This is to certify that the

dissertation entitled

Interval Estimation for the Difference of Two Binomial Proportions in Non-adaptive and Adaptive Designs

presented by

Yichuan Xia

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Statistics

Connie Page
Co- Major professor
Connie Page

Date July 12, 2002

Co-Major professor
Vincent Melfi

Interval Estimation for the Difference of Two Binomial Proportions in Non-adaptive and Adaptive Designs

By

Yichuan Xia

A DISSERTATION

Submitted to

Michigan State University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

Department of Statistics and Probability

2002

ABSTRACT

Interval Estimation for the Difference of Two Binomial Proportions in Non-adaptive and Adaptive Designs

By

Yichuan Xia

When comparing two treatments with dichotomous responses, the difference in proportions of successful responses of the two groups is often of primary interest. Confidence intervals are typically provided to estimate the treatment difference. This interval estimation problem for both non-adaptive and adaptive designs is studied in the dissertation.

Several methods of constructing confidence intervals for the difference of two proportions are evaluated in non-adaptive designs. We begin by exploring the poor performance of the most widely used confidence interval, the Wald interval. We show that the poor behavior mainly results from its inappropriate center: the coverage performance can be improved greatly by simply recentering the Wald interval. We then derive a formula which gives smooth approximation of the coverage probability of the Wald interval. Regardless of oscillation, this approximation shows how much the coverage probability of the Wald interval falls below the nominal level. Our analysis demonstrates the Wald interval is rather anti-conservative and often behaves much worse than people's expectation. As alternatives, the Wald interval with continuity correction, two confidence intervals with adjusted centers(a Bayesian interval derived from Beta priors and Agresti-Coull's adding 2 successes and 2 failures interval) and the profile likelihood based confidence interval are evaluated. We compare both their coverage performance and expected lengths with those of the standard Wald interval. To replace the Wald interval, intervals with adjusted centers are recommended. Adaptive designs are gaining more attention nowadays. For adaptive designs, the validity of constructing confidence intervals discussed in non-adaptive designs is verified. We evaluate the performance of those confidence intervals in two general categories of adaptive designs: allocation adaptive designs and response adaptive designs. We develop theorems concerning the connections between the coverage performance and expected lengths of confidence intervals based on non-adaptive and allocation adaptive designs. The theorems suggest that the Wald interval does not behave satisfactorily and that the intervals with adjusted centers should be used in allocation adaptive designs. Extensive simulation supports the same conclusion in response adaptive designs.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my dissertation advisors, Professor Vincent Melfi and Professor Connie Page, for your constant guidance, generous support and extreme patience during the writing of this dissertation. Your dedication and contribution to statistics have been my main source of inspiration and encouragement during my research. I am extremely grateful you suggested that I work in adaptive designs. I sincerely appreciate the time you put into our weekly meetings and the help you offered whenever I needed it. Your willingness to help at anytime encouraged me to keep going no matter what difficulties I met during the research of this dissertation. I also thank you for all the time you spent on proofreading my dissertation drafts. I understand that it was not comfortable to read and correct a non-native speaker's tedious and stiff writing on statistics. If time could flow backwards, I would not hesitate even a second to ask you to be my thesis advisors again.

I would also like to thank Professor Roy Erickson and Professor Habib Salehi for serving on my thesis committee. Your help is highly appreciated.

My special thanks go to Professor Roy Erickson and Professor Hira Koul who taught me Theory of Probability and Theory of Statistics during my first year of graduate study. The two courses turns out to be very important in my research. I also thank Professor Connie Page for training me to be a (good) statistical consultant which helped me a lot on my job searching and will benefit my future career. I also owe Professor Vincent Melfi many thanks for teaching me two courses: Modern statistics and Sequential analysis and guiding me on the usage of *Latex* and *C/C++* when I was in trouble with them.

I will not forget to express my sincere gratitude to Professor Stapleton. Thank you for being a true friend for so many years. Those scenarios when I got along with you are so lovely. I will not forget your warm but firm encouragement when I faced some trouble just before and after I came to the department; your suggestions on how to be a (good) teaching assistant; your correction on my English errors in emails; the jokes we played back and forth......

I also thank the department for offering me the assistantship for four years.

At last, but not the least, I want to thank all the professors and friends who ever helped me during my stay at Michigan State University. Life here became easy, meaningful and colorful because of you.

List of Figures

TABLE OF CONTENTS

Chapter 1

Literature Review

1.1 Introduction

In clinical trials and in industrial work, to compare a new treatment with a standard (control) treatment, the difference in probabilities of successful responses of the two groups is often of primary interest. Confidence intervals are typically provided to estimate the treatment difference. There exist quite a lot of methods for constructing confidence intervals for the difference of the two success probabilities.

The most widely used confidence interval, the Wald interval, which is an asymptotic confidence interval computed based on a normal approximation, does not behave satisfactorily. In this dissertation, the poor performance of the Wald interval and the reason for the poor coverage performance are explored in Chapter 2. In Chapter 3, some selected methods for constructing confidence intervals for the difference of two treatment proportions are evaluated and compared with the Wald interval. We restrict attention to non-adaptive designs in these two chapters.

Nowadays, adaptive designs, in which the allocation of next subject to a certain treatment depends on accumulating information, is more widely used. The interval estimation problem is studied for adaptive designs in Chapter 4.

In this dissertation, we use three types of "coverage" probabilities: exact, approximate and nominal coverage probabilities. The exact coverage probability of a confidence interval is the actual coverage probability of that interval. The approximate coverage probability is an approximation of the coverage probability. We will be using an Edgeworth expansion to derive the approximate coverage probability of the Wald interval. The nominal coverage probability is its named confidence level. For example, a 95% confidence interval has nominal coverage probability 0.95 though its exact coverage probability might be different from the claimed level. Sometimes we don't specify whether a coverage probability is exact, approximate or nominal if it is obvious in context.

Before presenting our findings, it is useful to give a survey of related literature.

1.2 Some Confidence Intervals and Comparisons

Though we are interested in confidence intervals for the difference of two proportions, it is worthwhile to mention two papers on confidence intervals for one proportion which have impacted our study.

Let us begin by introducing the Wald intervals for one proportion and for the difference of two proportions.

Let X denote the number of successes from n Bernoulli trials with success probability p and let $\hat{p}$ denote the sample proportion. For two independent treatments, let X_1, X_2 denote the numbers of successes from treatment 1 and treatment 2 respectively, so that $X_i \sim Bin(n_i, p_i)$ for $i = 1, 2$. Let z_α represent $1-\alpha$ percentile of the standard normal distribution.

1. The $100(1-\alpha)\%$ Wald confidence interval for p is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

2. The $100(1-\alpha)\%$ Wald confidence interval for $p_1 - p_2$ is

$$\hat{p_1} - \hat{p_2} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

One way to derive these confidence intervals is to invert large sample Wald tests, which evaluate standard errors at the maximum likelihood estimates. For instance, the interval for p is the set of p_0 values having P-value exceeding α in testing

$$H_0 : p = p_0 \text{ versus } H_a : p \neq p_0$$

using the approximately normal Wald test statistic. The Wald intervals are sometimes called the standard intervals. Although these two intervals are simple and applied most often, a considerable literature shows that they behave poorly.

Brown *et al.* (2002) consider confidence intervals for one proportion. They notice there is a widespread misconception that the problems of the Wald interval are serious only when p is close to either boundary, or when the sample size n is rather small. Brown *et al.* (2002) shows that virtually all of the conventional wisdom and popular prescriptions are misplaced because the Wald interval has a pronounced systematic bias due to its inappropriate center. They derive two-term Edgeworth expansions as an analytical tool to compare and rank the some selected intervals with regard to their coverage probabilities. They also give the two-term expansions for the expected lengths of the Wald interval and some alternative intervals.

When deriving the two-term Edgeworth expansions for the coverage probabilities of those intervals for p, Brown *et al.* (2002) express all the confidence intervals in a unified form:

$$\{p : l_* \leq \frac{n^{1/2}(\hat{p}-p)}{\sqrt{p(1-p)}} \leq u_*\},$$

where l_* and u_* are not related to the sample proportion $\hat{p}$. Since the statistic $n^{1/2}(\hat{p}-p)/\sqrt{p(1-p)}$ has lattice structure, a direct application of a Theorem of Bhattacharya and Ranga Rao (1976) gives the desired Edgeworth expansions. But this method does not apply in two treatment problems.

Brown *et al.* (2002) show that the Wilson confidence interval for p, due to Wilson (1927), behaves much better than the standard interval. The Wilson interval is based on inverting the test with standard error evaluated at the null hypothesis value, which is the score test approach. Given level of significance α, this interval contains all p_0 values for which

$$\frac{n^{1/2}|\hat{p}-p_0|}{\sqrt{p_0(1-p_0)}} < z_{\alpha/2}$$

and has the form

$$\frac{X + z_{\alpha/2}^2/2}{n + z_{\alpha/2}^2} \pm \frac{n^{1/2} z_{\alpha/2}}{n + z_{\alpha/2}^2}\sqrt{\hat{p}(1-\hat{p}) + \frac{z_{\alpha/2}^2}{4n}}. \tag{1.2.1}$$

This interval turns out to behave better than the Wald interval for p.

Some confidence intervals for $p1 - p2$ are motivated by the Wilson interval for p.

Agresti and Coull (1998) noticed that (1.2.1) can be rewritten in the following way:

$$\hat{p}\frac{n}{n+z_{\alpha/2}^2}+\frac{1}{2}\left(\frac{z_{\alpha/2}^2}{n+z_{\alpha/2}^2}\right)$$
$$\pm z_{\alpha/2}\sqrt{\frac{1}{n+z_{\alpha/2}^2}\left[\hat{p}(1-\hat{p})\left(\frac{n}{n+z_{\alpha/2}^2}\right)+\frac{1}{4}\left(\frac{z_{\alpha/2}^2}{n+z_{\alpha/2}^2}\right)\right]}.$$

Hence, the midpoint of the Wilson interval is a weighted average of $\hat{p}$ and $1/2$, and it equals the sample proportion after adding $z_{\alpha/2}^2$ pseudo observations, half of each type. The square of the coefficient of $z_{\alpha/2}$ in this formula, is a weighted average of the variance of a sample proportion when $p = \hat{p}$ and the variance of a sample proportion when $p = 1/2$, using $n + z_{\alpha/2}^2$ in place of the usual sample size n. Motivated by this decomposition of the Wilson interval, Agresti and Caffo (2000) proposed adding 4 pseudo observations, one success and one failure from each treatment, to get the confidence interval for $p_1 - p_2$,

Also motivated by the Wilson confidence interval for p, Newcombe (1998) proposed a method that performs substantially better than the Wald interval. This confidence interval results from the single-sample score intervals for p_1 and p_2. Specifically, let $l_i < u_i$ be the roots for p_i in

$$\frac{\hat{p}_i - p_i}{\sqrt{\frac{p_i(1-p_i)}{n_i}}} = z_{\alpha/2}$$

for $i = 1, 2$. Newcombe's hybrid score $100(1-\alpha)\%$ interval is defined as

$$\left((\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{l_1(1-l_1)}{n_1} + \frac{u_2(1-u_2)}{n_2}}, (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{u_1(1-u_1)}{n_1} + \frac{l_2(1-l_2)}{n_2}}\right).$$

Unlike quite a lot of other confidence intervals, the Newcombe's interval is not symmetric around $\hat{p}_1 - \hat{p}_2$. It has margins of errors different from those of the Wald interval.

Newcombe (1998) evaluate eleven methods of constructing confidence intervals for $p_1 - p_2$ through simulation. Some of those confidence intervals have relatively complicated expressions compared to intervals discussed so far such as the Wald interval and the Agresti-Coull interval. Newcombe (1998) suggests replacing the Wald interval with the Newcombe hybrid score interval.The profile likelihood method (introduced in detail later), involving

$$\{\Delta \in (-1, 1) : 2(l(\hat{\Delta}, \hat{p}_2) - \tilde{l}(\Delta)) \le \chi_1^2(\alpha)\},$$

where $\Delta = p_1 - p_2$, $\hat{\Delta} = \hat{p}_1 - \hat{p}_2$ and l denotes the log-likelihood function of (Δ, p_2), $\tilde{l}(\Delta) = \max_{p_2} l(\Delta, p_2)$, is also considered in his paper. Newcombe (1998) shows this interval has the best coverage and location properties among all the eleven confidence intervals while it displays an undesirable anomaly. Suppose X_1, n_1 and $\hat{p}_2$ are held constants, while $n_2 \to \infty$ through values which keep X_2 integer valued. One expects that a good method would produce a sequence of intervals, each nested within its predecessor, tending asymptotically towards some corresponding interval for the single proportion, shifted by the constant p_2. Yet the profile likelihood method gives a sequence of lower limits which increase up to a certain n_2, but subsequently decrease, violating the above consideration.

Agresti and Caffo (2000) evaluate the Wald interval, the Agresti-Coull interval, a Bayesian interval(considered in detail in Chapter 3) and Newcombe's hybrid score interval.They find their exact coverage probabilities and mean expected lengths at some specific pairs (n_1, n_2) with p_1 and p_2 taking values from the unit square. It is shown that the Agresti-Coull interval has better coverage performance than Newcombe's hybrid score interval.

The above results all involve non-adaptive designs with constant sample sizes. The sample sizes in adaptive designs are not constants but random variables. To distinguish from non-adaptive design, we use $N_i(k)$, $S_i(k)$ to denote the sample size and the number of successes from treatment i for $i = 1, 2$.

Wei *et al.* (1990) studied the interval estimation problem for $p_1 - p_2$ and a specific adaptive design: randomized play-the-winner design, which is due to Wei and Durham (1978) and tends to assign more study subjects to the better treatment. Wei *et al.* (1990) developed a network algorithm to find the joint distribution of $(N_1(k), S_1(k) + S_2(k), S_1(k))$, through which exact confidence intervals for $p_1 - p_2$ could be derived. The authors suggest using this method when the sample size n is small or moderate. There are two disadvantages of this method which limit its wide application . First, though the network algorithm can be easily modified to accommodate other adaptive designs, the computation of the joint distribution of $(N_1(k), S_1(k) + S_2(k), S_1(k))$ is not very easy. Second, the exact confidence interval does not have an explicit form. Wei *et al.* (1990) also evaluated the Wald interval and the profile likelihood based interval for $p_1 - p_2$ with randomized play-the-winner designs though simulation. They found that the Wald interval was rather anti-conservative. The profile likelihood method was recommended in Wei *et al.* (1990) for a moderate-sized or large sample design.

1.3 Application

All the confidence intervals studied in this dissertation are based on asymptotic theory. However, some confidence intervals behave rather well even when sample sizes are small or moderate. Therefore, one may use confidence intervals that will be suggested for a broad range of sample sizes.

Moreover, the simplicity of all the confidence intervals except the profile likelihood interval is an attractive feature from the point of view of applications.

People who wish to perform adaptive designs have a wide variety of adaptive allocation procedures at their disposal. And the corresponding asymptotic theories of quite a lot adaptive designs are also reported. There are numerous references on the interval estimation problem for $p_1 - p_2$. But most of them concentrate on non-adaptive designs. We hope this dissertation will be useful for constructing good confidence intervals for $p_1 - p_2$, especially in adaptive designs.

Chapter 2

Wald Interval Estimation for the Difference of two Binomial Proportions

2.1 Introduction

Interval estimation for a single binomial proportion and the difference of two binomial proportions are used extensively in practice and have been widely discussed in the literature. It is well known that the standard Wald intervals behave poorly. Brown *et al.* (2002) focused on the interval estimation for one binomial proportion and explored the reason why the coverage probabilities of the Wald interval for one binomial proportion are often far less than the nominal level even when the sample size is moderate or quite large. They evaluated the approximate coverage probabilities and expected lengths of the Wald interval for one binomial proportion and its candidate replacements. Inspired by their article, we studied interval estimation for the difference of two binomial proportions.

In Section 2, we focus on the poor performance of the Wald interval of the difference of two binomial proportions by exhibiting its behavior through a few examples. As will be shown, the Wald interval for the difference of two binomial proportions, defined in (2.2.1), shares some similar properties to those of one binomial proportion addressed in Brown *et al.* (2002). For example, the discreteness of the Binomial distribution leads to oscillatory coverage probabilities and the true coverage probabilities often differ significantly from the nominal level even when the two proportions are near 0.5 and sample sizes are moderate or large. We also note that unbalanced sample sizes, when the two proportions are close, among some other issues, may have severe effects on the coverage probabilities. In Section 3, we explore the reason for the poor performance of the Wald interval. Section 4 deals with a smooth approximation of the coverage probability of the Wald interval by applying Edgeworth Expansion methods.

2.2 Coverage Properties of the Wald Confidence Interval

Let X_1 and X_2 be two independent random variables, $X_i \sim Bin(n_i, p_i)$, where $p_i \in (0,1)$ for $i = 1,2$. Let $\hat{p}_i = X_i/n_i$. As mentioned in Chapter 1, the $100(1-\alpha)\%$

Wald confidence interval for $p_1 - p_2$ is

$$\hat{p_1} - \hat{p_2} \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}, \tag{2.2.1}$$

where $z_{\frac{\alpha}{2}}$ denotes the $100(1-\frac{\alpha}{2})$ percentile of the standard normal distribution. We will use $CI_W(n_1, n_2, X_1, X_2)$ to denote this interval and $CP_W(n_1, n_2, p_1, p_2)$ to denote its exact coverage probability. Then

$$\begin{aligned} & CP_W(n_1, n_2, p_1, p_2) \\ =& P\{p_1 - p_2 \in CI_W(n_1, n_2, X_1, X_2)\} \qquad (2.2.2) \\ =& \sum_{x_1=0}^{n_1}\sum_{x_2=0}^{n_2}\binom{n_1}{x_1}p_1^{x_1}(1-p_1)^{(n_1-x_1)}\binom{n_2}{x_2}p_2^{x_2}(1-p_2)^{(n_2-x_2)}I_{A_p}(x_1, x_2) \end{aligned}$$

where $A_p = \{(x_1, x_2) | p_1 - p_2 \in CI_W(n_1, n_2, x_1, x_2)\}$. We will present a few examples to show that the coverage probabilities of the Wald interval are typically lower than its nominal level.

The probabilities reported in the following plots and tables unless otherwise specified, are the result of exact probability calculations produced in S-Plus. Instead of using the algorithm given in equation 2.2.2, which contains two loops, we apply a more efficient one:

$$\begin{aligned} & CP_W(n_1, n_2, p_1, p_2) \\ =& \sum_{i=0}^{n_2} P\{L_W(n_1, n_2, p_1, p_2, i) < X_1 < U_W(n_1, n_2, p_1, p_2, i)\}P\{X_2 = i\} \qquad (2.2.3) \end{aligned}$$

where $L_W(n_1, n_2, p_1, p_2, i) < U_W(n_1, n_2, p_1, p_2, i)$ are two proper real roots of the equation as a function of X_1:

$$p_1 - p_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{X_1(1-X_1)}{(n_1)^3} + \frac{i(1-i)}{(n_2)^3}}.$$

The algorithm given in (2.2.3) contains only one loop.

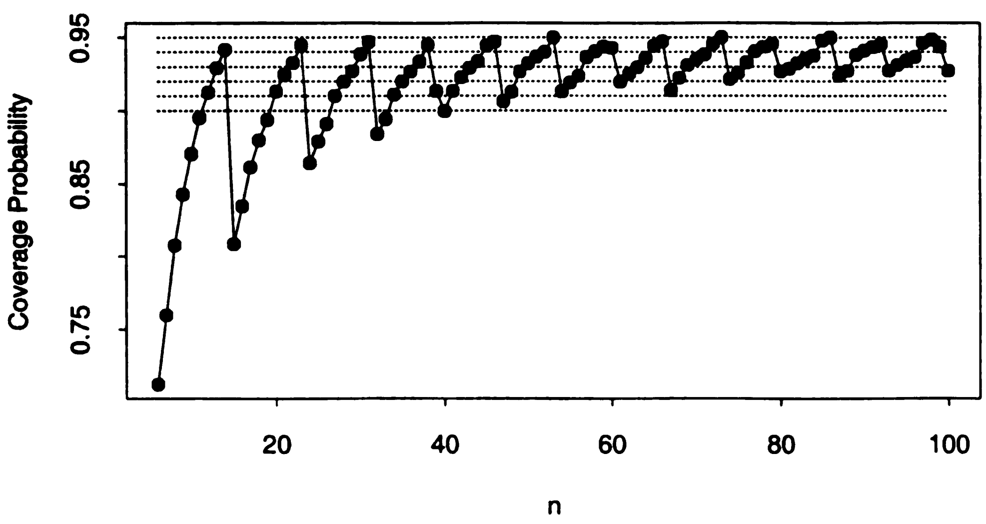
Example 1. Figure 1 plots the exact coverage probabilities of the nominal 95% standard Wald interval for $p_1 = 0.9, p_2 = 0.1$, $n_1 = n_2 = n$ when n varies from 6 to 100. Two important features of the Wald interval are exhibited in the figure.

First, there exists a very strong oscillation which is due to the discreteness of the binomial distribution. Therefore, the coverage probability does not at all get steadily closer to the nominal level though the magnitude of the oscillation tends to decrease. For example, at $n = 14$, the coverage probability is 0.942, but it is only 0.808 at $n = 15$. Even when n is as large as 67, the coverage probability is only 0.914. When $n = 100$, the coverage probability is still not satisfactory, it is only 0.927. Only after $n \geq 300$ does the coverage probability fluctuate above 0.94.

Second, the Wald interval is anti-conservative: the coverage probabilities at most values of n are less than the nominal level. Among all the coverage probabilities (for $n = 6$ up to $n = 100$), only three reach the nominal level. There are 50 coverage probabilities less than 0.93 and 31 less than 0.92.

Similar to the phenomenon pointed out in Brown *et al.* (2002) for one sample interval, the existence of the oscillation of the coverage probability makes some quadruples lucky and some unlucky. For instance, the quadruple $(n_1, n_2, p_1, p_2) = (53, 53, 0.9, 0.1)$ is lucky, with the exact coverage probability of the Wald interval equal to 0.9501. But $(n_1, n_2, p_1, p_2) = (54, 54, 0.9, 0.1)$ is unlucky, with the exact coverage probability 0.9132. Similarly, changing the proportions may result in some lucky or unlucky quadruples as well. Further, the lucky or unlucky quadruples

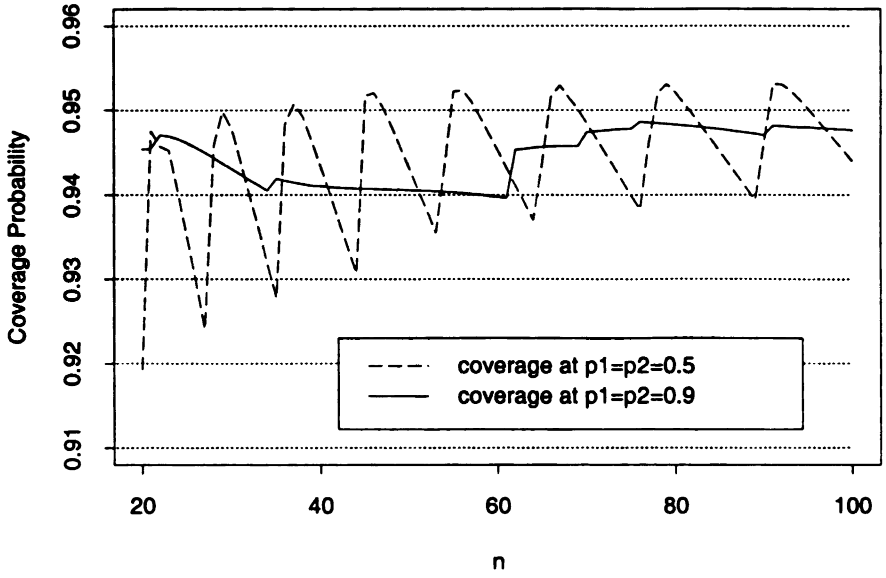
Figure 2.1: Exact coverage probability of the nominal 95% Wald interval for $p_1 = 0.9, p_2 = 0.1$ and $n1 = n2 = n = 6$ to 100



are not predictable. There is no obvious pattern to follow on telling whether a quadruple is lucky or not.

Example 2. Suppose the confidence level is 95%, $n_1 = n_2 = n$ and n varies from 20 to 100. Compare the coverage probabilities in two cases. Case one, $p_1 = p_2 = 0.9$; case two, $p_1 = p_2 = 0.5$. Conventional wisdom might suggest that the coverage probabilities in case 2 would be higher than those in case one. But this is not true. Figure 2 plots the coverage probabilities in the two cases. It is surprising to see $CP_W(n, n, 0.5, 0.5)$ is not obviously higher than $CP_W(n, n, 0.9, 0.9)$. When n varies from 20 to 100, $CP_W(n, n, 0.9, 0.9)$ has less oscillation. All $CP_W(n, n, 0.9, 0.9)$ are located between 0.940 and 0.949. The range of $CP_W(n, n, 0.5, 0.5)$ is $[0.919, 0.953]$. This example demonstrates we cannot judge the coverage probabilities of the Wald

Figure 2.2: Exact coverage probability of the nominal 95% Wald intervals for $p_1 = p_2 = 0.5$ and $p_1 = p_2 = 0.9$ with $n_1 = n_2 = n = 20$ to 100

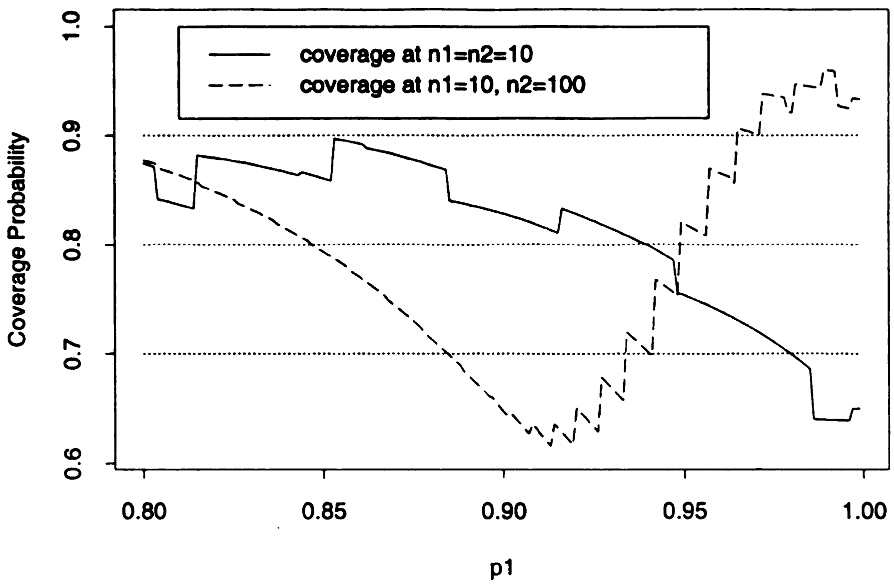


interval by whether p_1 and p_2 are close to center or not. The relative positions of p_1 and p_2 affect the coverage probability.

Since there are four quantities n_1, n_2, p_1 and p_2 affecting the coverage probability, considering only the magnitudes of proportions is not enough. In fact, not only the relative positions of p_1 and p_2 but also the relative sizes of n_1 and n_2 may influence $CP_W(n_1, n_2, p_1, p_2)$ significantly. Moreover, the four quantities interact.

Example 3. Fix $p_1 = p_2 = 0.9$. Consider the coverage probabilities at $n_1 = n_2 = 10$ and $n_1 = 10, n_2 = 100$ and nominal confidence level 95%. Which coverage probability is greater? It is striking to see that $CP_W(10, 10, 0.9, 0.9) = 0.8282$ and $CP_W(10, 100, 0.9, 0.9) = 0.6474$. The large sample size does not improve but

Figure 2.3: Exact coverage probability of the nominal 95% Wald interval at$n_1 = n_2 = 10$ and $n_1 = 10$, $n_2 = 100$ with $p_2 = 0.9$ and $p_1 = 0.8$ to 0.999 with jump size 0.001



reduce the coverage probability in this special case. And the big difference between the two coverage probabilities cannot be explained only by the phenomenon of oscillation. This suggests the drop on the coverage probability in case two is caused by unbalanced sample sizes. Figure 3 plots the coverage probabilities for p_1 varying between 0.8 and .999 with step size 0.001.

Table 1 and 2 list some coverage probabilities under different confidence levels for some p_1, p_2, n_1 and n_2. Observe how much the sample sizes might affect the coverage probability. From Figure 2.3 and the two tables, we may conclude that a larger sample size on one n_i could not guarantee a better coverage probability. The

Table 2.1: Exact coverage probability of the nominal 95% Wald interval

n_1	10	10	10	30	30	30	100	100	100
n_2	10	30	100	10	30	100	10	30	100
$p_1 = .9, p_2 = .5$	.911	.920	.804	.906	.934	.937	.896	.940	.947
$p_1 = .9, p_2 = .1$	.871	.849	.688	.849	.938	.908	.688	.908	.927
$p_1 = .8, p_2 = .3$	.894	.900	.877	.898	.940	.929	.893	.931	.939
$p_1 = .6, p_2 = .4$	.922	.913	.908	.913	.934	.941	.908	.941	.948
$p_1 = .9, p_2 = .8$	.949	.869	.647	.869	.945	.918	.647	.918	.948
$p_1 = .5, p_2 = .5$	.912	.917	.905	.917	.948	.939	.905	.939	.944

Table 2.2: Exact coverage probability of the nominal 99% Wald interval

n_1	10	10	10	30	30	30	100	100	100
n_2	10	30	100	10	30	100	10	30	100
$p_1 = .9, p_2 = .5$	.963	.963	.892	.968	.984	.978	.968	.982	.988
$p_1 = .9, p_2 = .1$	.878	.856	.754	.856	.946	.953	.754	.953	.982
$p_1 = .8, p_2 = .3$	.972	.948	.895	.954	.976	.979	.958	.981	.986
$p_1 = .6, p_2 = .4$	.974	.964	.954	.964	.982	.983	.954	.983	.988
$p_1 = .9, p_2 = .8$	.991	.973	.692	.973	.991	.956	.692	.956	.989
$p_1 = .5, p_2 = .5$	.958	.966	.967	.966	.986	.985	.967	.985	.987

relative magnitudes(balanced or not) of the two sample sizes is another issue that influence the coverage probability significantly.

It is obvious from above examples that the exact coverage probability of Wald interval seldom achieves the nominal level. We will examine the reason theoretically in next section.

At the end of this section, it is worthwhile to mention an issue that might cause a non-negligible loss of the coverage probability of the Wald interval. Unlike a lot of alternative intervals, the Wald interval is sensitive to whether a confidence interval is defined as open or closed. The next remark gives such an example. Neither Brown *et al.* (2002) nor Agresti and Coull (1998) specifically mentioned whether their confidence intervals were closed or not. But their results are consistent with open confidence intervals. In Wei *et al.* (1990), the authors specified open confidence intervals. In this report, we define a confidence interval to be open.

Remark 2.2.1. The shrinkage of the Wald interval to an empty set, (a, a), at some realizations of $(n_1, p_1; n_2, p_2)$ can cause its poor coverage performance, especially when both sample sizes are small and both proportions approach boundaries. The coverage probability of the Wald interval is at most $1 - (p_1^{n_1} + q_1^{n_1})(p_2^{n_2} + q_2^{n_2})$ regardless of the nominal level. For example, when $p_1 = p_2 = 0.95$ and both sample sizes are 20, the coverage probability of the Wald interval is at most 0.872 regardless of the nominal level. A more simple and instructive example is $n_1 = n_2 = 10$ and $p_1 = p_2 = 0.9$. If $X_1 = X_2 = 10$, then the confidence interval shrinks to $(0, 0)$. Though $p_1 - p_2 = 0$, but since $0 \notin (0, 0)$, $(10, 10)$ is not a proper pair in the A_p

defined at the beginning of this section. Note that $P\{X_1 = X_2 = 10\} = 0.1216$, which makes $CP_W(10, 10, 0.9, 0.9)$ at most 0.8784.

2.3 A Reason for Inadequate Coverage

Similar to the reason for the inadequate coverage of the Wald interval for one binomial proportion explored in Brown *et al.* (2002), we will show that the poor performance of the Wald interval is due mainly to the fact that the Wald confidence interval is symmetric about a "wrong" center. Although $\hat{p_1} - \hat{p_2}$ is the MLE and an unbiased estimator of $p_1 - p_2$, as the center of a confidence interval it causes a systematic loss of coverage from the nominal level. As we will see in next Chapter, by simply recentering the interval, one can improve the coverage performance significantly.

One way to derive the Wald interval is to invert the large-sample Wald test. The nominal $(1-\alpha)\%$ Wald interval for $p_1 - p_2$ is the set of δ_p for which

$$\frac{|\hat{p}_1 - \hat{p}_2 - \delta_p|}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}} \leq Z_{\alpha/2}$$

Hence, in deriving the Wald interval, the following consequence of the central limit theorem plays an important role:

$$\frac{\hat{p_1} - \hat{p_2} - (p_1 - p_2)}{\sqrt{\hat{p_1}(1-\hat{p_1})/n_1 + \hat{p_2}(1-\hat{p_2})/n_2}} \stackrel{\mathcal{L}}{\Longrightarrow} N(0,1)$$

For simplicity, denote the left side above W_{n_1,n_2}. Even for quite large values of n_1 and n_2, the actual distribution of W_{n_1,n_2} can be far from the the standard normal distribution for many p_1 and p_2 as we will show next. Thus the very premise on which the Wald interval is based is seriously flawed for moderate and even quite large values of n_1 and n_2.

The bias of W_{n_1,n_2}, which is EW_{n_1,n_2}, from the mean of standard normal distribution can be analytically computed by doing standard expansions. Denote $\omega_{n_i} = \hat{p}_i - p_i$ for $i = 1, 2$. Then simple algebra gives

$$W_{n_1,n_2} = \frac{\omega_{n_1} - \omega_{n_2}}{\sqrt{\frac{p_1q_1+(q_1-p_1)\omega_{n_1}-\omega_{n_1}^2}{n_1} + \frac{p_2q_2+(q_2-p_2)\omega_{n_2}-\omega_{n_2}^2}{n_2}}}$$

where $q_i = 1 - p_i$ for $i = 1, 2$. Let $u = \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}$. Denote the denominator b, then $1/b$ can be expressed as

$$\begin{aligned} &u^{-1/2}\left(1 + \frac{(q_1 - p_1)\omega_{n_1}}{n_1u} + \frac{(q_2 - p_2)\omega_{n_2}}{n_2u} - \left(\frac{\omega_{n_1}^2}{n_1u} + \frac{\omega_{n_2}^2}{n_2u}\right)\right)^{-\frac{1}{2}} \\ =&u^{-1/2}(1+x)^{-1/2}, \end{aligned}$$

where $x = \frac{(q_1-p_1)\omega_{n_1}}{n_1u} + \frac{(q_2-p_2)\omega_{n_2}}{n_2u} - \left(\frac{\omega_{n_1}^2}{n_1u} + \frac{\omega_{n_2}^2}{n_2u}\right)$. Since $\omega_{n_i} = O_p(n_i^{-1/2})$, a Taylor expansion yields

$$W_{n_1,n_2} = (\omega_{n_1} - \omega_{n_2})u^{-1/2}(1 - \frac{x}{2} + \frac{3x^2}{8} - \frac{5x^3}{16} + o_p((n_1 \wedge n_2)^{-3/2}))$$

The formulas for central moments of the binomial distribution then yield an approximation to the bias:

$$\begin{aligned} EW_{n_1,n_2} =& \frac{p_1 - 1/2}{n_1u_1u^{1/2}}\left(1 + \frac{1}{n_1}(\frac{9}{2u_1} - 1)\right) - \frac{p_2 - 1/2}{n_2u_2u^{1/2}}\left(1 + \frac{1}{n_2}(\frac{9}{2u_2} - 1)\right) \\ &+ \frac{3((p_1 - 1/2) - (p_2 - 1/2))}{2n_1n_2u_1u_2u^{1/2}} \\ &- \frac{15(p_1 - 1/2)(p_2 - 1/2)}{2n_1n_2u_1u_2u^{3/2}}\left(\frac{p_2 - 1/2}{n_2} - \frac{p_1 - 1/2}{n_1}\right) \\ &+ o((n_1 \wedge n_2)^{-3/2}) \qquad (2.3.1) \end{aligned}$$

where

$$u_1 = 1 + \frac{p_2 q_2}{p_1 q_1}\frac{n_1}{n_2} \qquad \text{and} \qquad u_2 = 1 + \frac{p_1 q_1}{p_2 q_2}\frac{n_2}{n_1}$$

From (2.3.1), it can be seen that when both p_1 and p_2 approach $1/2$ for fixed sample sizes, the bias tends to decrease. Therefore, ignoring the oscillation effect, one can expect to increase the coverage probability by shifting the terms of the center of Wald confidence interval from $\hat{p_1}$ and $\hat{p_2}$ toward $1/2$ and $1/2$. When the two proportions are close for comparable sample sizes, the effect from p_1 could counteract that from p_2. On the other hand, it also explains why the Wald confidence interval behaves poorly when the sample sizes are extremely unbalanced for some p_1 and p_2: the effect from p_1 cannot cancel out most effect from p_2. In general, equation (2.3.1) can be used as a rule of thumb to explain how interaction among n_1, n_2, p_1 and p_2 affects the bias and thus the coverage probability.

2.4 A smoothing formula obtained by Edgeworth Expansion methods

In this section, we will not justify Edgeworth Expansions, but rather will use Edgeworth expansion techniques to derive a formula that works well in approximating coverage probabilities of the Wald interval in a variety of settings. See Bhattacharya and Ranga Rao (1976) and Hall (1992) for more details on Edgeworth Expansions.

First a theorem from Hall (1992) on Edgeworth expansion is presented. It gives general conditions under which the Edgeworth expansion is valid and will be used as a tool to derive the smooth approximation of the coverage probability of the Wald interval.

Theorem 2.4.1. ***(Hall, 1992, page 56)*** *Let* $\mathbf{X}, \mathbf{X_1}, \mathbf{X_2}, ...,$ *be independent and identically distributed random column d-vectors with mean* μ*, and put* $\overline{\mathbf{X}} = n^{-1}\sum_{i=1}^{n}\mathbf{X_i}$ *. Assume the function* $A : \mathbb{R}^d \to \mathbb{R}$ *has* $j+2$ *continuous derivatives in a neighborhood of* $\mu = E(\mathbf{X})$*, that* $A(\mu) = 0$*, that* $E(\|\mathbf{X}\|^{j+2}) < \infty$*, and that the characteristic function* χ *of* $\mathbf{X}$ *satisfies*

$$\limsup_{\|\mathbf{t}\|\to\infty} |\chi(\mathbf{t})| < 1. \tag{2.4.1}$$

The above inequality is called Cramer's condition. Denote the asymptotic variance of $n^{1/2}A(\overline{\mathbf{X}})$ *by* σ^2*. Suppose* $\sigma > 0$*. Then for* $j \geq 0$*,*

$$\begin{aligned} P(n^{1/2}A(\overline{\mathbf{X}})/\sigma \leq x) =& \Phi(x) + n^{-1/2}r_1(x)\phi(x) + n^{-1}r_2(x)\phi(x) + \cdots \\ & + n^{-j/2}r_j(x)\phi(x) + o(n^{-j/2}) \end{aligned} \tag{2.4.2}$$

uniformly in x*, where* r_j *is a polynomial of degree at most* $3j-1$*, odd for even* j *and even for odd* j*, with coefficients depending on moments of* $\mathbf{X}$ *up to order* $j+2$*.*

According to the arguments (pages 47, 48) in Hall (1992), the r_j for $j = 1, 2$ in Theroem 2.4.1 are given by

$$r_1(x) = -\left\{k_{1,2} + \frac{1}{6}k_{3,1}(x^2-1)\right\} \tag{2.4.3}$$

and

$$r_2(x) = -x\left\{\frac{1}{2}(k_{2,2}+k_{1,2}^2) + \frac{1}{24}(k_{4,1}+4k_{1,2}k_{3,1})(x^2-3) + \frac{1}{72}k_{3,1}^2(x^4-10x^2+15)\right\}, \tag{2.4.4}$$

where those k's can be determined through the following expressions of κ's that may are expanded in terms of k's as a power series in n^{-1}

$$\kappa_{j,n} = n^{-(j-2)/2}(k_{j,1} + n^{-1}k_{j,2} + n^{-2}k_{j,3} + ...), j \geq 1 \qquad (2.4.5)$$

Let $S_n = n^{1/2}(\hat{\theta} - \theta_0)/\hat{\sigma}$. The κ's are defined by

$$\kappa_{1,n} = E(S_n)$$

$$\kappa_{2,n} = E(S_n^2) - (E(S_n))^2 = var(S_n)$$

$$\kappa_{3,n} = E(S_n^3) - 3E(S_n^2)E(S_n) + 2(E(S_n))^3 = E(S_n - ES_n)^3$$

$$\kappa_{4,n} = E(S_n^4) - 4E(S_n^3)E(S_n) - 3(E(S_n^2)^2 + 12E(S_n^2)(E(S_n))^2 - 6(E(S_n))^4$$

$$= E(S_n - ES_n)^4 - 3(var(S_n))^2 \qquad (2.4.6)$$

To derive the smooth approximation of the coverage probability of the Wald interval, we define some notation. Let $\{Y_{1,j} : j = 1, 2, \dots\}$ and $\{Y_{2,j} : j = 1, 2, \dots\}$ be two independent sequences of independent Bernoulli random variables, $Y_{i,j} \sim Bernoulli(p_i)$ and let $X_i = \sum_{j=1}^{n_i} Y_{i,j}$, where $i = 1, 2$. Let γ and κ stand for the skewness and kurtosis of $D = Y_{1,1} - Y_{2,1}$ respectively. Then

$$\gamma = E(D - ED)^3 = E\left(Y_{1,1} - Y_{2,1} - (p_1 - p_2)\right)^3$$

$$= p_1q_1(q_1 - p_1) - p_2q_2(q_2 - p_2)$$

and

$$\kappa = E(D - ED)^4 - 3(var(D))^2$$

$$= E\left(Y_1 - Y_2 - (p_1 - p_2)\right)^4 - 3(var(Y_1 - Y_2))^2$$

$$= p_1q_1(q_1 - p_1)^2 + p_2q_2(q_2 - p_2)^2 - 2p_1^2q_1^2 - 2p_2^2q_2^2$$

We do not have appropriate random variables to apply Theorem 2.4.1 directly since Bernoulli random variables do not satisfy Cramer's condition. In general, absolutely continuous random variables satisfy Cramer's condition. Therefore, we need to smooth Bernoulli random variables first. However, there is another problem arising after smoothing: the Wald test statistic, through which we may define the exact coverage probability of the Wald interval, is

$$W_{n_1,n_2} = \frac{\hat{p_1} - \hat{p_2} - (p_1 - p_2)}{\sqrt{\hat{p_1}(1-\hat{p_1})/n_1 + \hat{p_2}(1-\hat{p_2})/n_2}},$$

on the set

$$\Pi_{n_1,n_2} = \left\{ \frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2} > 0 \right\}$$

and has no definition on

$$\overline{\Pi}_{n_1,n_2} = \left\{ \frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2} = 0 \right\}.$$

Consequently, we need to consider the smoothing random variables on Π_{n_1,n_2}. But Theorem 2.4.1 does not apply to random variables defined on a proper subset.

Since $P(\overline{\Pi}_{n,n}) = (p_1^n + q_1^n)(p_2^n + q_2^n)$, which is of higher order of $O(n^{-3/2})$. Hence, the probability that W_{n_1,n_2} has no definition can be absorbed in $O(n^{-3/2})$. And a smooth approximation of the coverage probability of the Wald interval will be given in an expression with error term $O(n^{-3/2})$.

What would happen if the Edgeworth Expansions were theoretically valid on the subset Π_{n_1,n_2}? We will focus on Π_{n_1,n_2} henceforth.

For simplicity, we consider the case when $n_1 = n_2 = n \geq 2$.

The procedure of deriving the smooth approximation contains four steps. First, we create two sequences of random variables and define a statistic $T_{n,n}$ by using those created random variables. We then show statistic $T_{n,n}$ can be used to approximate the exact coverage probability of the Wald interval. In step 2, we verify the validity of doing Edgeworth expansion for statistic $T_{n,n}$ if the expansions were valid on a subset. The Edgeworth expansion for $T_{n,n}$ is derived in step 3. Last, in step 4, the smoothing formula of the coverage probability of the Wald interval is given by applying results from the first three steps.

Step 1. We first create two sequences of random variables to be used in the Edgeworth expansion and show the exact coverage probability of the Wald interval can be approximated using a statistic defined through the created random variables.

Suppose $\xi_{i,j}$ and $\eta_{i,j}$ are two independent sequences of i.i.d random variables for $j = 1, 2, \cdots$, both are independent of $Y_{i,j}$ for $i = 1, 2$ and $\xi_{i,j} \sim U(-1/n^4, 1/n^4)$, $\eta_{i,j} \sim U(1 - 1/n^4, 1 + 1/n^4)$. For $i = 1, 2$ and $j = 1, 2, \cdots$, define

$$T_{i,j} = \xi_{i,j}[Y_{i,j} = 0] + \eta_{i,j}[Y_{i,j} = 1]. \tag{2.4.7}$$

Then

$$Y_{i,j} - \frac{1}{n^4} < T_{i,j} < Y_{i,j} + \frac{1}{n^4} \tag{2.4.8}$$

Put $\overline{T_i} = \sum_{j=1}^{n} T_{i,j}/n$. The following inequality holds by applying inequality (2.4.8)

$$\overline{T_i} - \frac{1}{n^4} < \hat{p}_i < \overline{T_i} + \frac{1}{n^4}. \tag{2.4.9}$$

Consider the quantity under the square root in $W_{n,n}$.

$$\begin{aligned}
&\frac{\hat{p_1}(1-\hat{p_1})}{n}+\frac{\hat{p_2}(1-\hat{p_2})}{n} \\
<&\frac{(\overline{T_1}+1/n^4)(1-\overline{T_1}+1/n^4)}{n}+\frac{(\overline{T_2}+1/n^4)(1-\overline{T_2}+1/n^5)}{n} \\
<&\frac{\overline{T_1}(1-\overline{T_1})}{n}+\frac{\overline{T_2}(1-\overline{T_2})}{n}+\frac{3}{n^5}.
\end{aligned} \tag{2.4.10}$$

Similarly,

$$\frac{\hat{p_1}(1-\hat{p_1})}{n}+\frac{\hat{p_2}(1-\hat{p_2})}{n} > \frac{\overline{T_1}(1-\overline{T_1})}{n}+\frac{\overline{T_2}(1-\overline{T_2})}{n}-\frac{3}{n^5} \tag{2.4.11}$$

Note that $\frac{\overline{T_1}(1-\overline{T_1})}{n}+\frac{\overline{T_2}(1-\overline{T_2})}{n}-\frac{3}{n^5} \leq 0$ if and only if $\frac{\hat{p_1}(1-\hat{p_1})}{n}+\frac{\hat{p_2}(1-\hat{p_2})}{n}=0$, which is out of our consideration according to previous analysis. Then, on $\Pi_{n,n}$, define

$$T_{n,n}=\frac{\overline{T_1}-\overline{T_2}-(p_1-p_2)}{\sqrt{\frac{\overline{T_1}(1-\overline{T_1})}{n}+\frac{\overline{T_2}(1-\overline{T_2})}{n}}}.$$

Therefore, applying inequalities (2.4.10) and (2.4.11), the following inequality chain holds,

$$\frac{\overline{T_1}-\overline{T_2}-(p_1-p_2)-2/n^4}{\sqrt{\frac{\overline{T_1}(1-\overline{T_1})}{n}+\frac{\overline{T_2}(1-\overline{T_2})}{n}+\frac{3}{n^5}}} < W_{n,n} < \frac{\overline{T_1}-\overline{T_2}-(p_1-p_2)+2/n^4}{\sqrt{\frac{\overline{T_1}(1-\overline{T_1})}{n}+\frac{\overline{T_2}(1-\overline{T_2})}{n}-\frac{3}{n^5}}}.$$

Further more, a few lines of expansions and algebra yield,

$$T_{n,n}(1-2/n^3)-n^{-2} < W_{n,n} < T_{n,n}(1+2/n^3)+n^{-2}$$

Therefore, the coverage probability of Wald interval satisfies

$$P_W = P\left(|W_{n,n}| \le Z_{\alpha/2}\right)$$
$$=P\left(W_{n,n} \le Z_{\alpha/2}\right) - P\left(W_{n,n} < -Z_{\alpha/2}\right)$$
$$\le P\left(T_{n,n} \le (Z_{\alpha/2} + n^{-2})/(1-2/n^3)\right) - P\left(T_{n,n} < -(Z_{\alpha/2} + n^{-2})/(1+2/n^3)\right) \tag{2.4.12}$$

Similarly,

$$P\left(|W_{n,n}| \le Z_{\alpha/2}\right) \tag{2.4.13}$$
$$\ge P\left(T_{n,n} \le (Z_{\alpha/2} - n^{-2})/(1+2/n^3)\right) - P\left(T_{n,n} < -(Z_{\alpha/2} - n^{-2})/(1-2/n^3)\right) \tag{2.4.14}$$

Step 2, verify the validity of performing an Edgeworth expansion for $T_{n,n}$. Define a sequence of random vectors $\mathbf{Z}_j = \begin{pmatrix} T_{1,j} \\ T_{2,j} \end{pmatrix}$, where $j = 1, 2, \cdots$. Note that

$$ET_{i,j} = E(\xi_{i,j}[Y_{i,j}=0] + \eta_{i,j}[Y_{i,j}=1])$$
$$= E(\xi_{i,j})(1-p_i) + E(\eta_{i,j})p_i = p_i$$

so the $\mathbf{Z_j}'s$ are *i.i.d* random vectors satisfying $\mu = E(\mathbf{Z}_j) = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$. Let $\overline{\mathbf{Z}} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{Z}_j = \begin{pmatrix} \overline{T_1} \\ \overline{T_2} \end{pmatrix}$. For any vector $\mathbf{x} = (x^{(1)}, x^{(2)}) \in (0,1)^2$, define

$$A(\mathbf{x}) = \frac{x^{(1)} - x^{(2)} - (p_1 - p_2)}{\sqrt{(x^{(1)}(1-x^{(1)}) + x^{(2)}(1-x^{(2)}))}}$$

Then $A(\mathbf{x})$ is a infinitely differentiable function with $A(\overline{\mathbf{Z}}) = n^{1/2}T_{n,n}$ and $A(\mu) = 0$.

Note that the asymptotic variance of $n^{1/2}A(\overline{\mathbf{Z}})$ is 1, according to Theorem 2.4.1,

$$
\begin{aligned}
P\{n^{1/2}A(\overline{\mathbf{Z}}) \leq x\} &= P\{n^{1/2}T_{n,n,p_1,p_2} \leq x\} \\
&= \Phi(x) + n^{-1/2}r_1(x)\phi(x) + n^{-1}r_2(x)\phi(x) + \cdots \\
&+ n^{-k/2}r_k(x)\phi(x) + o(n^{-k/2}).
\end{aligned}
$$

The above expansion is valid as an asymptotic series to k term if for any positive integer j,

$$E(||\mathbf{Z}_j||^{k+2}) < \infty \tag{2.4.15}$$

and

$$\limsup_{|t^{(1)}|+|t^{(2)}|\to\infty} |E\exp(it^{(1)}Z_j^{(1)} + it^{(2)}Z_j^{(2)})| < 1. \tag{2.4.16}$$

By an argument (page 65) in Hall (1992), *Cramer*'s condition (2.4.16) holds if the distribution of the random vector $\mathbf{Z}_j$ has a non-degenerate, absolutely continuous component, which is satisfied in current settings. The former inequality (2.4.15) is guaranteed by

$$E(||\mathbf{Z}_j||^{k+2}) = E(|Y_{1,j}|^2 + |Y_{2,j}|^2)^{\frac{k+2}{2}} \leq 4^{\frac{k+2}{2}}.$$

Step 3, derive the two-term Edgeworth expansion of $n^{1/2}A(\overline{\mathbf{Z}})$.

For simplicity, use S_n to denote $n^{1/2}A(\overline{\mathbf{Z}})$. Let $W_i = \overline{T_i} - p_i$, for $i = 1, 2$, then

$$
\begin{aligned}
E(W_i) &= 0 \qquad E(W_i^2) = \frac{p_i q_i}{n} + o(n^{-8}) \\
E(W_i^3) &= \frac{p_i q_i (q_i - p_i)}{n^2} + o(n^{-8}) \\
E(W_i^4) &= \frac{p_i q_i (p_i^3 + q_i^3) + 3(n-1)p_i^2 q_i^2}{n^3} + o(n^{-8}) \sim O(n^{-2}) \\
E(W_i^5) &= \frac{10 p_i^2 q_i^2 (q_i - p_i)}{n^3} + o(n^{-4}) + o(n^{-8}) \sim O(n^{-3}) \\
E(W_i^6) &= \frac{15 p_i^3 q_i^3}{n^3} + o(n^{-4}) + o(n^{-8}) \sim O(n^{-3})
\end{aligned}
\tag{2.4.17}
$$

Then with $S_n = n^{1/2}A(\overline{\mathbf{Z}})$ and $W_i = O_p(n^{-1/2})$,

$$
\begin{aligned}
S_n &= n^{1/2}(W_1 - W_2)\left((W_1 + p_1)(-W_1 + q_1) + (W_2 + p_2)(-W_2 + q_2)\right)^{-1/2} \\
&= n^{1/2}(W_1 - W_2)\left(p_1 q_1 + p_2 q_2 + (q_1 - p_1)W_1 + (q_2 - p_2)W_2 - (W_1^2 + W_2^2)\right)^{-1/2} \\
&= n^{1/2}(W_1 - W_2)\tau^{-1/2}\left(1 + \frac{(q_1 - p_1)}{\tau}W_1 + \frac{(q_2 - p_2)}{\tau}W_2 - \frac{1}{\tau}(W_1^2 + W_2^2)\right)^{-1/2} \\
&= n^{1/2}\tau^{-1/2}(W_1 - W_2)\{1 - \frac{1}{2}\tau^{-1}\sum_{i=1}^{2}(q_i - p_i)W_i \\
&+ \frac{1}{2}\tau^{-1}\sum_{i=1}^{2}[1 + \frac{3}{4}\tau^{-1}(q_i - p_i)^2]W_i^2 + \frac{3}{4}\tau^{-2}(q_1 - p_1)(q_2 - p_2)W_1 W_2\} \\
&+ O_p(n^{-3/2})
\end{aligned}
\tag{2.4.18}
$$

Therefore, apply the moment equations (2.4.17) and the independence of W_1 and

W_2, we have

$$E(S_n) = n^{1/2}\tau^{-1/2}E\left((W_1 - W_2)(1 - \frac{1}{2}\tau^{-1}\sum_{i=1}^{2}(q_i - p_i)W_i)\right) + O(n^{-1})$$

$$= -\frac{1}{2}n^{-1/2}\tau^{-3/2}\gamma + O(n^{-1})$$

$$E(S_n^2) = n\tau^{-1}E\{(W_1 - W_2)^2(1 - \tau^{-1}\sum_{i=1}^{2}(q_i - p_i)W_i$$

$$+ \tau^{-1}\sum_{i=1}^{2}(1 + \tau^{-1}(q_i - p_i)^2)W_i^2$$

$$+ 2\tau^{-2}(q_1 - p_1)(q_2 - p_2)W_1W_2)\} + O(n^{-3/2})$$

$$= 1 + n^{-1} + 2n^{-1}\tau^{-2}(p_1^2q_1^2 + p_2^2q_2^2) + 2\tau^{-3}\gamma^2 + O(n^{-3/2})$$

$$E(S_n^3) = n^{3/2}\tau^{-3/2}E\left((W_1 - W_2)^3(1 - \frac{3}{2}\tau^{-1}\sum_{i=1}^{2}(q_i - p_i)W_i)\right) + O(n^{-1})$$

$$= -\frac{7}{2}n^{-1/2}\tau^{-3/2}\gamma + O(n^{-1})$$

and

$$E(S_n^4) = n^2\tau^{-2}E\{(W_1 - W_2)^4(1 - 2\tau^{-1}\sum_{i=1}^{2}(q_i - p_i)W_i$$

$$+ \tau^{-1}\sum_{i=1}^{2}(2 + 3\tau^{-1}(q_i - p_i)^2)W_i^2$$

$$+ 6\tau^{-2}(q_1 - p_1)(q_2 - p_2)W_1W_2)\} + O(n^{-3/2})$$

$$= 3 + 6n^{-1} + 18n^{-1}\tau^{-2}(p_1^2q_1^2 + p_2^2q_2^2) - 2n^{-1}\tau^{-2}\kappa + 28\tau^{-3}\gamma^2 + O(n^{-3/2})$$

Hence, by equations (2.4.6)

$$\kappa_{1,n} = E(S_n) = -\frac{1}{2}n^{-1/2}\tau^{-3/2}\gamma + O(n^{-1})$$

$$\kappa_{2,n} = E(S_n^2) - (E(S_n))^2$$

$$= 1 + n^{-1} + 2n^{-1}\tau^{-2}(p_1^2q_1^2 + p_2^2q_2^2) + \frac{7}{4}n^{-1}\tau^{-3}\gamma^2 + O(n^{-3/2})$$

$$\kappa_{3,n} = E(S_n^3) - 3E(S_n^2)E(S_n) + 2(E(S_n))^3$$

$$= -2n^{-1/2}\tau^{-3/2}\gamma + O(n^{-1})$$

and

$$\kappa_{4,n} = E(S_n^4) - 4E(S_n^3)E(S_n) - 3(E(S_n^2)^2 + 12E(S_n^2)(E(S_n))^2 - 6(E(S_n))^4$$

$$= 6n^{-1}\tau^{-2}(p_1^2q_1^2 + p_2^2q_2^2) - 2n^{-1}\tau^{-2}\kappa + 12n^{-1}\tau^{-3}\gamma^2 + O(n^{-3/2})$$

Therefore, in the notation of (2.4.5), the two-term Edgeworth expansion for $T_{n,n}$ is

$$P(T_{n,n} < a) = \Phi(a) + n^{-1/2}r_1(a)\phi(a) + n^{-1}r_2(a)\phi(a) + O(n^{-3/2})$$

where $r_1(a)$ and $r_2(a)$ are given in equation (2.4.3) equation (2.4.4) with

$$k_{1,2} = -\frac{1}{2}\tau^{-3/2}\gamma$$

$$k_{2,2} = 1 + 2\tau^{-2}(p_1^2q_1^2 + p_2^2q_2^2) + \frac{7}{4}\tau^{-3}\gamma^2$$

$$k_{3,1} = -2\tau^{-3/2}\gamma$$

$$k_{4,1} = 6\tau^{-2}(p_1^2q_1^2 + p_2^2q_2^2) - 2\tau^{-2}\kappa + 12\tau^{-3}\gamma^2$$

in which $\tau = p_1q_1 + p_2q_2$.

Step 4, compute the smooth approximation of the coverage probability of the Wald interval.

By equation (2.4.12),

$$
\begin{aligned}
P_W \leq & P\left(T_{n,n} \leq (Z_{\alpha/2}+n^{-2})/(1-2/n^3)\right) - P\left(T_{n,n} < -(Z_{\alpha/2}+n^{-2})/(1+2/n^3)\right) \\
= & \Phi((Z_{\alpha/2}+n^{-2})/(1-2/n^3)) - \Phi(-(Z_{\alpha/2}+n^{-2})/(1+2/n^3)) \\
& + n^{-1/2} r_1((Z_{\alpha/2}+n^{-2})/(1-2/n^3))\phi((Z_{\alpha/2}+n^{-2})/(1-2/n^3)) \\
& - n^{-1/2} r_1(-(Z_{\alpha/2}+n^{-2})/(1+2/n^3))\phi(-(Z_{\alpha/2}+n^{-2})/(1+2/n^3)) \\
& + n^{-1} r_2((Z_{\alpha/2}+n^{-2})/(1-2/n^3))\phi((Z_{\alpha/2}+n^{-2})/(1-2/n^3)) \\
& - n^{-1} r_2(-(Z_{\alpha/2}+n^{-2})/(1+2/n^3))\phi(-(Z_{\alpha/2}+n^{-2})/(1+2/n^3)) \\
& + O(n^{-3/2}) \\
= & (1-\alpha) + 2n^{-1} r_2(Z_{\alpha/2})\phi(Z_{\alpha/2}) + O(n^{-3/2}) \qquad (2.4.19)
\end{aligned}
$$

The cancellation is valid because all functions appeared in the two-term Edgeworth expansion of S_n are continuous and the n^{-2} terms can be absorbed in $O(n^{-3/2})$. That $r_1(x)$ and $\phi(x)$ are even functions and $r_W(x)$ is an odd function also guarantees the last two steps.

Similarly, it can be shown that

$$P_W \geq (1-\alpha) + 2n^{-1} r_2(Z_{\alpha/2})\phi(Z_{\alpha/2}) + O(n^{-3/2}). \qquad (2.4.20)$$

Combine inequalities (2.4.19) and (2.4.20), then we have the **smoothing formula** for the coverage probability of the Wald interval:

The coverage probability of the Wald interval is at most $1-(p_1^n+q_1^n)(p_2^n+q_2^n)$ *and can be expanded as*

$$P_W = (1-\alpha) + 2n^{-1} r_2(Z_{\alpha/2})\phi(Z_{\alpha/2}) + O(n^{-3/2}) \qquad (2.4.21)$$

where $r_2(Z_{\alpha/2}) = r_2(Z_{\alpha/2})$ *in equation (2.4.4) with*

$$k_{1,2} = -\frac{1}{2}\tau^{-3/2}\gamma$$

$$k_{2,2} = 1 + 2\tau^{-2}(p_1^2q_1^2 + p_2^2q_2^2) + \frac{7}{4}\tau^{-3}\gamma^2$$

$$k_{3,1} = -2\tau^{-3/2}\gamma$$

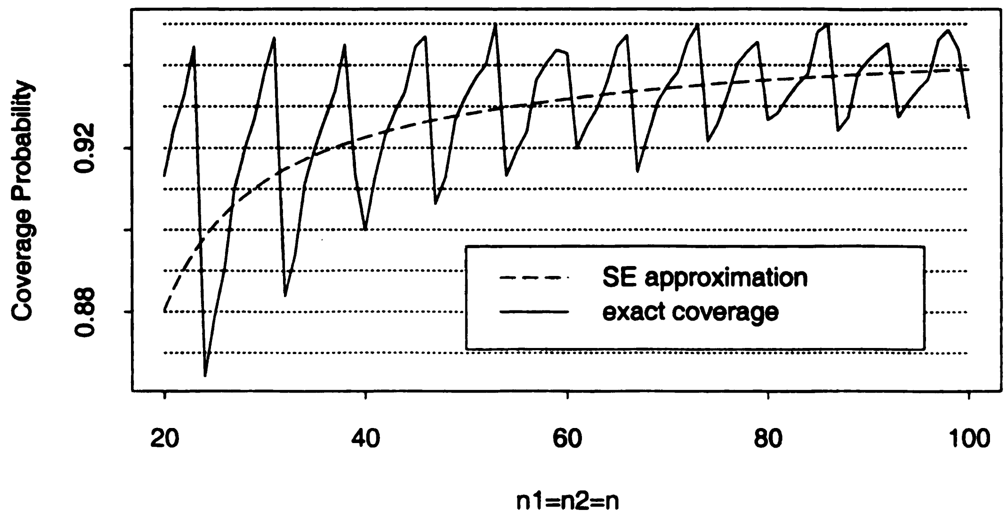
$$k_{4,1} = 6\tau^{-2}(p_1^2q_1^2 + p_2^2q_2^2) - 2\tau^{-2}\kappa + 12\tau^{-3}\gamma^2$$

in which $\tau = p_1q_1 + p_2q_2$ *when* $0 < \alpha < 1$ *and* $n_1 = n_2 = n \geq 2$.

Remark 2.4.1. Neglecting the error term $O(n^{-3/2})$, one can see the approximate coverage probability given by the smoothing formula is a smooth function with respect to sample size n and the proportions, thus it does not inherit the oscillation of the exact coverage probability. Because we used random variables with absolutely continuous distribution functions instead of the discrete binomial random variables in deriving the expansion, the oscillation is likely caused by the discreteness of Bernoulli distribution.

Remark 2.4.2. Compare the approximate coverage probabilities from the expansion and the exact coverage probabilities, for instance, see Figure 4 and 5, in which the nominal levels are 95% and the SE stands for the smooth expansion. We notice that the coverage probabilities approximated by the smoothing formula follows the trend of the exact coverage regardless of the oscillation. The approximate coverage given by the smooth formula is lower than the nominal level. Our further study shows that at 95% nominal level, the $r_W(Z_{\alpha/2})$ term is always negative unless both proportions are either less than 0.028 or greater than 0.972. When the

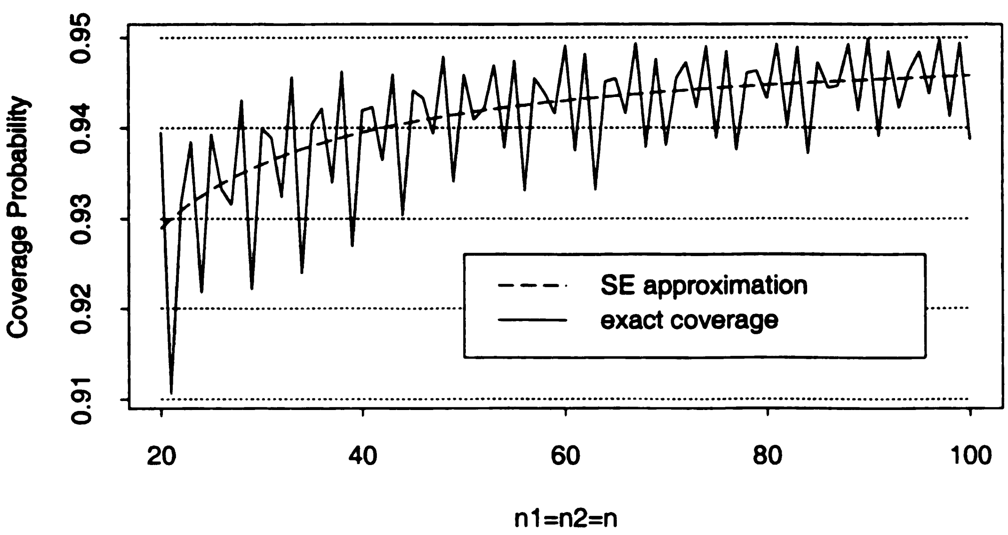
Figure 2.4: Exact and SE approximate coverage probabilities of the nominal 95% Wald intervals for $p_1 = 0.9$, $p_2 = 0.1$ and $n_1 = n_2 = n$



nominal level is 90%, if $0.001 \le p_i \le 0.999$ for $i = 1$ or $i = 2$, then $r_W(Z_{\alpha/2}) < 0$. When the two proportions take some extreme values, the $(p_1^{n_1} + q_1^{n_1})(p_2^{n_2} + q_2^{n_2})$ term may achieve non-negligible amount even for quite large sample sizes, say $0.999^{200} = 0.819$, $0.99^{200} = 0.134$. Therefore, formula 2.4.21 explains why the Wald interval has typically lower exact coverage probabilities than the nominal level and the formula gives the order of the negative deviation of the coverage probability of the Wald interval from the nominal level.

Remark 2.4.3. The smooth approximation of the exact coverage probability of the Wald interval at different sample sizes is still valid when $n_i = \pi_i n$ for $i = 1, 2$, where π_1 and π_2 are two relatively prime positive integers and n is an positive integer. Figure 6 plots the exact coverage probabilities and their approximations

Figure 2.5: Exact and SE approximate coverage probabilities of the nominal 95% Wald intervals for $p_1 = 0.8$, $p_2 = 0.3$ and $n_1 = n_2 = n$

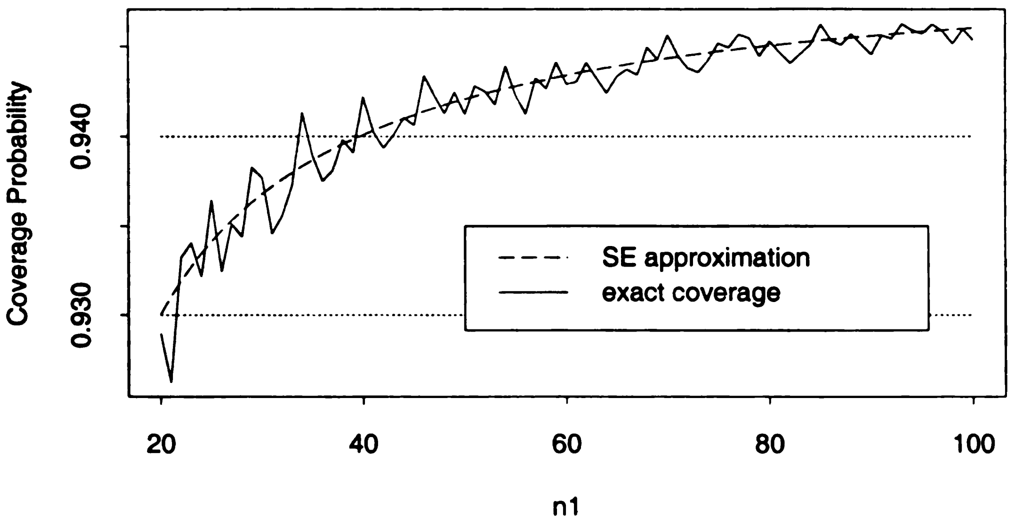


given by the smoothing formula when $p_1 = 0.9$, $p_2 = 0.8$, $n_2 = 2n_1$ and n_1 varies from 20 to 100 at 95% level. The smoothing formula approximates well in this case too.

Remark 2.4.4. For any discrete random variables that have finitely many possible values, a corresponding formula can be derived through the method applied above. For other discrete random variables, the constant 4 in defining ξ and η may take some other value.

According to our analysis, we conclude that the Wald interval behaves much worse than people's expectation and should be used with caution. Alternative intervals will be evaluated in next Chapter.

Figure 2.6: Exact and SE approximate coverage probabilities of the nominal 95% Wald intervals for $p_1 = 0.8$, $p_2 = 0.3$ and $n_2 = 2n_1$



Chapter 3

Interval Estimation for the Difference of two Binomial Proportions

3.1 Introduction

The poor performance of the Wald interval for the difference of two binomial proportions has been addressed in last chapter. Consequently, there are quite a lot of methods of developing alternative intervals suggested. Their performances differ significantly.

In Section 2, we present several interval estimation methods as candidates to replace the Wald interval, each with its motivation. The candidate intervals are classified into three groups: (1) The Wald interval with continuity correction. It has the same center as the original Wald interval. (2) Confidence intervals with adjusted centers. We select two of such intervals. One is derived through Bayesian approach with *Beta* prior distributions and then using normal approximation. we identify it the (approximate) Bayes interval in the report. Another one is proposed by Agresti and Coull (1998). The main idea is to add four pseudo observations. Both intervals have different centers from the Wald interval. (3) The profile likelihood based confidence interval. Which, unlike the other intervals, does not have an explicit form.

In Section 3, the performances of the above intervals with explicit forms along with the Wald interval are evaluated. We assess those intervals on two aspects. One is their coverage probabilities. All the coverage probabilities in this section are computed exactly rather than by simulation. The other is their expected lengths.

The profile likelihood based interval is taken into consideration in Section 4. We compare the coverage probabilities and lengths of all the alternative intervals along with the Wald interval through simulation.

According to our analysis, we recommend the intervals with adjusted centers as substitutes for the Wald interval.

We concentrate on intervals with 95% nominal level in this chapter. The conclusions also hold for intervals with other nominal levels.

3.2 Some Alternative Intervals

The following are some alternatives to the Wald interval.

3.2.1 The Wald interval with continuity correction

There are a few intervals with different correction terms in this category. The most widely used one is

$$\hat{p}_1 - \hat{p}_2 \pm \left(z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} + \frac{1}{2n_1} + \frac{1}{2n_2} \right) \tag{3.2.1}$$

It results from inverting the Wald test: when computing the p-value, a continuity correction is applied for improving the accuracy of the central limit theorem approximation. This interval has the same center as the Wald interval and a greater margin of error.

3.2.2 Intervals with adjusted center

Approximate Bayes interval

The method is motivated by using the Bayesian estimates instead of the maximum likelihood estimates when deriving confidence intervals. For $i = 1, 2$, the independent conjugate $Beta(a, b)$ prior distribution results in the posterior distribution of p_i is $Beta(a + X_i, b + n_i - X_i)$ with mean $\tilde{p}_i = (X_i + a)/(n_i + a + b)$ and variance $\tilde{p}_i(1 - \tilde{p}_i)/(n_i + a + b + 1)$. Using a normal approximation for the distribution of the difference of the posterior beta variate leads to the approximate Bayes interval

$$\tilde{p}_1 - \tilde{p}_2 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1 + a + b + 1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2 + a + b + 1}}$$

In particular, if $a = b$, the estimators $\tilde{p}_1$ and $\tilde{p}_2$ are driven to be closer to 1/2 than $\hat{p}_1$ and $\hat{p}_2$ respectively unless $p_i = 1/2$. Suggested by Berry (1996) (p.291), the approximate Bayes interval in the report is specified to take $a = b = 1$, which leads to

$$\tilde{p}_1 - \tilde{p}_2 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+3} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+3}} \tag{3.2.2}$$

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

The Agresti-Coull interval

As mentioned in Chapter 1, motivated by the Wilson interval for one binomial proportion as shown in Wilson (1927), Agresti and Coull (1998) suggested an interval with $Z_{0.025}^2 \approx 4$ pseudo observations, one success and one failure from each binomial population. Then the sample proportions are $\tilde{p}_i = (X_i + 1)/(n_i + 2)$ for $i = 1, 2$. Replacing $\hat{p}_i$ with $\tilde{p}_i$ and n_i with $n_i + 2$ for $i = 1, 2$ in the ordinary Wald interval yields the Agresti-Coull interval

$$CI_A = \tilde{p}_1 - \tilde{p}_2 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}} \tag{3.2.3}$$

The above two intervals have the same center. The approximate Bayes interval is a subset of the Agresti-Coull interval. Since they have a different center from the Wald interval with or without continuity correction but a similar form, we call them intervals with adjusted centers.

3.2.3 The profile likelihood based intervals

Unlike the other intervals discussed so far, profile likelihood based intervals do not have explicit forms. Suppose the log-likelihood function of $\theta = (\Delta, p_2)$ is $l(\Delta, p_2)$, where $\Delta = p_1 - p_2$ is the parameter of interest and p_2 is regarded as a nuisance parameter. Let $\tilde{l}(\Delta) = \max_{p_2} l(\Delta, p_2)$, which is called the profile likelihood for Δ, where the range of p_2 for the maximization is $(0, 1-\Delta)$ if $\Delta \geq 0$ and $(-\Delta, 1)$ otherwise. Then an approximate $100(1-\alpha)\%$ profile likelihood interval for $p_1 - p_2$ is

$$\{\Delta \in (-1,1) : 2(l(\hat{\Delta}, \hat{p}_2) - \tilde{l}(\Delta)) \leq \chi_1^2(\alpha)\} \tag{3.2.4}$$

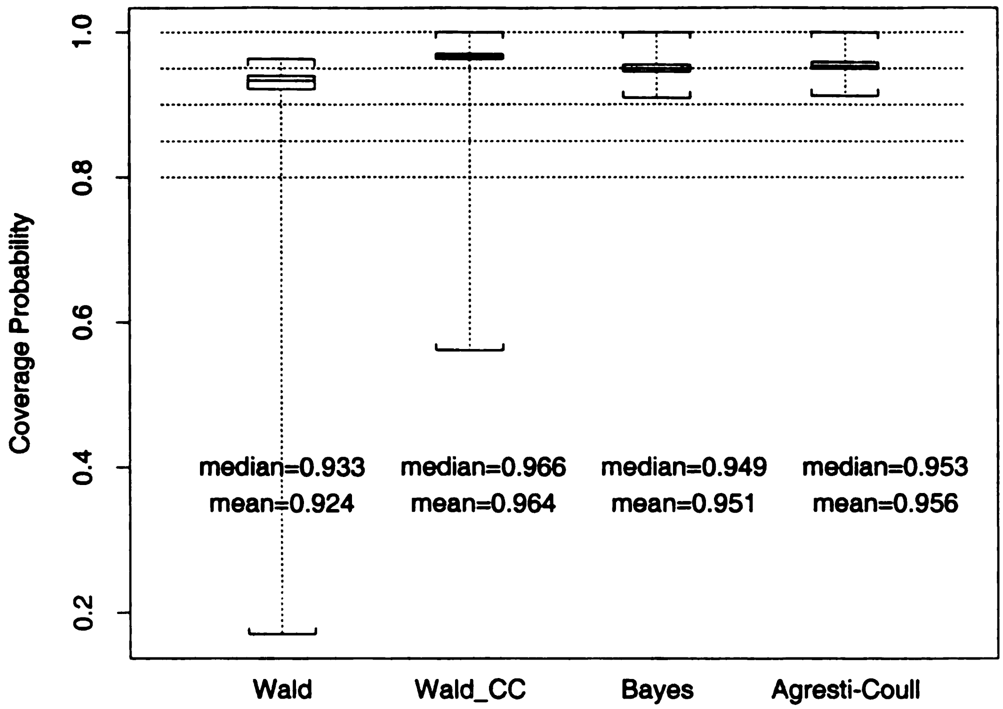
where $\hat{\Delta} = \hat{p}_1 - \hat{p}_2$ and $\chi_1^2(\alpha)$ is the 100α upper percentage point of χ_1^2. This interval follows from the fact that for $\Delta = \Delta_0$, $2(l(\hat{\Delta}, \hat{p}_2) - \tilde{l}(\Delta_0))$ is asymptotically chi-squared distributed with 1 degree of freedom as shown in Cox and Hinkley (1974)

.

3.3 Comparison of Intervals with Explicit Forms

We address the comparisons on two aspects: coverage properties and expected lengths. For convenience, we define some unified notations: CI_* represents interval $*$, CP_* and EL_* refer to its coverage probability and expected length respectively, where $*$ may be W, WCC, B, AC and PLB that indicate the Wald interval, the Wald interval with continuity correction, the approximate Bayes interval, the Agresti-Coull interval and the profile likelihood based interval respectively.

Figure 3.1: Exact coverage probability Boxplots of some 95% nominal intervals



3.3.1 Coverage Properties

Since the Wald interval with continuity correction contains the Wald interval, its coverage probability is always no less than that of the Wald interval. Similarly, the coverage probability of approximate Bayes interval cannot exceed that of the Agresti-Coull interval.

To explore the average performance of the four intervals for small to moderate sample sizes, we randomly sampled 10,000 values of $(n_1, p_1; n_2, p_2)$, taking p_1 and p_2 independently from $U(0,1)$ and taking n_1 and n_2 independently from the uniform distribution over $\{10, 11, \ldots, 50\}$. We evaluated the exact coverage probabilities of the four intervals for each realization of $(n_1, p_1; n_2, p_2)$ at the 95% nominal level.

Figure 3.1 shows the coverage probability boxplots of the four intervals. The means and medians of the coverage probabilities of the four intervals are listed in the figure as well. The performance of the Wald interval is very poor. Both the mean and the median of its coverage probabilities are much lower than the nominal level and those of other intervals. Its behavior is not stable either. It suffers from occasionally very low coverage probabilities. For example, the minimum coverage probability of the Wald interval from our evaluation is only 0.170.

Contrary to the Wald interval, the coverage probabilities of the Wald interval with continuity correction tends to be much higher than the nominal level. It has 81.4% of its coverage probabilities greater than 0.96. Since it is just a simple spread of the Wald interval, it inherits some disadvantages of the Wald interval such as unstable and occasionally very low coverage probabilities.

Compared with the Wald interval and its variate, coverage probabilities of approximate Bayes interval and Agresti-Coull interval lie more closely to the nominal level, which makes them more reliable. This can also be seen from the mean distances of the coverage probabilities of the four intervals from the nominal level, which are 0.026,0.019,0.07 and 0.07 for the Wald, Wald with continuity correction, Bayes and Agresti-Coull intervals respectively.

According to the analysis in Chapter 2, when both p_1 and p_2 are getting closer to 1/2, EW_{n_1,n_2}, the bias of the test statistic W_{n_1,n_2} from the mean of standard normal distribution tends to decrease, where

$$W_{n_1,n_2} = \frac{\hat{p_1} - \hat{p_2} - (p_1 - p_2)}{\sqrt{\hat{p_1}(1-\hat{p_1})/n_1 + \hat{p_2}(1-\hat{p_2})/n_2}},$$

through which the Wald interval can be derived. This explains why the coverage behaviors of the intervals with adjusted centers are better than that of the Wald interval: $\tilde{p_i}$ is closer to $1/2$ than $\hat{p_i}$ unless $\hat{p_i} = 1/2$.

Average performance over the unit square for (p_1, p_2) and $\{10, 11, \ldots, 50\} \times \{10, 11, \ldots, 50\}$ for (n_1, n_2) can mask certain behaviors of the four intervals in some regions. In particular, some pairs of (p_1, p_2) are of more interest, say $|p_1 - p_2|$ small. Similarly, some pairs of (n_1, n_2) may be more important, for example, proportional sampling may result in some favorable (n_1, n_2). Hence, it is necessary to focus on some special and common cases.

Figure 3.2 shows how the coverage probabilities of the four intervals vary for $p_2 = 0.5$ and p_1 varying between 0.01 and 0.99 with step size 0.01 at $n_1 = n_2 = 20$ and nominal level 95%. In this case, the coverage probabilities of the Wald interval never achieve the nominal level. Its mean coverage is 0.933. The coverage of the Wald interval with continuity correction is always above 0.96 with mean 0.969. Most coverage probabilities of the Bayes interval and the Agresti-Coull interval fluctuates between 0.94 and 0.96. The mean coverage probabilities are 0.946 and 0.951 respectively.

Figure 3.3 demonstrates the behaviors of the four intervals when $p_1 = p_2$ vary between 0.01 and 0.99 with step size 0.01 at $n_1 = n_2 = 20$ and nominal level 95%. Again, the coverage probability of the Wald interval never reaches the nominal

Figure 3.2: Comparison of exact coverage probabilities for $p_2 = 0.5$, $n_1 = n_2 = 20$ at 95% nominal level

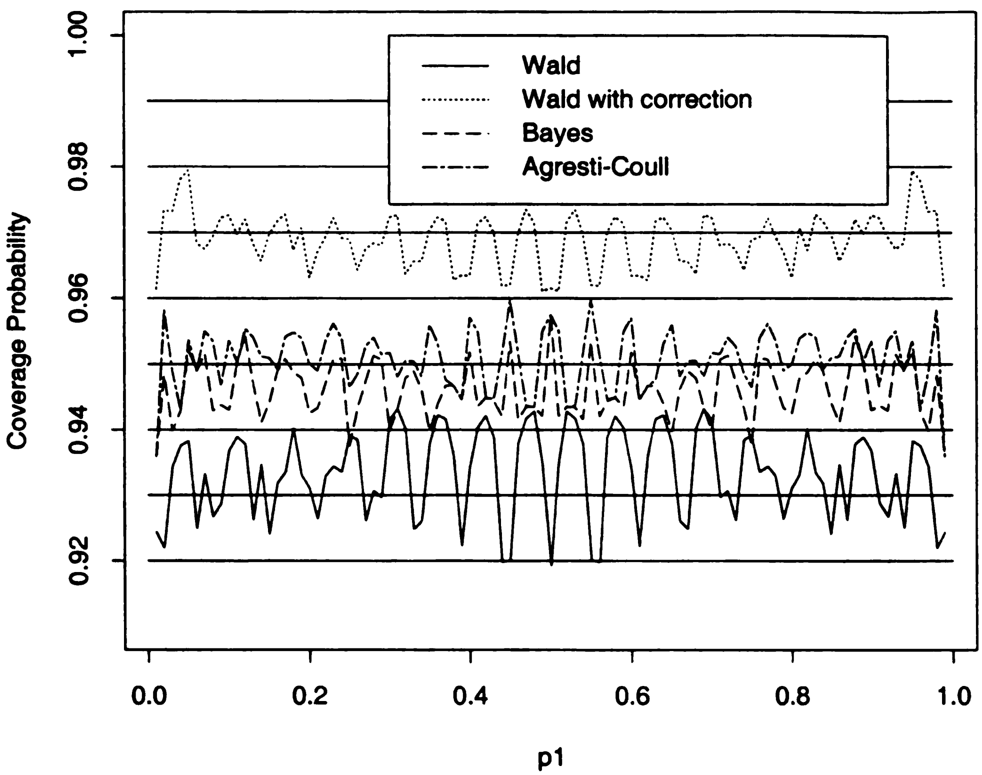


Figure 3.3: Comparison of exact coverage probabilities for $p_1 = p_2 = 0.01$ to $0.99, n_1 = n_2 = 20$ at 95% nominal level

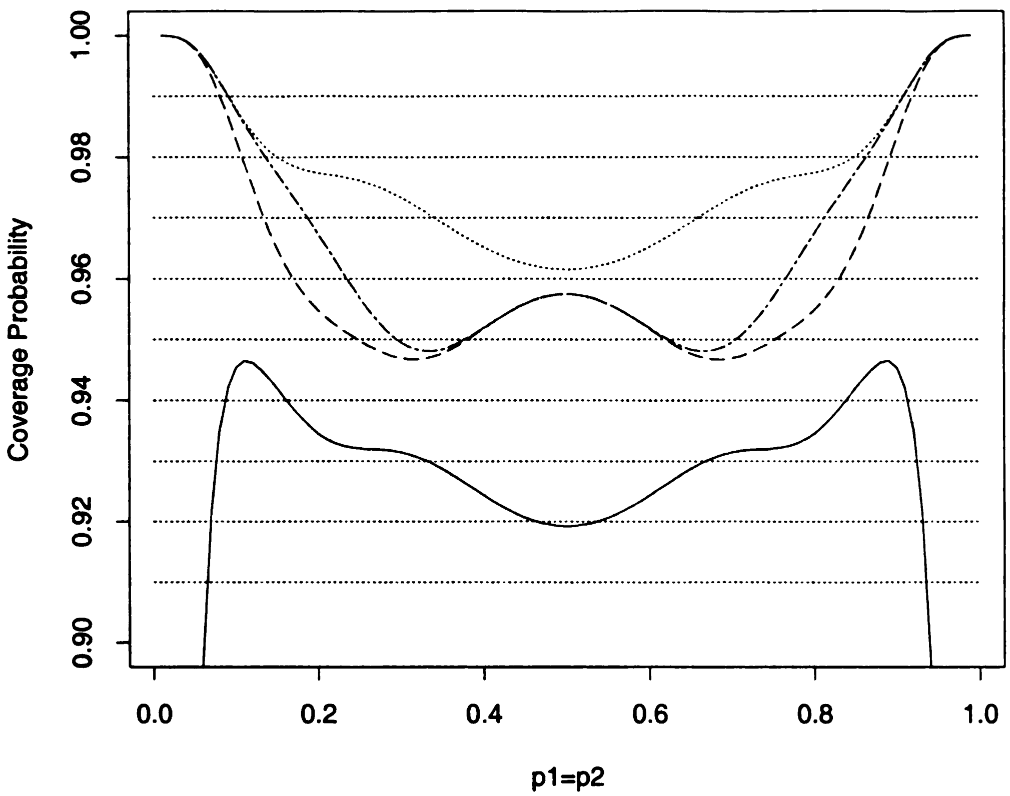
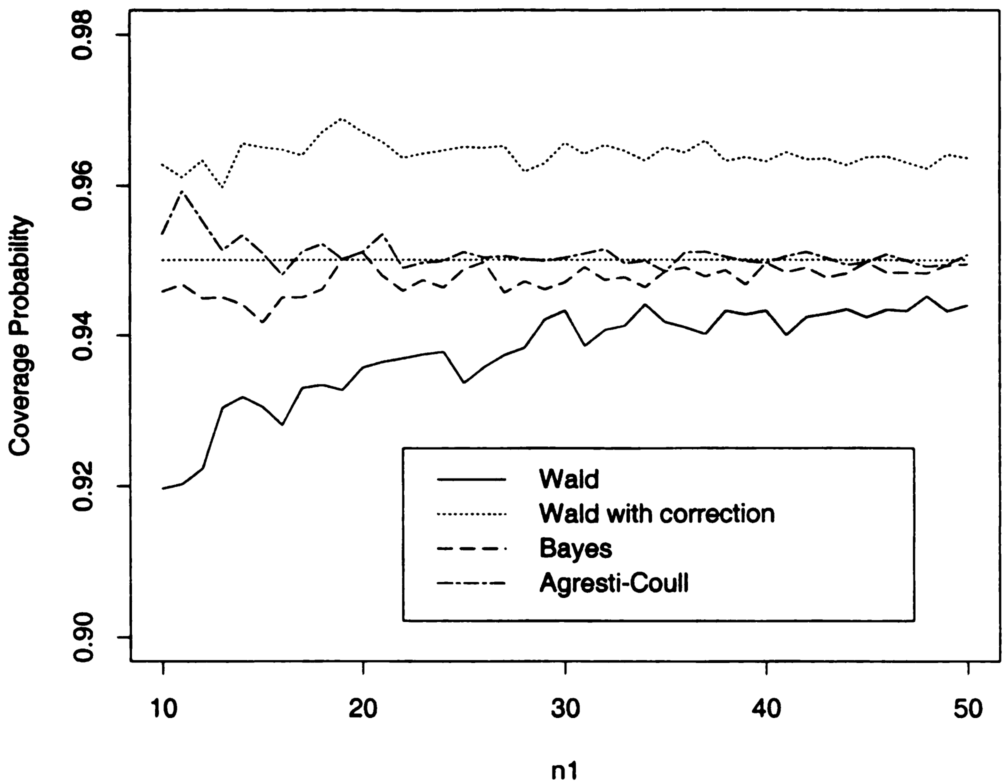


Figure 3.4: Comparison of exact coverage probabilities at $p_1 = 0.7, p_2 = 0.5$, $n_2 = 2n_1$ at 95% nominal level



level and the coverage of its variate remains above 0.96. It is striking to see that all the coverage probabilities of the three alternative intervals converge to 1 and the coverage of the Wald interval drops dramatically as $p_1 = p_2$ approach either boundary. This is because the Wald interval suffers from shrinkage to an empty set for some realizations of $(n_1, p_1; n_2, p_2)$ while the other intervals do not. When both sample sizes approach boundaries and sample sizes are small or moderate, the chance that the Wald interval is empty might be non-negligible or quite large.

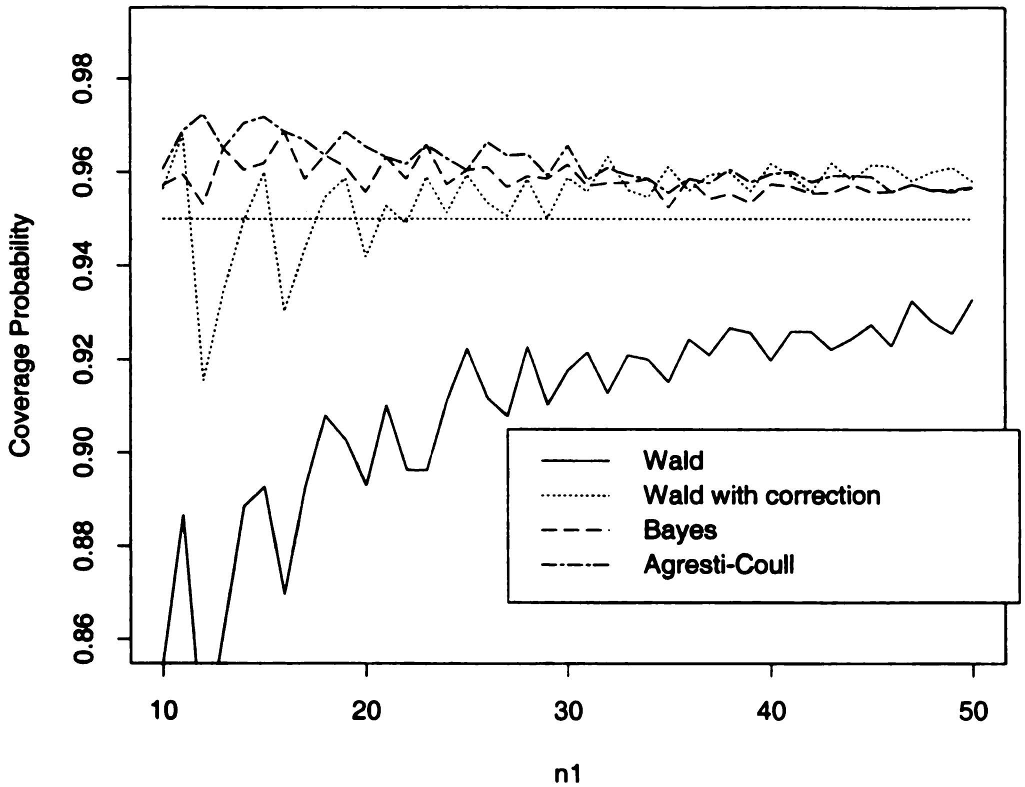
The coverage performances for $p_1 = 0.7$, $p_2 = 0.5$ and $n_2 = 2n_1$ are plotted in figure 3.4. When sample sizes are small, the coverage probabilities of the Wald interval are much lower than the nominal level while the other three intervals have coverage probabilities much closer to the nominal level. As sample sizes increase, all the coverage probabilities are driven to the nominal level. Though the exact coverage probabilities of the Wald interval with continuity correction and the intervals with adjusted centers are much higher than the 95% nominal level, this does not persist for all p_1 and p_2 when $n_2 = 2n_1$. For example, if $p_1 = 0.9$, $p_2 = 0.1$, $n_1 = 12, n_2 = 24$, the coverage probability of the Wald interval with continuity correction is only 0.915. More coverage probabilities for $p_1 = 0.9$, $p_2 = 0.1$ are plotted in figure 3.5.

The discussion so far shows that the coverage of Wald interval is too low and the coverage of Wald interval with continuity correction is often too high. The intervals with adjusted centers have coverage probabilities around the nominal level.

3.3.2 Expected Lengths

In addition to coverage probability, parsimony in length is another important issue for evaluating a confidence interval. We have shown that the coverage probabilities of intervals with adjusted centers are much higher than the Wald interval in a frequentist sense, but the gain on coverage probability is not due to greater lengths. On the contrary, the intervals with adjusted centers often have smaller lengthes than the Wald interval. But for the Wald interval with continuity correction, the improvement of coverage probability is completely through widening the Wald in-

Figure 3.5: Comparison of exact coverage probabilities at $p_1 = 0.9$, $p_2 = 0.8$, $n_2 = 2n_1$ at 95% nominal level



terval.

Theorem 3.3.1. *Denote* $\frac{1}{2z_{\frac{\alpha}{2}}}$ *by c. Then*

$$cEL_W = u^{1/2} - \frac{u^{-1/2}}{8}\left(\frac{4(p_1q_1 + r^2p_2q_2)}{n^2} + \frac{p_1q_1(q_1-p_1)^2 + r^3p_2q_2(q_2-p_2)^2}{n^2(p_1q_1 + rp_2q_2)}\right)$$

$$+O(n^{-2}) \tag{3.3.1}$$

$$cEL_{WCC} = cEL_W + \frac{1+r}{2n} + O(n^{-2}) \tag{3.3.2}$$

$$cEL_B = cEL_W + u^{1/2}\frac{1 + r^2 - 7(p_1q_1 + r^2p_2q_2)}{2n(p_1q_1 + rp_2q_2)} + O(n^{-2}) \tag{3.3.3}$$

$$cEL_{AC} = cEL_W + u^{1/2}\frac{1 + r^2 - 6(p_1q_1 + r^2p_2q_2)}{2n(p_1q_1 + rp_2q_2)} + O(n^{-2}) \tag{3.3.4}$$

where $r = n_1/n_2$, $n = n_1$ *and*

$$u = \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2} = \frac{1}{n}(p_1q_1 + rp_2q_2)$$

Proof. Define $\omega_i = (\hat{p}_i - p_i)/p_i$, then $\hat{p}_i = p_i\omega_i + p_i$. The length of the Wald interval, denoted by L_W, is

$$\begin{aligned} L_W =& 2z_{\frac{\alpha}{2}}\left(\sum_{i=1}^{2}\frac{\hat{p}_i(1-\hat{p}_i)}{n_i}\right)^{\frac{1}{2}} \\ =& 2z_{\frac{\alpha}{2}}[u^{\frac{1}{2}} - \frac{1}{2}u^{-\frac{1}{2}}\sum_{i=1}^{2}\left(\frac{1}{4}u^{-1}\frac{p_i^2(q_i-p_i)^2}{n_i^2} + \frac{p_i^2}{n_i}\right)\omega_i^2 \\ &+ f_W(\omega_1,\omega_2)] + O_p(n^{-2}) \end{aligned} \tag{3.3.5}$$

where $f_W(\omega_1, \omega_2)$ only contains terms of ω_i and $\omega_1\omega_2$ and has mean 0. The second step is achieved by multivariate Taylor Expansion and $\omega_i = O_p(n^{-\frac{1}{2}})$. Equation (3.3.1) follows from taking expected value of equation (3.3.5) with respect to ω_1 and ω_2.

Since the length of Wald interval with continuity correction is

$$L_{WCC} = L_W + 2z_{\frac{\alpha}{2}}(1/2n_1 + 1/2n_2),$$

some elementary algebra results in (3.3.2).

The length, L_B, of approximate Bayes interval, is

$$\begin{aligned}
L_B =& 2z_{\frac{\alpha}{2}} \left(\sum_{i=1}^{2} \frac{\tilde{p}_i(1-\tilde{p}_i)}{n_i+3} \right)^{\frac{1}{2}} \\
=& 2z_{\frac{\alpha}{2}} (t^{\frac{1}{2}} - \frac{1}{2} t^{-\frac{1}{2}} \sum_{i=1}^{2} \left(\frac{1}{4} t^{-1} \frac{n_i^4 p_i^2 (q_i - p_i)^2}{(n_i+2)^4 (n_i+3)^2} + \frac{n_i^2 p_i^2}{(n_i+2)^2 (n_i+3)} \right) \omega_i^2 \\
& + f_B(\omega_1, \omega_2)) + O_p(n^{-2}) \qquad (3.3.6)
\end{aligned}$$

where $f_B(\omega_1, \omega_2)$ only contains terms of ω_i and $\omega_1\omega_2$ and has mean 0. The second step is again achieved by multivariate Taylor Expansion and $\omega_i = O_p(n^{-\frac{1}{2}})$. The t has the expression

$$\begin{aligned}
t =& \frac{\tilde{p}_1(1-\tilde{p}_2)}{n_1+3} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+3} \\
=& \frac{p_1 q_1}{n_1} + \frac{1 - 7p_1 q_1}{n_1^2} + \frac{p_2 q_2}{n_2} + \frac{1 - 7p_2 q_2}{n_2^2} + O(n^{-3}) \\
=& u \left(1 + \frac{1 + r^2 - 7(p_1 q_1 + r^2 p_2 q_2)}{n(p_1 q_1 + r p_2 q_2)} \right) + O(n^{-3}) \qquad (3.3.7)
\end{aligned}$$

Therefore, replacing the t with the expression given by (3.3.7) and taking expectation of equation (3.3.6) with respect to ω_1 and ω_2 gives the desired result (3.3.3).

The proof of (3.3.4) is very similar to the proof for the approximate Bayes interval and is omitted. □

A direct conclusion of Theorem 3.3.1 is the comparison results of the expected lengths of the intervals.

Corollary 3.3.1. *Denote $r = n_1/n_2$ and $n_1 = n$, then up to an error of $O(n^{-2})$, $EL_W \geq EL_B$ if and only if*

$$7(p_1q_1 + r^2p_2q_2) \geq 1 + r^2$$

and $EL_{WCC} \geq EL_B$ if and only if

$$1 + r \geq \frac{1 + r^2 - 7(p_1q_1 + r^2p_2q_2)}{\sqrt{n(p_1q_1 + rp_2q_2)}}.$$

Remark 3.3.1. The above corollary can be applied to the Agresti-Coull interval after replacing the 7's by 6's.

Remark 3.3.2. Based on the corollary, when both p_1 and p_2 are in $(0.173, 0.827)$, the Bayes interval is shorter than the Wald interval and it is the shortest among the four intervals if the $O(n^{-2})$ error is neglected. In addition, if $p_i \in \left(\frac{1}{2} - \frac{1}{2}\sqrt{\frac{n_i-4}{n_i}}\right)$ is satisfied for either $i = 1$ or $i = 2$, the Bayes interval is shorter than the Wald interval with continuity correction if the $O(n^{-2})$ error is neglected. Therefore, when one sample size is not too small and the corresponding proportion is not too close to the boundaries, the approximate Bayes interval is shorter than the Wald interval with continuity correction.

Remark 3.3.3. The Wald interval with continuity correction is often much longer than the other three intervals. Figure 3.6 and 3.7 plot the approximate expected lengths of the four intervals under different conditions when nominal confidence level is 95%. They demonstrate that the expected lengths of those intervals except the Wald interval with continuity correction are comparable.

Figure 3.6: Comparison of approximate expected lengths of some confidence intervals for $p_2 = 0.5$ and $n_1 = n_2 = 25$

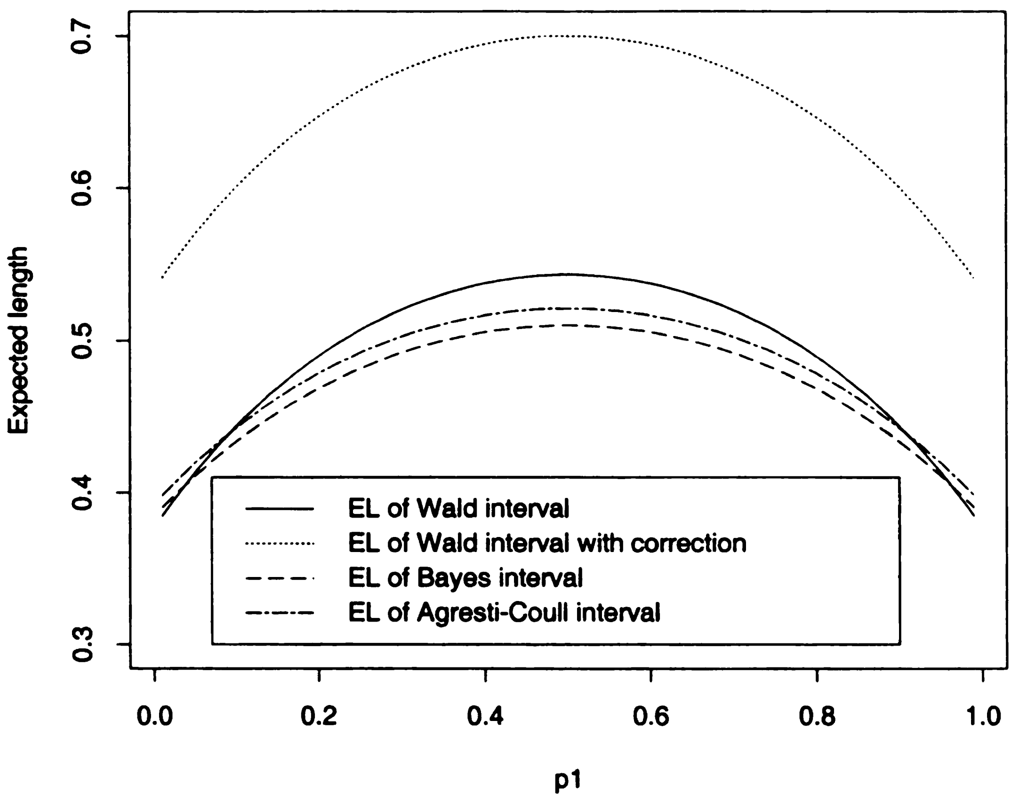
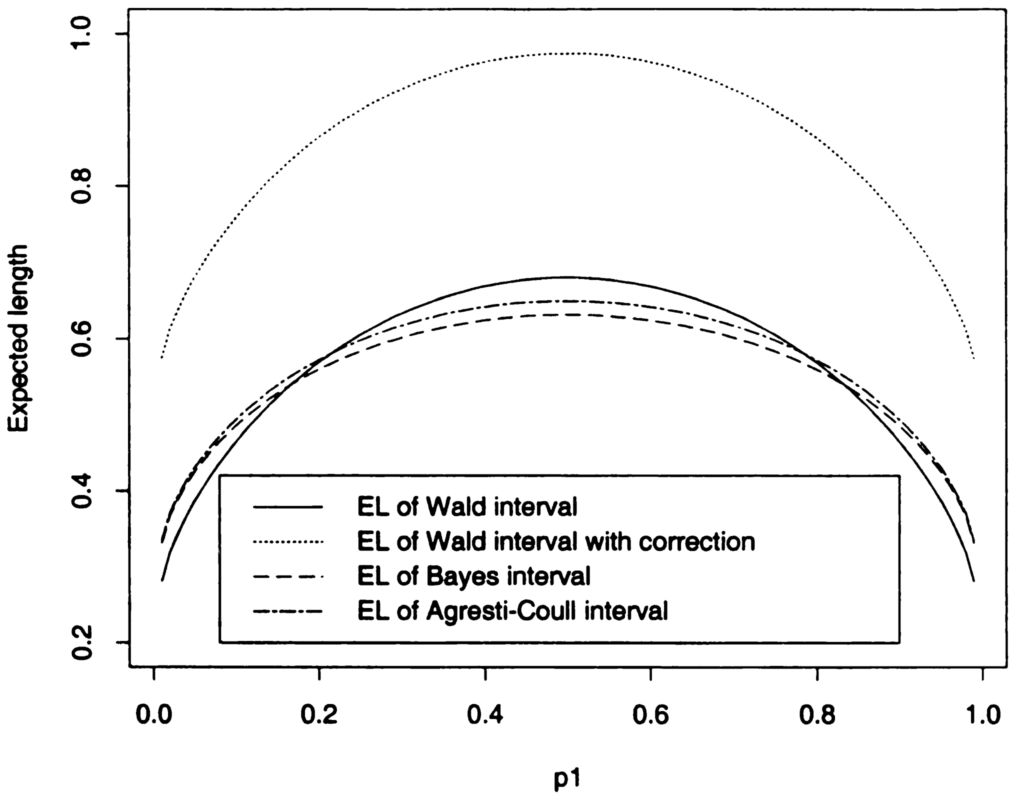


Figure 3.7: Comparison of approximate expected lengths of some confidence intervals for $p_2 = 0.1$ and $n_1 = 10$, $n_2 = 20$



Now we may conclude that the poor coverage performance of the Wald interval is not because it is short. On the contrary, compared to intervals with adjusted centers, the Wald interval is often longer than them but with less coverage probability. The high coverage probability of Wald interval with continuity correction is achieved by widening the Wald interval dramatically. Hence, in replacing the Wald interval, intervals with adjusted centers are much more preferable.

3.4 Comparison of the Wald Interval and all Proposed Alternatives

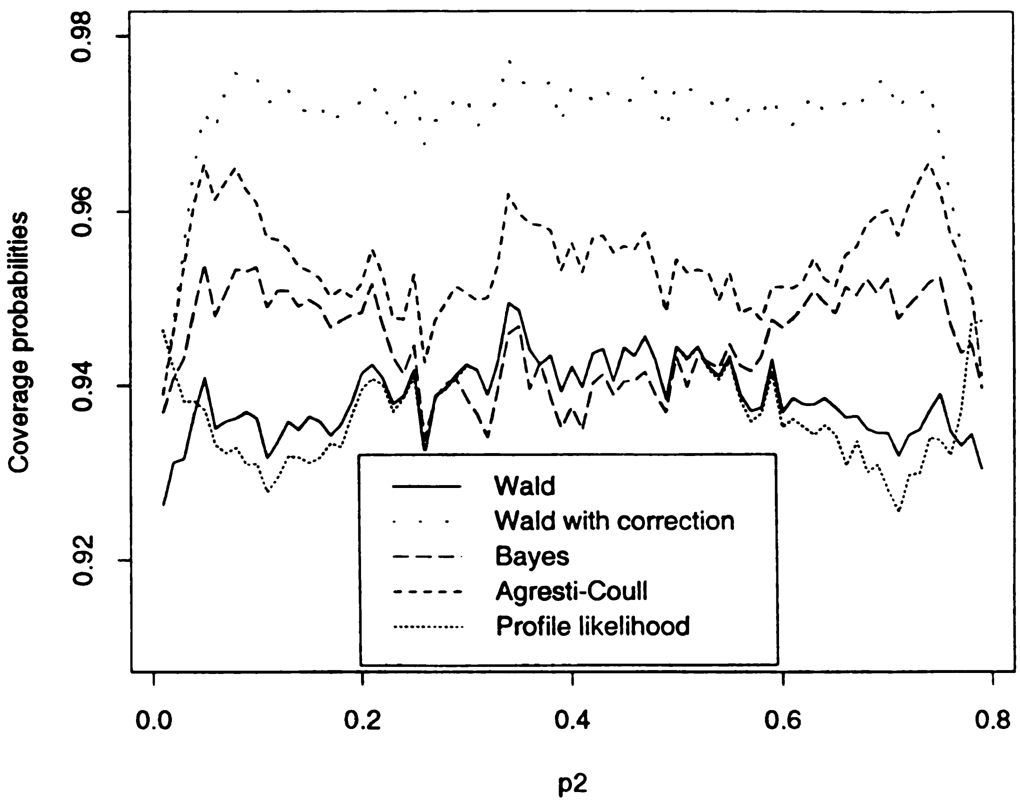
The comparison is based on a simulation with 10000 iterations for each selected $(n_1, p_1; n_2, p_2)$. The simulation results are summarized in next table, in which we use WCC and PLB to indicate the Wald interval with continuity correction and the profile likelihood based interval respectively. Since the Wald interval with continuity correction is not as good as the intervals with adjusted centers, we will not compare it with other intervals.

Through table 1, we can see that the profile likelihood based interval does improve upon the Wald interval on coverage probabilities in a frequentist sense. As listed in the table, its coverage probabilities are (much) higher than the coverage probabilities of Wald interval except a few points of $(n_1, p_1; n_2, p_2)$. Hence, the coverage of this interval is more reliable than the Wald interval. This suggests that

Table 3.1: Comparison of Confidence intervals at 95% level

				Coverage Probability					Length				
n_1	n_2	p_1	p_2	Wald	WCC	Bayes	AC	PLB	Wald	WCC	Bayes	AC	PLB
10	10	.9	.1	.870	.878	.953	.953	.857	.458	.658	.556	.579	.467
		.9	.8	.874	.989	.976	.985	.921	.567	.767	.601	.626	.638
		.8	.3	.896	.973	.951	.953	.949	.707	.907	.671	.699	.708
		.8	.7	.917	.975	.956	.968	.913	.707	.907	.671	.699	.728
		.6	.4	.919	.972	.955	.961	.941	.812	1.01	.730	.760	.812
		.6	.5	.907	.966	.955	.957	.941	.821	1.02	.736	.766	.822
		.5	.5	.910	.955	.954	.954	.945	.830	1.03	.741	.771	.831
15	15	.9	.1	.809	.955	.962	.971	.930	.396	.529	.450	.463	.406
		.9	.8	.933	.979	.960	.970	.923	.479	.613	.496	.511	.527
		.8	.3	.932	.955	.937	.956	.940	.591	.724	.568	.584	.596
		.8	.7	.931	.976	.953	.958	.933	.591	.724	.568	.584	.614
		.6	.4	.932	.973	.956	.959	.949	.676	.810	.626	.644	.676
		.6	.5	.930	.974	.933	.933	.931	.684	.817	.631	.649	.684
		.5	.5	.939	.954	.951	.954	.951	.691	.824	.636	.654	.689
20	20	.9	.1	.916	.920	.956	.956	.967	.352	.452	.387	.396	.367
		.9	.8	.943	.972	.960	.971	.931	.423	.523	.433	.445	.463
		.8	.3	.938	.965	.954	.954	.931	.517	.617	.501	.512	.523
		.8	.7	.940	.970	.949	.951	.939	.517	.617	.501	.512	.544
		.6	.4	.942	.973	.938	.956	.942	.591	.691	.556	.569	.596
		.6	.5	.931	.963	.949	.955	.949	.598	.698	.561	.574	.601
		.5	.5	.918	.958	.954	.954	.954	.604	.704	.566	.578	.606
50	50	.9	.1	.932	.969	.949	.949	.953	.231	.271	.240	.242	.248
		.9	.8	.944	.967	.953	.957	.941	.273	.313	.276	.279	.294
		.8	.3	.945	.969	.951	.951	.946	.333	.373	.329	.332	.354
		.8	.7	.940	.966	.944	.948	.943	.333	.373	.329	.332	.355
		.6	.4	.941	.957	.944	.944	.944	.380	.420	.370	.374	.394
		.6	.5	.939	.966	.940	.940	.940	.384	.424	.374	.377	.396
		.5	.5	.939	.962	.939	.939	.939	.388	.428	.377	.381	.397
20	10	.9	.1	.856	.955	.956	.960	.930	.409	.559	.479	.495	.434
		.9	.8	.873	.945	.972	.972	.906	.520	.670	.530	.549	.557
		.8	.3	.908	.966	.943	.951	.937	.632	.782	.597	.618	.631
		.8	.7	.913	.966	.949	.949	.939	.632	.782	.597	.618	.644
		.6	.4	.921	.964	.946	.951	.944	.710	.860	.649	.671	.701
		.6	.5	.917	.966	.947	.947	.945	.720	.870	.655	.677	.710
		.5	.5	.926	.969	.944	.947	.944	.725	.875	.659	.682	.714
10	20	.9	.1	.856	.958	.953	.957	.932	.410	.560	.479	.495	.434
		.9	.8	.945	.983	.970	.984	.927	.475	.626	.516	.533	.540
		.8	.3	.911	.963	.946	.955	.924	.604	.754	.587	.607	.609
		.8	.7	.921	.966	.954	.963	.923	.604	.754	.587	.607	.629
		.6	.4	.921	.962	.945	.950	.941	.710	.860	.645	.671	.701
		.6	.5	.919	.962	.944	.948	.939	.715	.865	.653	.675	.706
		.5	.5	.925	.967	.941	.945	.941	.725	.875	.659	.681	.714

Figure 3.8: Comparison of coverage probabilities for $p_1 = 0.21$ to 0.99, $p_2 = p_1 - 0.2$, $n_1 = n_2 = 20$ at 95% nominal level



the profile likelihood interval might be a good alternative to the Wald interval.

However, for small or moderate balanced sample sizes, the coverage behavior of the profile likelihood based interval is questionable. Figure 3.8 plots the coverage probabilities of the five 95% nominal intervals at $n_1 = n_2 = 20$ and $p_1 = 0.21$ to 0.99 with step-size 0.01 and $p_2 = p_1 - 0.2$. In this specific case, though the lengths of the profile likelihood based interval are always greater than those of the Wald interval, its coverage behavior is even worse than that of the Wald interval .

In general, one disadvantage of the profile likelihood based interval is, for balanced sample sizes, the lengths of the profile likelihood intervals are greater than those of the Wald intervals.

There is an "outlier" among the coverage probabilities of the profile likelihood based interval in the table. When $n_1 = n_2 = 10$ and $p_1 = 0.9$, $p_2 = 0.1$, the coverage probability of the profile likelihood based interval is only 0.857 while all the other coverage probabilities of this interval listed in the table are greater than 0.90. This is due to the discrete nature of the binomial distribution. Some quadruples (n_1, p_1, n_2, p_2) are lucky and some are unlucky. The quadruple $(n_1, p_1, n_2, p_2) = (10, 0, 9, 10, 0.1)$ is an unlucky one for the profile likelihood interval.

According to table 1, compared to the intervals with adjusted centers, the profile likelihood interval does not behave better. The coverage probabilities of the Bayes interval and the Agresti-Coull interval are seldom less than those of the profile likelihood interval and have less deviation according to the above table. For small or moderate sample sizes, when p_1 and p_2 are close, the coverage probabilities of intervals with adjusted centers tend to be greater than or equal to those of the profile likelihood interval. This property makes the intervals with adjusted centers more attractive because it is more common that the difference of the two proportions of interest is small or not very large.

In addition, except that p_1 and p_2 are close to boundaries, intervals with adjusted centers are always shorter than the corresponding profile likelihood based intervals. The other disadvantage of the profile likelihood based interval is that it does not have an explicit form.

Based on our evaluation, all the candidate intervals improve the coverage probabilities greatly upon the Wald interval in a frequentist sense. The phenomenon of over nominal coverage probability occurs quite often to the Wald interval with continuity correction, which tends to have the largest expected length. The profile likelihood based interval has a better coverage performance but greater expected length than the Wald interval when the binomial proportions are not close to the boundaries, and the computation of this interval is complex. Moreover, our extensive simulation shows the performance of the intervals with adjusted center is better than other intervals.

With respect to the five confidence interval methods discussed for constructing approximate $100(1-\alpha)\%$ two-sided intervals, we recommend intervals with adjusted centers as substitutes for the Wald interval. Because of their stable coverage behaviors, they have relatively reliable coverage performance even when n_1, n_2 are very small and p_1, p_2 are close to boundaries. Their simple expressions make the computation easier. Moreover, their lengths are not longer than other intervals in a frequentist sense. Especially, when the proportions are not very close to the boundaries, the lengths of the intervals with adjusted centers tend to be smaller than the others. As for which interval to choose between the Bayes interval and the Agresti-Coull interval, it depends on one's favor. The former is shorter and a little bit less conservative.

Chapter 4

Interval Estimation for the Difference of Two Binomial Proportions in Adaptive Designs

4.1 Introduction

In clinical trials and in industrial work, adaptive designs which use accumulating information to assign subjects to different treatments, are often highly desirable. People apply adaptive designs for two possible aims: first, to draw reliable statistical inferences for the benefit of future subjects, which can be thought of as an utilitarian goal. Second, to assign each subject to the treatment with better performance, which is the individualistic goal.

In this chapter, the approaches of constructing confidence intervals for non-adaptive designs will be applied to adaptive designs. A sequential adaptive model is considered in which two treatments are compared and the responses are binary: success or failure.

In section 4.2, notation and some existing adaptive designs will be introduced. The validity of extending non-adaptive methods to adaptive designs will be checked in section 4.3. As will be explained in more details, adaptive designs are classified into two categories: allocation adaptive designs and response adaptive designs. In section 4.4, the connection between the coverage performance and expected lengths of a confidence interval derived from a non-adaptive design and its counterpart from allocation adaptive design is stated and proved. In section 4.5, simulation results are given for response adaptive designs.

4.2 Notation and Some Adaptive Designs

The two populations to be compared are referred to as Population A and Population B, and $\{X_k : k \geq 1\}$ and $\{Y_k : k \geq 1\}$ denote the potential independent observations from populations A and B respectively. For each $k \geq 1$, exactly one of (X_k, Y_k) is actually observed. It is assumed that $(X_1, Y_1), (X_2, Y_2), \ldots$ are i.i.d., where $X_1 \sim \mathit{Bernoulli}(p_A)$ and $Y_1 \sim \mathit{Bernoulli}(p_B)$. The total sample size is n, the number of observations from populations A and B. For each $k > 1$, define δ_k to be 1 or 0 according to whether the kth object is assigned to population A or B. The symbols $N_A(k)$ and $N_B(k)$ indicate the numbers of the first k observations that are allocated to population A and B through stage k . Then

$$N_A(k) = \sum_{i=1}^{k} \delta_i$$

and

$$N_B(k) = \sum_{i=1}^{k} (1 - \delta_i) = k - N_A(k).$$

Further, define $S_A(k)$ and $S_B(k)$ to be the numbers of successes from populations A and B through stage k. Then

$$S_A(k) = \sum_{i=1}^{k} \delta_i X_i$$

and

$$S_B(k) = \sum_{i=1}^{k} (1 - \delta_i) Y_i.$$

As stated in Geraldes (1999), most adaptive designs fit into one of two general categories: allocation adaptive designs and response adaptive designs. The former encompasses those approaches for which the allocation of each subject does not depend on the responses of previous subjects but only depends on the subject's covariate levels (when covariate information is taken into consideration) and the allocations and covariate levels of the previous subjects. The second category includes those approaches for which the allocation of each subject depends also on the responses of the previous subjects. Hence, the main difference between allocation adaptive designs and response adaptive designs is that $(X_1, Y_1), (X_2, Y_2), \ldots$ are independent of the δ's in allocation adaptive designs and they are dependent in response adaptive designs.

There are a lot of adaptive designs in the literature, for example, the doubly adaptive biased coin design proposed by Eisele (1994), the play-the-winner design proposed by Smythe and Rosenberger (1995). Woodroofe (1982) considers the problem of sequentially allocating patients to treatments when covariate information is present. We will introduce some adaptive designs in the next two subsections.

4.2.1 Some Allocation-Adaptive Designs

The *Biased Coin Design*, proposed by Efron (1971), allocates the next subject to one of the two populations, A or B, according to the following rule. Let D_k denote the difference of $N_A(k)/k$ and $N_B(k)/k$. Let p_0 be a constant in $[0.5, 1]$. Then

$$P(\delta_{k+1} = 1) = \begin{cases} 1 - p_0, & \text{if } D_k > 0; \\ 1/2, & \text{if } D_k = 0; \\ p_0, & \text{if } D_k < 0. \end{cases}$$

This allocation policy tends to balance the number of observations from both populations.

The *Adaptive Biased Coin Design*, proposed by Wei (1978), allocates subjects to A or B according to the following rule. Let D_k denote the difference of $N_A(k)/k$ and $N_B(k)/k$. Let $h : [-1, 1] \to [0, 1]$ be a non-increasing function such that $h(x) = 1 - h(-x)$ for any $x \in [-1, 1]$. Then $P(\delta_{k+1} = 1) = h(D_k)$. This allocation policy may force an extremely imbalanced experiment to be balanced very quickly.

4.2.2 Some Response-Adaptive Designs

The *Randomized Play-the-Winner Rule*, proposed by Wei and Durham (1978), tends to allocate more subjects to the population with higher success proportion. This rule can be described with an urn model. An urn has balls of two different types, marked A or B. We start with α balls of each type. When a subject enters the study, a ball is drawn at random and replaced. If it is type A, then the subject is assigned to A. It is assigned to B otherwise. If the observation of the subject is a success, then β balls of the same type are added. Otherwise, β balls of the other type added to the urn. This rule is denoted by $RPW(\alpha, \beta)$.

The *Randomized Adaptive Design*, proposed by Melfi and Page (1995), tends to allocate subjects to both populations according to an optimal proportion. Suppose $\{U_k : k \geq 1\}$ are a sequence of i.i.d. random variables, whose common distribution is $U(0,1)$, and which are independent of both $\{X_k : k \geq 1\}$ and $\{Y_k : k \geq 1\}$. To minimize the variance of the estimator$\hat{p}_A(k) - \hat{p}_B(k)$, the desired proportion is

$$\pi(p_A, p_B) = \frac{\sqrt{p_A(1-p_A)}}{\sqrt{p_A(1-p_A)} + \sqrt{p_B(1-p_B)}}.$$

Let $\hat{\pi}_k = \pi(\hat{p}_A(k), \hat{p}_B(k))$, where $\hat{p}_A(k), \hat{p}_A(k)$ are two estimators of the success probabilities p_A and p_B. Then

$$\delta_{k+1} = I\{U_{k+1} < \hat{\pi}(k)\}.$$

4.3 The Confidence Intervals in Adaptive Designs

Because of the adaptive nature of the design, the distribution of $S_A(k)$ may no longer be Binomial $Bin(N_A(k), p_A)$. And $S_A(k)$, $S_B(k)$ are no longer independent. Therefore, the validity of constructing confidence intervals using the non-adaptive formulas needs to be verified.

The maximum likelihood estimators, at stage k, of the success probabilities, p_A and p_B, are

$$\hat{p_A}(k) = \frac{S_A(k)}{N_A(k)}$$

and

$$\hat{p_B}(k) = \frac{S_B(k)}{N_B(k)}.$$

Some statisticians have studied asymptotic properties in some adaptive design settings, such as Eisele and Woodroofe (1995), Bai *et al.* (2002), Rosenberger (1993) and Rosenberger *et al.* (1997).

Melfi *et al.* (2001) prove some theorems and applied them to show that

$$\left(\frac{N_A(k)^{1/2}(\hat{p_A}(k) - p_A)}{(p_A q_A)^{1/2}}, \frac{N_B(k)^{1/2}(\hat{p_B}(k) - p_B)}{(p_B q_B)^{1/2}}\right) \xrightarrow{\mathcal{L}} (Z_1, Z_2) \qquad (4.3.1)$$

under a wide range of adaptive design rules, where Z_1 and Z_2 are independent standard normal random variables. Wei *et al.* (1990) proved the same result under randomized play the winner rule using martingale technique. Therefore, the adaptive version of the Wald confidence interval and the Wald interval with continuity correction up to stage k with nominal level $100(1-\alpha)\%$ are

$$CI_W^A(k) = \hat{p_A}(k) - \hat{p_B}(k) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p_A}(k)\hat{q_A}(k)}{N_A(k)} + \frac{\hat{p_B}(k)\hat{q_B}(k)}{N_B(k)}} \qquad (4.3.2)$$

and

$$CI^A_{WCC}(k) = \hat{p_A}(k) - \hat{p_B}(k) \pm \left(z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p_A}(k)\hat{q_A}(k)}{N_A(k)} + \frac{\hat{p_B}(k)\hat{q_B}(k)}{N_B(k)}} + \frac{1}{2N_A(k)} + \frac{1}{2N_B(k)} \right) \tag{4.3.3}$$

When $\Delta = \Delta_0$ for $\Delta = p_A - p_B$, it follows from (4.3.1) and the arguments in Cox and Hinkley(1974, page 322-323) that the variable $2\{l(\hat{\Delta}(k), \hat{p}_B(k)) - \tilde{l}(\Delta_0)\}$ is asymptotically chi-squared distributed with one degree of freedom, where $\hat{\Delta}(k) = \hat{p}_A(k) - \hat{p}_B(k)$. Thus an approximate $100(1-\alpha)\%$ profile likelihood based confidence interval for $p_A - p_B$ of adaptive designs is:

$$CI^A_{PLB}(k) = \{\Delta \in (-1,1) : 2(l(\hat{\Delta}(k), \hat{p}_B(k)) - \tilde{l}(\Delta)) \leq \chi^2_1(\alpha)\}. \tag{4.3.4}$$

To derive the confidence intervals with adjusted centers for adaptive designs, we define two estimators for p_A and p_B:

$$\tilde{p_A}(k) = \frac{S_A(k) + 1}{N_A(k) + 2}$$

and

$$\tilde{p_B}(k) = \frac{S_B(k) + 1}{N_B(k) + 2}.$$

Theorem 4.3.1. *In the above adaptive setting, if* $\frac{N_A(k)}{a_k} \to 1$ *and* $\frac{N_B(k)}{b_k} \to 1$ *in probability as* $k \to \infty$, *where* $\{a_k, b_k\}$ *are positive constants with* a_k *and* b_k *tending to infinity. Then,*

$$\left(\frac{(N_A(k) + c)^{1/2}(\tilde{p_A}(k) - p_A)}{\sqrt{p_A q_A}}, \frac{(N_B(k) + c)^{1/2}(\tilde{p_B}(k) - p_B)}{\sqrt{p_B q_B}} \right) \stackrel{\mathcal{L}}{\Longrightarrow} (Z_1, Z_2), \tag{4.3.5}$$

where c *is a constant and* Z_1, Z_2 *are independent standard normal random variables.*

Proof. The desired conclusion is a direct result of Corollary 3.1 in Melfi *et al.* (2001). □

This theorem gives the validity of the confidence intervals with adjusted centers for adaptive designs. Hence, the nominal level $100(1-\alpha)\%$ Bayes and Agresti-Coull confidence intervals for adaptive designs are

$$CI_B^A(k) = \tilde{p_A}(k) - \tilde{p_B}(k) \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\tilde{p_A}(k)\tilde{q_A}(k)}{N_A(k)+3} + \frac{\tilde{p_B}(k)\tilde{q_B}(k)}{N_B(k)+3}} \qquad (4.3.6)$$

and

$$CI_{AC}^A(k) = \tilde{p_A}(k) - \tilde{p_B}(k) \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\tilde{p_A}(k)\tilde{q_A}(k)}{N_A(k)+2} + \frac{\tilde{p_B}(k)\tilde{q_B}(k)}{N_B(k)+2}}. \qquad (4.3.7)$$

We will consider the performance of the above confidence intervals in the next section.

4.4 Comparison of Confidence Intervals in Allocation Adaptive Designs

For convenience, we define some notation: $CI_*^A(k)$ is the confidence interval derived by method $*$ and based on a certain adaptive design up to stage k, where $*$ might be W, WCC, B, AC and PLB that indicate the Wald interval, the Wald interval with continuity correction, the approximate Bayes interval, the Agresti-Coull interval and the profile likelihood based interval respectively. If necessary, we may replace A by a specific adaptive design. And $CI_*(i,j)$ is the counterpart of $CI_*^A(k)$ from a non-adaptive design with i observations from population A and j observations from population B, where $i+j = k$. Similarly, use $EL_*^A(k)$ to denote the expected length of the confidence interval derived by method $*$ and based on an adaptive design. $EL_*(i,j)$ represents the counterpart of EL_*^A with i observations from population A and j observations from population B.

The following theorem explores the connection of the coverage probabilities between confidence intervals based on allocation adaptive designs and non-adaptive designs.

Theorem 4.4.1. *In allocation adaptive designs,*

$$P(p_A - p_B \in CI_*^A(k)) = \sum_{j=0}^{k} P(p_A - p_B \in CI_*(j, k-j))P(N_A(k) = j)$$

where the $*$ may be any confidence interval that the non-adaptive version $CI_*(j, k-j)$ only involves sufficient statistics: S_A and S_B.

The proof of this theorem is based on the next Lemma.

Lemma 4.4.1. *In allocation adaptive designs, suppose a and b are any non-negative integers satisfying $a \leq j$ and $b \leq k-j$, then*

1. $P(S_A(k) = a | N_A(k) = j) = \binom{j}{a} p_A^a (1-p_A)^{j-a}$;

2. $P(S_B(k) = b | N_A(k) = j) = \binom{k-j}{b} p_B^b (1-p_B)^{k-j-b}$ *and*

3. $P(S_A(k) = a, S_B(k) = b | N_A(k) = j)$

   $= P(S_A(k) = a | N_A(k) = j) P(S_B(k) = b | N_A(k) = j)$.

Proof. Let $\overrightarrow{\delta} = (\delta_1, \ldots, \delta_k)$.

For any $j \in \{0, 1, \ldots, k\}$,

$$\{N_A(k) = j\} = \bigcup_{t=1}^{C_k^j} \{\overrightarrow{\delta} = \overrightarrow{\delta_t}\}$$

where $\overrightarrow{\delta_t}$ is such a k-dimension vector that has j elements with value 1 and the other $k - j$ elements with value 0. There are $C_k^j = \binom{k}{j}$ different such vectors. We put them in order.

Note that

$$\begin{aligned}
&P(S_A(k) = a \,|\, N_A(k) = j) \\
=&P\left(\sum_{i=1}^{k} \delta_i X_i = a \,\Big|\, \sum_{i=1}^{k} \delta_i = j\right) \\
=&\sum_{t=1}^{C_k^j} P\left(\sum_{i=1}^{k} \delta_i X_i = a, \overrightarrow{\delta} = \overrightarrow{\delta_t} \,\Big|\, \sum_{i=1}^{k} \delta_i = j\right) \\
=&\sum_{t=1}^{C_k^j} P\left(\sum_{i=1}^{j} X_i = a, \overrightarrow{\delta} = \overrightarrow{\delta_t} \,\Big|\, \sum_{i=1}^{k} \delta_i = j\right) \\
=&P\left(\sum_{i=1}^{j} X_i = a \,\Big|\, \sum_{i=1}^{k} \delta_i = j\right) && (4.4.1) \\
=&\binom{j}{a} p_A^a (1 - p_A)^{j-a}
\end{aligned}$$

(4.4.2)

The third step is valid because $\{\delta_l, l \geq 1\}$ is independent of the $i.i.d$ sequence $\{X_i, i = 1, 2, \ldots\}$ in allocation adaptive design. Hence, those δ's only indicate when to take observations from A and B. We have proved that $S_A(k)$ has a conditionally binomial distribution.

Similarly, we can prove the conclusion related to $P(S_B(k) = b \,|\, N_A(k) = j)$.

Next to prove that given $N_A(k) = j$, $S_A(k)$ and $S_B(k)$ are independent.

The conditionally joint distribution of $S_A(k)$ and $S_B(k)$ is

$$\begin{aligned}
&P(S_A(k) = a, S_B(k) = b \mid N_A(k) = j) \\
=&P\left(\sum_{i=1}^{k} \delta_i X_i = a, \sum_{i=1}^{k}(1-\delta_i)Y_i = b \mid \sum_{i=1}^{k} \delta_i = j\right) \\
=&\sum_{t=1}^{C_k^j} P\left(\sum_{i=1}^{j} X_i = a, \sum_{i=1}^{k-j} Y_i = b, \overrightarrow{\delta} = \overrightarrow{\delta_t} \mid \sum_{i=1}^{k} \delta_i = j\right) \\
=&P\left(\sum_{i=1}^{j} X_i = a, \sum_{i=1}^{k-j} Y_i = b, \mid \sum_{i=1}^{k} \delta_i = j\right)
\end{aligned} \tag{4.4.3}$$

Therefore, by the independence of the responses and the allocations, equation (4.4.3) can be rewritten in the following way:

$$\begin{aligned}
&P(S_A(k) = a, S_B(k) = b \mid N_A(k) = j) \\
=&P\left(\sum_{i=1}^{j} X_i = a \mid N_A(k) = j\right) P\left(\sum_{i=1}^{k-j} Y_i = b \mid N_A(k) = j\right) \\
=&P\left(S_A(k) = a | N_A(k) = j\right) P\left(S_B(k) = b | N_A(k) = j\right)
\end{aligned} \tag{4.4.4}$$

Hence, the lemma holds. □

Proof. (of Theorem 4.4.1).

For any confidence interval,

$$\begin{aligned}
&P(p_A - p_B \in CI_*^A(k)) \\
=&\sum_{j=0}^{k} P(p_A - p_B \in CI_*^A(k), N_A(k) = j) \\
=&\sum_{j=0}^{k} P(p_A - p_B \in CI_*^A(k) | N_A(k) = j) P(N_A(k) = j)
\end{aligned} \tag{4.4.5}$$

If the non-adaptive version of the confidence interval only involves sufficient statistics: S_A and S_B, the desired conclusion is achieved by applying Lemma 4.4.1 to (4.4.5). □

Remark 4.4.1. The condition that $\frac{N_A(k)}{a_k} \to 1$ and $\frac{N_B(k)}{b_k} \to 1$ in probability as $k \to \infty$ is not needed for the proof procedure, but does guarantee the validity of the asymptotic normality needed in constructing those confidence intervals in general adaptive designs.

There is a similar theorem concerning the connection of the expected lengths of confidence intervals in non-adaptive designs and allocation adaptive designs.

Theorem 4.4.2. *In allocation adaptive designs,*

$$EL_*^A = \sum_{j=0}^{k} EL_*(j, k-j) P(N_A = j),$$

where the $*$ *may be any confidence interval that the non-adaptive version* $CI_*(j, k-j)$ *only involves the sufficient statistics:* S_A *and* S_B.

Proof. This proof is similar to the one of Theorem 4.4.1. □

Remark 4.4.2. The five confidence intervals considered in the dissertation satisfy the requirements of the two theorems.

Remark 4.4.3. The two theorems imply that for allocation adaptive designs, a confidence interval should behave well if it behaves well in non-adaptive designs.

4.5 Comparison of Confidence Intervals in Response Adaptive Designs

For response adaptive designs, we do not have simple results as we do in allocation adaptive designs. The main reason is Lemma 4.4.1 does not hold in response adaptive designs because δ's are not independent of the responses X's and Y's.

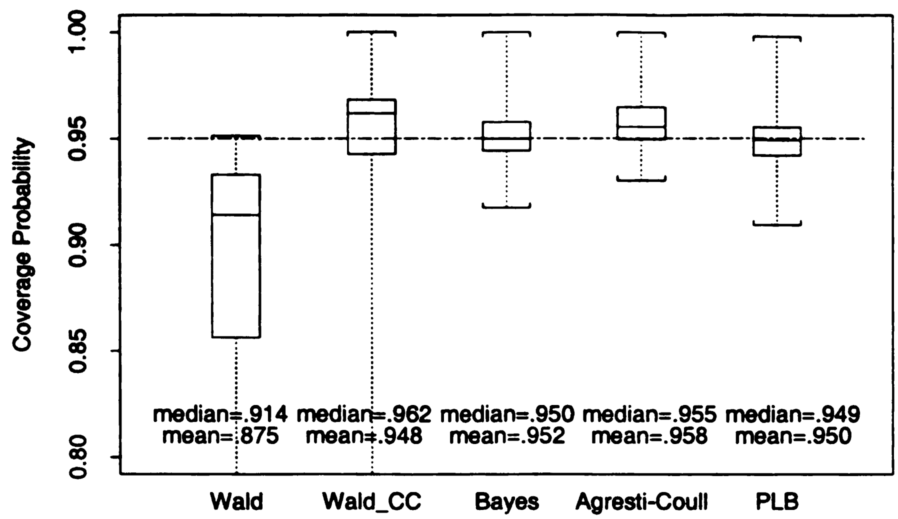
However, for response adaptive designs, we still have the same conclusion via simulation: if a confidence interval behaves well in non-adaptive designs, one may expect this confidence interval to behave well in response adaptive designs. We obtain this conclusion through extensive simulation studies on some response adaptive designs. All the results shown use a simulation with 10000 iterations for each realized (n, p_A, p_B).

We concentrate on RPW(1,1), the randomized play the winner design with $\alpha = 1$ and $\beta = 1$, in this dissertation. Similar conclusions hold for some other response adaptive designs such as the randomized adaptive designs.

As we did in non-adaptive designs, to explore the average performance of the five confidence intervals, we randomly sampled 10,000 values of (n, p_A, p_B), taking p_A and p_B independently from $U(0, 1)$ and taking n from uniform distribution over $\{10, 11, \ldots, 100\}$. We then applied $RPW(1, 1)$ rule to the sampled n to achieve the sample sizes from the two treatments.

Figure 4.1 shows the average coverage performance of the five intervals with means and medians of the coverage probabilities listed. Similar to the results in

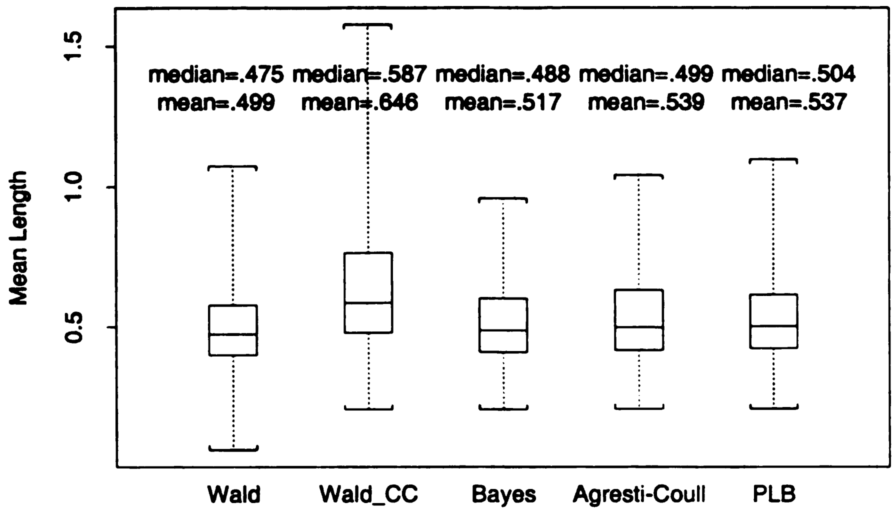
Figure 4.1: Coverage probability Boxplots of some 95% nominal intervals upon RPW(1,1)



non-adaptive designs, the Wald interval behaves poorly: the coverage probability is unstable and very low with median 0.914 and mean 0.875 at the 95% nominal level. It also occasionally has very low coverage probabilities. Though the coverage probabilities of the Wald interval with continuity corrections are higher than those of the Wald interval, it inherits some disadvantages of the Wald interval too: occasional very low coverage probabilities and unstable performance. The average coverage behaviors of the Bayes interval, the Agresti-Coull interval and the Profile likelihood based interval are very similar: their means and medians are close to the 95% nominal level. The profile likelihood interval is not as stable as the intervals with adjusted centers.

When comparing Figure 3.1 and 4.1 or simply comparing the corresponding

Figure 4.2: Expected Length Boxplots of some 95% nominal intervals under RPW(1,1)



mean and median coverage probabilities, we notice one interesting point. The average coverage performance of the Wald interval and the Wald interval with continuity correction in RPW(1,1) is much worse than it is in non-adaptive designs. However, this is not true for intervals with adjusted centers, which makes the intervals with adjusted centers desirable with RPW(1,1) and some other adaptive designs because of their stable performance.

We also plot the mean length boxplots of the five intervals with RPW(1,1) in Figure 4.2. Since the Wald interval with continuity correction is too wide compared to other intervals, we discard it in the following comparisons. And because the two intervals with adjusted centers are very similar, we will only consider the Agresti-Coull interval henceforth.

Figure 4.3: Coverage probabilities of three 95% nominal intervals for $n = 20$ and $p_A = 0.5$ under RPW(1,1)

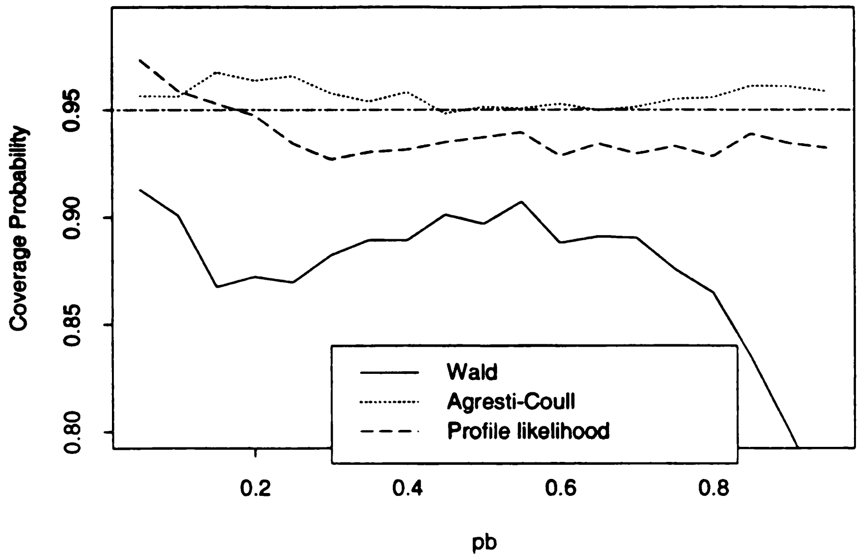


Figure 4.4: Expected lengths of three 95% nominal intervals for $n = 20$ and $p_A = 0.5$ upon RPW(1,1)

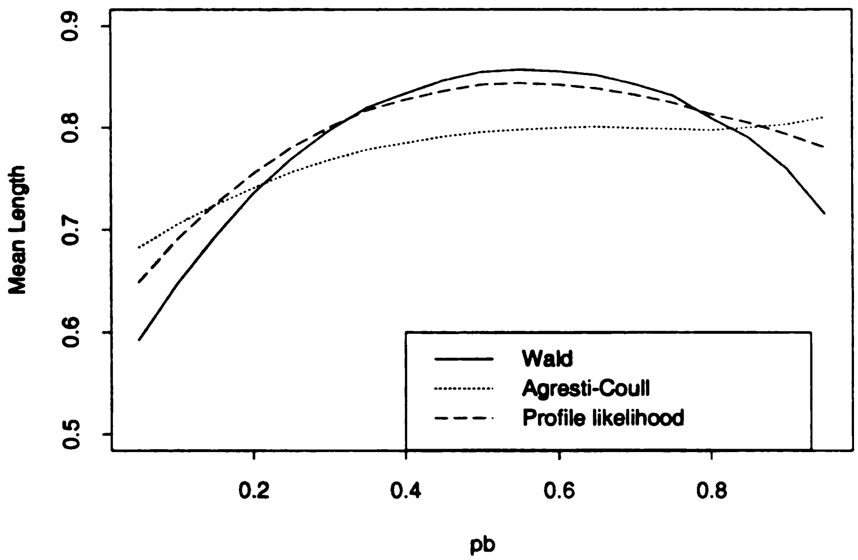


Figure 4.5: Coverage probabilities of three 95% nominal intervals for $n = 20$ and $p_A = 0.9$ upon RPW(1,1)

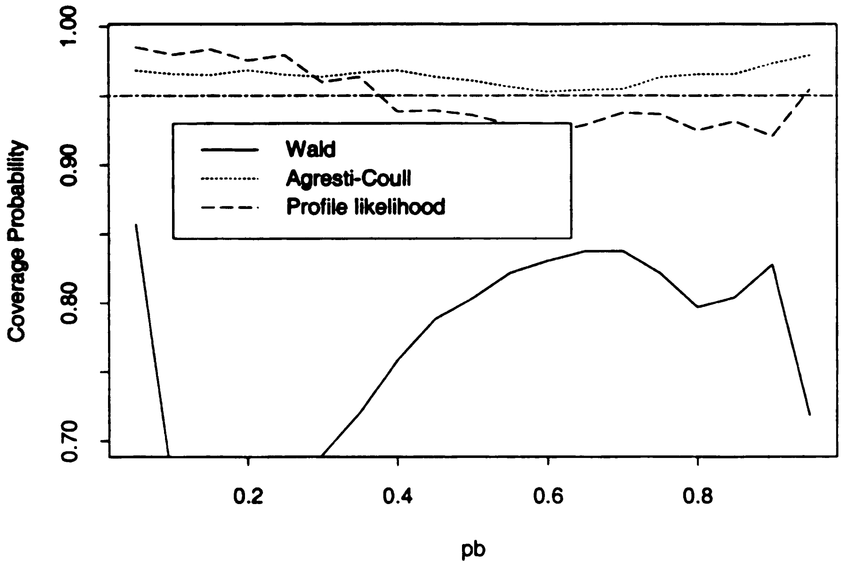
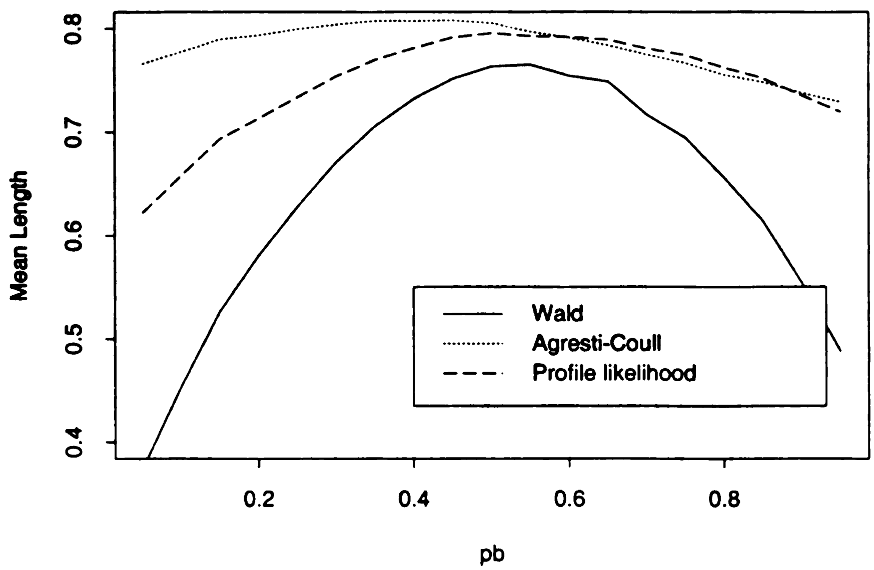


Figure 4.3 plots the coverage probabilities of the Wald interval, the Agresti-Coull interval and the profile likelihood based interval at the 95% nominal level for $n = 20$, $p_A = 0.5$, $p_B = 0.05$ through 0.95 with step-size 0.05 with RPW(1,1). And Figure 4.4 plots the corresponding mean lengths of those three intervals. We see that the Agresti-Coull interval has both satisfactory coverage performance and the expected length in this setup. Its coverage probabilities are almost all (right) above the nominal level and it has the shortest length unless p_B is close to either boundary, i.e, the difference of $p_A - p_B$ is not very big.

Though the values p_B taken are symmetric around $p_A = 0.5$, Figure 4.3 and Figure 4.4 do not exhibit any symmetry. This is due to the adaptive nature of the RPW(1,1) design and p_A/p_B being not symmetric around $p_A = 0.5$.

Figure 4.6: Expected lengths of three 95% nominal intervals for $n = 20$ and $p_A = 0.9$ upon RPW(1,1)



Different from the setup of Figure 4.3 and Figure 4.4, let $p_A = 0.9$ in Figure 4.5 and Figure 4.6. When p_A is far from p_B, the coverage probability of the profile likelihood interval is rather high. It drops when p_A and p_B gets closer. Contrary to the profile likelihood interval, the coverage probability of Agresti-Coull interval is much less sensitive to the relative positions of p_A and p_B. The coverage remains above nominal level. Though the expected length of the Agresti-Coull interval is much greater than that of the profile likelihood based interval when $p_A - p_B$ is large, it is close to the latter when $p_A - p_B$ is not very large. This is verified through our extensive simulation. Actually, when the total sample size n increases, the disadvantage of the expected length of the Agresti-Coull interval when $p_A - p_B$ is large decreases. For example, when $n = 100$ and keep p_A and p_B same as in Figure

4.6, the mean lengths of the three intervals are comparable. The expected length of the Agresti-Coull interval is less than that of the profile likelihood based interval most of the time and it has the smallest length when p_B is not very close to either boundary.

In Figure 4.6, one may notice that the expected length of the Wald interval is much smaller of those of the other two intervals, especially when p_B is close to 0 or 1. This is due to the high frequency of the occurrence of the empty Wald interval when the sample size is small and the success proportions are close to boundaries. This is also the reason for the low coverage probability of the Wald interval in Figure 4.5. The feature of the much lower expected length of the Wald interval is not so obvious or does not exist for moderate(say, $n = 50$) or large sample size(say, $n = 100$).

Let us compare the three confidence intervals from another point of view: let the total sample size n vary and keep p_A and p_B as constants.

Figure 4.7 and Figure 4.8, respectively, plot the coverage probabilities and mean lengths of the three confidence intervals for n varying from 10 through 100 with $p_A = 0.7$ and $p_B = 0.4$. The Agresti-Coull interval has both the highest coverage probability and the shortest expected length for most values of n. This makes the Agresti-Coull interval very attractive in application. Another advantage of the Agresti-Coull interval is it may achieve the nominal level for very small sample sizes. When the sample size increases, the coverage probability of the Agresti-Coull interval tends to go down and fluctuate around the nominal level which may be

Figure 4.7: Coverage probabilities of three 95% nominal intervals for $n = 10 - 100$ and $p_A = 0.7$, $p_B = 0.4$ upon RPW(1,1)

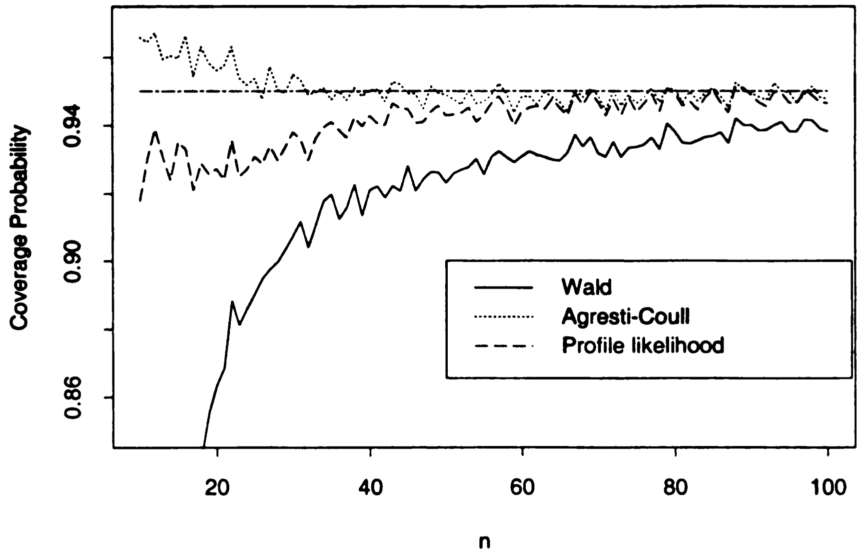
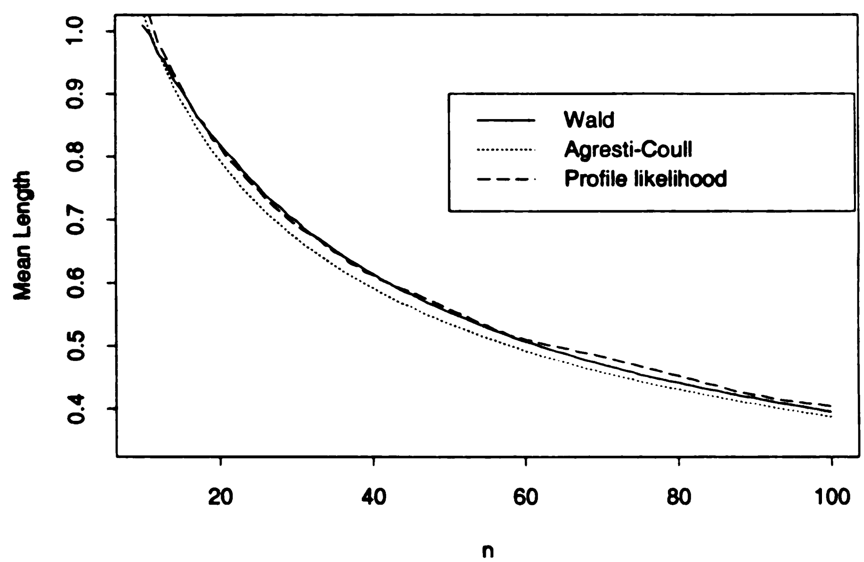


Figure 4.8: Expected lengths of three 95% nominal intervals for $n = 10 - 100$ and $p_A = 0.7$, $p_B = 0.4$ upon RPW(1,1)



explained by the central limit theory for adaptive designs.

Our extensive simulation shows that the Agresti-Coull interval always has the most satisfactory coverage probability when p_A and p_B are not far apart from each other (say, $|p_A - p_B| < 0.5$). When $|p_A - p_B|$ is very large, the profile likelihood based interval has the highest coverage probability. The expected length of the Agresti-Coull interval is also satisfactory unless the two proportions are close to boundaries.

4.6 Conclusion

In summary, compared to other intervals discussed, the intervals with adjusted centers behave best with RPW(1,1). They have both stable and satisfactory coverage probabilities and expected lengths in a frequentist sense. The stableness of the two intervals makes them good intervals in other adaptive designs. Our simulation with some other adaptive designs such as the randomized adaptive designs and adaptive weighted difference designs, due to Geraldes (1999), confirms this conclusion. Therefore, we suggest the intervals with adjusted centers to be used in adaptive designs.

One may expect to improve the coverage performance of the intervals with adjusted centers for large sample size by adjusting the weights of $\hat{p}_A(k)$ and $1/2$ when defining $\tilde{p}_A(k)$ for large k's. We may adjust $\tilde{p}_B(k)$ the same way.

Bibliography

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. *Amer. Statist.* **54** 280–288.

AGRESTI, A. and COULL, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *Amer. Statist.* **52** 119–126.

BAI, Z. D., HU, F. and ROSENBERGER, W. F. (2002). Asymptotic properties of adaptive designs for clinical trials with delayed response. *Ann. Statist.* **30** 122–139.

BERRY, D. (1996). *Statistics: A Bayesian Perspective*. Belmont, CA:Wadsworth.

BHATTACHARYA, R. N. and RANGA RAO, R. (1976). *Normal approximation and asymptotic expansions*. John Wiley & Sons, New York-London-Sydney. Wiley Series in Probability and Mathematical Statistics.

BROWN, L. D., CAI, T. T. and DASGUPTA, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *Ann. Statist.* **30** 160–201.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical statistics*. Chapman and Hall, London.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika* **58** 403–417.

Eisele, J. R. (1994). The doubly adaptive biased coin design for sequential clinical trials. *J. Statist. Plann. Inference* **38** 249–261.

Eisele, J. R. and Woodroofe, M. B. (1995). Central limit theorems for doubly adaptive biased coin designs. *Ann. Statist.* **23** 234–254.

Geraldes, M. (1999). Covariates in adaptive designs for clinical trials. *Ph.D Dissertation, Michigan State University* .

Hall, P. (1992). *The bootstrap and Edgeworth expansion*. Springer-Verlag, New York.

Melfi, V. F. and Page, C. (1995). *Randomized adaptive designs*. Inst. Math. Statist., Hayward, CA.

Melfi, V. F., Page, C. and Geraldes, M. (2001). An adaptive randomized design with application to estimation. *Canad. J. Statist.* **29** 107–116.

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

ROSENBERGER, W. F. (1993). Asymptotic inference with response-adaptive treatment allocation designs. *Ann. Statist.* **21** 2098–2107.

ROSENBERGER, W. F., FLOURNOY, N. and DURHAM, S. D. (1997). Asymptotic normality of maximum likelihood estimators from multiparameter response-driven designs. *J. Statist. Plann. Inference* **60** 69–76.

SMYTHE, R. T. and ROSENBERGER, W. F. (1995). Play-the-winner designs, generalized Pólya urns, and Markov branching processes. In *Adaptive designs (South Hadley, MA, 1992).* Inst. Math. Statist., Hayward, CA 13–22.

WEI, L. J. (1978). The adaptive biased coin design for sequential experiments. *Ann. Statist.* **6** 92–100.

WEI, L. J. and DURHAM, S. (1978). The randomized play-the-winner rule in medical trials. *J. Amer. Statist. Assoc.* **73** 840–843.

WEI, L. J., SMYTHE, R. T., LIN, D. Y. and PARK, T. S. (1990). Statistical inference with data-dependent treatment allocation rules. *J. Amer. Statist. Assoc.* **85** 156–162.

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

WOODROOFE, M. (1982). Sequential allocation with covariates. *Sankhyā Ser. A* **44** 403–414.