining cases of $(n_1, n_2)$, the benefit of using confidence intervals based upon $p_1$ becomes more apparent. For these cases, the majority of sample points (up to 65%) are captured by $||CI_{p_1}|| < ||CI_{p_2}||$. The difference in $AADIL$ values for these cases is noteworthy as well. The ratio of the $AADIL$ for the $||CI_{p_1}|| < ||CI_{p_2}||$ category to that of the $||CI_{p_1}|| \neq ||CI_{p_2}||$ case for $(n_1, n_2) = (27, 13), (53, 27), (66, 34)$ are 4.67, 3.14, and 3.09 respectively. Thus, for these cases, the majority of $CI_{p_1}$ lengths are shorter and by a significantly greater degree as compared to when they are not shorter.

For $n_1 : n_2 = 3:1$, we find the trend continues in favor of the $p_1$ based confidence intervals. Except for the $(n_1, n_2) = (15, 5)$ and (30, 10) cases, a majority of sample points are captured in $||CI_{p_1}|| < ||CI_{p_2}||$. It is worth noting, however, that over 38% of points in the $(n_1, n_2) = (30, 10)$ case are in the $||CI_{p_1}|| < ||CI_{p_2}||$ category and the $AADIL$ values are different by over a factor of 4, in favor of the $p_1$ based intervals. For the remaining $(n_1, n_2)$ cases, the majority of sample points (up to ~75%) are captured by $||CI_{p_1}|| < ||CI_{p_2}||$. Again, it is important to note the difference in magnitudes of the $AADIL$ values. The ratio of the $AADIL$ for the $||CI_{p_1}|| < ||CI_{p_2}||$ category to that of the $||CI_{p_1}|| \neq ||CI_{p_2}||$ case for $(n_1, n_2) = (45, 15), (60, 20),$ and (75, 25) are 4.86, 5.88, and 4.21 respectively. So we see again that, for these cases, the majority of $CI_{p_1}$ lengths are shorter and by a significantly greater degree as compared to when they are not shorter.

### 4.3 Comparison of Relative Difference of Expected Length

The expected length plots we will discuss are based on a fixed $\pi_2$ approach. We again considered $\pi_2$ restricted to the set $D$ which was defined previously under the coverage probability discussion. Although the direct comparison of expected lengths is useful in determining how often (and to what extent) the $CI_{p_1}$ lengths are comparatively shorter, we would like to be able to quantify the difference between the $CI_{p_1}$ and $CI_{p_2}$ expected lengths in relation to their overall magnitudes. Such a quantification can be important since a given expected length difference can be more meaningful for smaller intervals as opposed to larger intervals.

For notational convenience, let us define

$$\epsilon_{p_1}(\pi_1, \pi_2) = \sum_{(x,y) \in \mathcal{G}} ||CI_{p_1}(x,y)|| \text{bin}(x, n_1, \pi_1) \text{bin}(y, n_2, \pi_2)$$

and

$$\epsilon_{p_2}(\pi_1, \pi_2) = \sum_{(x,y) \in \mathcal{G}} ||CI_{p_2}(x,y)|| \text{bin}(x, n_1, \pi_1) \text{bin}(y, n_2, \pi_2)$$

where $||CI(x, y)||$ denotes the length of the confidence interval $CI(x, y)$.

We define the relative difference of expected lengths ($RDEL$) as:

$$RDEL = \frac{\epsilon_{p_1} - \epsilon_{p_2}}{\frac{1}{2}(\epsilon_{p_1} + \epsilon_{p_2})}$$
Based on its definition, note that $RDEL$ is positive when the $p_e$ expected length is shorter, negative when the $p_u$ expected length is longer. We will evaluate the $RDEL$ for various sample size settings.

### 4.3.1 $n_1 : n_2 = 1:1$

For $n_1 : n_2 = 1:1$, the relative difference is negative in the majority of plots. In other words, the $p_e$ expected length is greater than the $p_u$ expected length in a majority of cases. In Figure 4, where $(n_1, n_2) = (10, 10)$ and $\pi_2 = 0.20$, we see that the relative difference is strictly negative exhibiting a typical value between -0.007 and -0.008. As $n_1$ and $n_2$ increase, we find that more and more plots yield relative differences that are positive. However, for such plots, the range of $\pi_1$ values for which the relative difference is positive is often not as large as the range of $\pi_1$ values where the relative difference is negative. Another important note is that, in most plots where the relative difference is positive, the maximum value of $RDEL$ is, at most, on the same order of magnitude as the minimum value of $RDEL$. That is, it is often the case that $|\max(RDEL)| \leq |\min(RDEL)|$ for a given plot. Such observations suggest that the $p_e$ confidence intervals do not perform noticeably better than the corresponding $p_u$ confidence intervals with respect to expected length when $n_1 : n_2 = 1:1$.

![Figure 4: Relative difference of expected lengths ($RDEL$) for $n_1 = 10$, $n_2 = 10$ where $\pi_2 = 0.20$.](image)

### 4.3.2 $n_1 : n_2 = 2:1$

For $n_1 : n_2 = 2:1$, there is a marked improvement in favor of the $p_e$ expected lengths. Except for the smallest sample size setting of $(n_1, n_2) = (13, 7)$, we note two major differences from what we observed in the 1:1 ratio case. First, although the relative difference is not uniformly positive across all $\pi_1$, it is positive over a wider region of the range of $\pi_1$. This is evident especially in the cases where $n_1 + n_2 \geq 60$. Second, in most plots where the relative difference is positive, the maximum value of $RDEL$ noticeably exceeds the magnitude of the minimum value of $RDEL$. That is, $|\max(RDEL)| > |\min(RDEL)|$ for a majority of plots. Note, for example, Figure 5 where $(n_1, n_2) = (66, 34)$ and $\pi_2 = 0.20$.

![Figure 5: Relative difference of expected lengths ($RDEL$) for $n_1 = 66$, $n_2 = 34$ where $\pi_2 = 0.20$.](image)

### 4.3.3 $n_1 : n_2 = 3:1$

Finally, for $n_1 : n_2 = 3:1$, we find the performance of the $p_e$ expected lengths improves even further. Again, aside from the smallest sample size setting of $(n_1, n_2) = (15, 5)$, the two important differences mentioned above in the 2:1 case are further pronounced here. In a majority of the plots, the relative difference is almost uniformly positive across all $\pi_1$. This is especially evident in the cases where $n_1 + n_2 \geq 60$. Also, in a majority of the plots, the maximum value of $RDEL$ significantly exceeds the magnitude of the minimum value of $RDEL$. That is, $|\max(RDEL)| > |\min(RDEL)|$ for almost all plots. Consider, for example, Figure 6 where $(n_1, n_2) = (75, 25)$ and $\pi_2 = 0.15$.

![Figure 6:](image)

### 4.3.4 All Sample Size Combinations

As an overall summary across all 15 combinations of $(n_1, n_2)$, in general the relative difference of expected lengths are positive and significantly large for unbalanced
sample size ratios. For the balanced sample size ratio cases, the results are mixed since the relative difference either fluctuated between positive and negative values or was primarily negative. However, among all 15 combinations of \((n_1, n_2)\), when the relative difference assumed negative values, the magnitude of these values were typically quite small. On the other hand, when the relative difference assumed positive values, the magnitude of such values were comparatively large. This was especially evident among the unbalanced sample size ratios for \(n_1 + n_2 \geq 60\).

5 Conclusion

When comparing the coverage probability properties, we see that the confidence intervals based on the restricted exact unconditional p-value is often less conservative than the confidence intervals based on the unrestricted exact unconditional p-value. Although \(p_r\) was not found to be uniformly less conservative than \(p_u\), we found that, especially for the unbalanced sample size settings, \(p_r\) was found to be less conservative to a significantly greater degree as compared to when \(p_r\) was found to be more conservative.

In terms of average length comparisons, the benefits of the restricted exact unconditional method were found especially in the non-1:1 ratios of \(n_1\) to \(n_2\). In the 1:1 cases, the majority of \(CI_{p_r}\) lengths were not shorter than corresponding \(CI_{p_u}\) lengths, however except for \((n_1, n_2) = (10, 10)\), the \(AADIL\) values for \(||CI_{p_r}|| < ||CI_{p_u}||\) were larger than that for \(||CI_{p_r}|| \geq ||CI_{p_u}||\). The \(p_r\) based confidence intervals performed progressively better in the 2:1 and 3:1 ratios where the difference in \(AADIL\) values were noticeably different in many cases.

Finally, with respect to expected length comparisons we find that, in general, the \(p_r\) expected lengths are shorter than \(p_u\) expected lengths in the case of unbalanced sample size ratios. As the samples sizes \(n_1\) and \(n_2\) increase, so does the benefit of \(p_r\) based confidence intervals as the difference, and relative difference, between the two expected lengths grows. For a large number of \(RDEL\) plots in the non-1:1 cases we found that \(RDEL\) was positive for a relatively large portion of the range of \(\pi_1\) and, in those instances where \(RDEL\) was negative, the extent to which \(RDEL\) fell below zero was quite small relative to the extent to which \(RDEL\) rose above zero.

Overall, the restricted exact unconditional p-value confidence interval method does not offer significant benefits in the balanced sample size cases. However, the restricted method offers noticeable benefits in the unbalanced sample size cases. Although its performance was not uniformly better than the corresponding performance of the competing exact unconditional method, the gains of using the \(p_r\) confidence intervals strongly outweigh the losses.

References


Table 1: Average length comparisons for CI based on $p_r$ ($CI_{p_r}$) versus the CI based on $p_u$ ($CI_{p_u}$) where $n_1 : n_2 = 1:1$. $N = (n_1 + 1)(n_2 + 1)$ is the total number of confidence intervals possible given $n_1$ and $n_2$. $\parallel CI_{p_r} \parallel \not< \parallel CI_{p_u} \parallel$ indicates the category where the length of $CI_{p_r}$ for a given $(x, y)$ is not less than the corresponding length of $CI_{p_u}$. $\parallel CI_{p_r} \parallel < \parallel CI_{p_u} \parallel$ indicates the category where the length of $CI_{p_r}$ for a given $(x, y)$ is less than the corresponding length of $CI_{p_u}$.

$n$ accounts for the number of confidence intervals that are found under a particular category.

AADIL is the average of the absolute value of the difference of $CI_{p_r}$ and $CI_{p_u}$ interval lengths for all sample points in a particular category.

<table>
<thead>
<tr>
<th>$n_1 : n_2 = 1:1$</th>
<th>$n_1 : n_2 = 2:1$</th>
<th>$n_1 : n_2 = 3:1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\parallel CI_{p_r} \parallel \not&lt; \parallel CI_{p_u} \parallel$</td>
<td>$\parallel CI_{p_r} \parallel \not&lt; \parallel CI_{p_u} \parallel$</td>
<td>$\parallel CI_{p_r} \parallel \not&lt; \parallel CI_{p_u} \parallel$</td>
</tr>
<tr>
<td>$n (n/N)$</td>
<td>$n (n/N)$</td>
<td>$n (n/N)$</td>
</tr>
<tr>
<td>$AADIL$</td>
<td>$AADIL$</td>
<td>$AADIL$</td>
</tr>
<tr>
<td>(10,10)</td>
<td>(13,7)</td>
<td>(15,5)</td>
</tr>
<tr>
<td>$N = 121$</td>
<td>$N = 112$</td>
<td>$N = 96$</td>
</tr>
<tr>
<td>121 (100%)</td>
<td>94 (83.9%)</td>
<td>90 (93.8%)</td>
</tr>
<tr>
<td>0.006</td>
<td>0.008</td>
<td>0.006</td>
</tr>
<tr>
<td>(20,20)</td>
<td>(27,13)</td>
<td>(30,10)</td>
</tr>
<tr>
<td>$N = 441$</td>
<td>$N = 392$</td>
<td>$N = 341$</td>
</tr>
<tr>
<td>401 (90.9%)</td>
<td>228 (58.2%)</td>
<td>209 (61.3%)</td>
</tr>
<tr>
<td>0.004</td>
<td>0.006</td>
<td>0.006</td>
</tr>
<tr>
<td>(30,30)</td>
<td>(40,20)</td>
<td>(45,15)</td>
</tr>
<tr>
<td>$N = 961$</td>
<td>$N = 861$</td>
<td>$N = 736$</td>
</tr>
<tr>
<td>777 (80.9%)</td>
<td>397 (46.1%)</td>
<td>300 (40.8%)</td>
</tr>
<tr>
<td>0.004</td>
<td>0.003</td>
<td>0.005</td>
</tr>
<tr>
<td>(40,40)</td>
<td>(53,27)</td>
<td>(60,20)</td>
</tr>
<tr>
<td>$N = 1681$</td>
<td>$N = 1512$</td>
<td>$N = 1281$</td>
</tr>
<tr>
<td>1257 (74.8%)</td>
<td>596 (39.4%)</td>
<td>375 (29.3%)</td>
</tr>
<tr>
<td>0.003</td>
<td>0.003</td>
<td>0.005</td>
</tr>
<tr>
<td>(50,50)</td>
<td>(66,34)</td>
<td>(75,25)</td>
</tr>
<tr>
<td>$N = 2601$</td>
<td>$N = 2345$</td>
<td>$N = 1976$</td>
</tr>
<tr>
<td>1699 (65.3%)</td>
<td>815 (34.8%)</td>
<td>500 (25.3%)</td>
</tr>
<tr>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>$\parallel CI_{p_r} \parallel &lt; \parallel CI_{p_u} \parallel$</td>
<td>$\parallel CI_{p_r} \parallel &lt; \parallel CI_{p_u} \parallel$</td>
<td>$\parallel CI_{p_r} \parallel &lt; \parallel CI_{p_u} \parallel$</td>
</tr>
<tr>
<td>$n (n/N)$</td>
<td>$n (n/N)$</td>
<td>$n (n/N)$</td>
</tr>
<tr>
<td>$AADIL$</td>
<td>$AADIL$</td>
<td>$AADIL$</td>
</tr>
<tr>
<td>(10,10)</td>
<td>(13,7)</td>
<td>(15,5)</td>
</tr>
<tr>
<td>$N = 121$</td>
<td>$N = 112$</td>
<td>$N = 96$</td>
</tr>
<tr>
<td>0 (%)</td>
<td>18 (16.1%)</td>
<td>6 (6.2%)</td>
</tr>
<tr>
<td>0.005</td>
<td>0.016</td>
<td>0.011</td>
</tr>
<tr>
<td>(20,20)</td>
<td>(27,13)</td>
<td>(30,10)</td>
</tr>
<tr>
<td>$N = 441$</td>
<td>$N = 392$</td>
<td>$N = 341$</td>
</tr>
<tr>
<td>184 (19.1%)</td>
<td>164 (41.8%)</td>
<td>132 (38.7%)</td>
</tr>
<tr>
<td>0.004</td>
<td>0.016</td>
<td>0.024</td>
</tr>
<tr>
<td>(30,30)</td>
<td>(40,20)</td>
<td>(45,15)</td>
</tr>
<tr>
<td>$N = 961$</td>
<td>$N = 861$</td>
<td>$N = 736$</td>
</tr>
<tr>
<td>424 (25.2%)</td>
<td>464 (53.9%)</td>
<td>436 (59.2%)</td>
</tr>
<tr>
<td>0.004</td>
<td>0.016</td>
<td>0.023</td>
</tr>
<tr>
<td>(40,40)</td>
<td>(53,27)</td>
<td>(60,20)</td>
</tr>
<tr>
<td>$N = 1681$</td>
<td>$N = 1512$</td>
<td>$N = 1281$</td>
</tr>
<tr>
<td>902 (34.7%)</td>
<td>916 (60.6%)</td>
<td>906 (70.7%)</td>
</tr>
<tr>
<td>0.006</td>
<td>0.010</td>
<td>0.018</td>
</tr>
<tr>
<td>(50,50)</td>
<td>(66,34)</td>
<td>(75,25)</td>
</tr>
<tr>
<td>$N = 2601$</td>
<td>$N = 2345$</td>
<td>$N = 1976$</td>
</tr>
<tr>
<td>1530 (65.2%)</td>
<td>1530 (65.2%)</td>
<td>1476 (74.7%)</td>
</tr>
<tr>
<td>0.006</td>
<td>0.008</td>
<td>0.012</td>
</tr>
</tbody>
</table>