8

The Paired 2×2 *Table*

8.1 Introduction

This chapter considers tests for association, effect measures, and confidence intervals for paired binomial probabilities. Paired binomial probabilities arise in study designs such as matched and cross-over clinical trials, longitudinal studies, and matched case-control studies. The data consist of two samples of dichotomous events: Event A and Event B. Each observation of Event A is matched with one observation of Event B. The two observations in a matched pair may come from the same subject, such as in cross-over clinical trials, where each subject is measured twice (treatment A and treatment B). In matched case-control studies, however, each matched pair refers to two different subjects: one case (Event A) and one matching control (Event B). The purpose of a case-control study is to compare the exposure history between cases and controls. In both situations, the outcomes in the two samples are dependent. The paired 2×2 table may also the result of the measurements of two raters, and if inter-rater agreement is of interest, the methods in Section 13.2 should be used.

The results of studies of paired binomial probabilities can be summarized in a paired 2×2 table, as shown in Table 8.1. The possible outcomes for each event is either success or failure. As usual, success does not necessarily indicate a favorable outcome but rather the outcome of interest, which may, for instance, be the presence of a certain disease. The paired 2×2 table may look like the unpaired 2×2 table in Chapter 4 (see Table 4.1), but the statistical methods used to analyze unpaired and paired 2×2 tables are not the same. Because the two samples of observations in a paired 2×2 table are matched, the statistical methods used to analyze paired 2×2 tables must account for dependent data. We also note that Table 8.1 (unlike Table 4.1) contains $2N$ observations, because each count consists of a pair of observations.

Section 8.2 gives examples of published studies with paired 2×2 table data that illustrate different study designs, and Section 8.3 introduces the notation and the relevant sampling distribution. Two main categories of statistical models (marginal and subject-specific models) are described in Section 8.4. Tests for association are described in Section 8.5. The next four sections present confidence intervals for the difference between probabilities (Section 8.6), the number needed to treat (Section 8.7), the ratio of probabilities (Section 8.8),

TABLE 8.1
The observed counts of a paired 2×2 table

	Event B		
Event A	**Success**	**Failure**	**Total**
Success	n_{11}	n_{12}	n_{1+}
Failure	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	N

Additional notation:
$\mathbf{n} = \{n_{11}, n_{12}, n_{21}, n_{22}\}$: the observed table
$\mathbf{x} = \{x_{11}, x_{12}, x_{21}, x_{22}\}$: any possible table

and the odds ratio (Section 8.9). Section 8.10 gives recommendations for the practical use of the methods in Sections 8.5–8.9. This chapter is partly based on Fagerland et al. (2013) and Fagerland et al. (2014).

8.2 Examples

8.2.1 Airway Hyper-Responsiveness Status before and after Stem Cell Transplantation

Stem cell transplantation (SCT) is a recognized treatment option for patients with hematological (and various other) malignancies (Bentur et al., 2009). SCT is, however, associated with pulmonary complications. In a prospective longitudinal study, Bentur et al. (2009) measured the airway hyper-responsiveness (AHR) status of 21 children before and after SCT. The purpose of the study was to investigate whether the prevalence of AHR increases following SCT. The results of the study are summarized in Table 8.2. Two children (9.5%) had AHR before SCT and eight (38%) children had AHR after SCT.

TABLE 8.2
Airway hyper-responsiveness (AHR) status before and after stem cell transplantation (SCT) in 21 children (Bentur et al., 2009)

	After SCT		
Before SCT	**AHR**	**No AHR**	**Total**
AHR	1	1	2 (9.5%)
No AHR	7	12	19 (91%)
Total	8 (38%)	13 (62%)	21 (100%)

The two measurements of AHR that constitute a matched pair come from the same patient, and the matching is on the exposure variable (SCT). We can analyze Table 8.2 in several ways. We can formulate a null hypothesis that the probabilities of AHR before and after SCT are equal. Section 8.5 considers tests for association that can be used to test this hypothesis. We can also estimate the strength of the relationship between SCT and AHR status with four different effect measures: the difference between probabilities (Section 8.6), the number needed to treat (Section 8.7), the ratio of probabilities (Section 8.8), and the odds ratio (Section 8.9). We shall return to this example when we illustrate the statistical methods later in this chapter.

8.2.2 Complete Response before and after Consolidation Therapy

The study in the previous example had a small sample size with only 21 pairs of observations. We now consider a similar but larger study with 161 pairs of observations. Cavo et al. (2012) report the results of a randomized clinical trial of two induction therapy treatments before autologous stem cell transplantation for patients with multiple myeloma. A secondary endpoint of the trial was to assess the efficacy and safety of subsequent consolidation therapy. The results for one of the treatment arms are shown in Table 8.3. The outcome was complete response (CR), confirmed from bone marrow biopsy samples, and each patient was measured before and after consolidation therapy. The study design (longitudinal) is the same as in the previous example, and each matched pair consists of two measurements from one patient. An increase in the proportion of patients with CR following consolidation therapy can be observed: sixty-five (40%) patients had CR before consolidation therapy, and 75 (47%) patients had CR after consolidation therapy. Table 8.3 can be analyzed with tests for association and—because the matching is on the exposure variable (consolidation therapy)—with the same effect measures as the previous example.

TABLE 8.3
Complete response (CR) before and after consolidation therapy (Cavo et al., 2012)

	After consolidation		
Before consolidation	**CR**	**No CR**	**Total**
CR	59	6	65 (40%)
No CR	16	80	96 (60%)
Total	75 (47%)	86 (53%)	161 (100%)

8.2.3 The Association between Floppy Eyelid Syndrome and Obstructive Sleep Apnea-Hypopnea Syndrome

We now turn to a different study design: the matched case-control study. In a study reported by Ezra et al. (2010), 102 patients with floppy eyelid syndrome (FES, the cases) were 1:1 matched to 102 patients without FES (the controls). The patients were matched according to age, gender, and body mass index. One of the aims of the study was to investigate the association between FES (the disease) and obstructive sleep apnea-hypopnea syndrome (OSAHS, the exposure). Table 8.4 shows the results. Each pair of observations now consists of the OSAHS status of one case and the OSAHS status of one matching control. Thirty-two (31%) of the 102 cases had OSAHS, whereas only nine (8.8%) of the 102 controls had OSAHS.

TABLE 8.4
The observed association between floppy eyelid syndrome (FES, the disease) and obstructive sleep apnea-hypopnea syndrome (OSAHS, the exposure) in a matched case-control study (Ezra et al., 2010)

	Controls (no FES)		
Cases (FES)	**OSAHS**	**No OSAHS**	**Total**
OSAHS	7	25	32 (31%)
No OSAHS	2	68	70 (69%)
Total	9 (8.8%)	93 (91%)	102 (100%)

In this example, matching is on the outcome (disease) variable (FES) and not on the exposure variable, as in the previous examples. We thus have information on the distribution of the exposure given the disease but not the other way around. The odds ratio is an appropriate effect measure, because the odds ratio for the association of disease given exposure is equal to the odds ratio of the association of exposure given disease. It is, however, noteworthy that the ordinary unconditional maximum likelihood is inconsistent as an estimate of the odds ratio, and it is crucial to use the conditional maximum likelihood estimate, see Section 8.9.1.

Table 8.4 can also be analyzed with tests for association, and we shall revisit this example in Section 8.5.6 (tests for association) and in Section 8.9.5 (estimation of the conditional odds ratio).

8.3 Notation and Sampling Distribution

Suppose that we have observed N pairs of dichotomous events (A and B), and let Y_1 denote the outcome of Event A, with $Y_1 = 1$ for a success and $Y_1 = 0$ for a failure. Likewise, let Y_2 denote the outcome of Event B, with $Y_2 = 1$ for a success and $Y_2 = 0$ for a failure. Each n_{ij} in Table 8.1 ($i, j = 1, 2$) corresponds to the number of pairs with outcomes $Y_1 = 2-i$ and $Y_2 = 2-j$. The $n_{11}+n_{22}$ pairs with identical outcomes are referred to as *concordant pairs*, whereas the $n_{12} + n_{21}$ pairs with unequal outcomes are referred to as *discordant pairs*. Sometimes, we use "subject" to mean a pair of observations, independent of whether the two observations originate from the same study participant or two matched participants.

Let π_{ij} denote the joint probability that $Y_1 = 2 - i$ and $Y_2 = 2 - j$, for $i, j = 1, 2$, such that we have the probability structure in Table 8.5. The joint sampling distribution for the paired 2×2 table is the multinomial distribution with probabilities $\boldsymbol{\pi} = \{\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}\}$ and N:

$$f(\mathbf{x} \,|\, \boldsymbol{\pi}, N) = \frac{N!}{x_{11}!x_{12}!x_{21}!x_{22}!}\pi_{11}^{x_{11}}\pi_{12}^{x_{12}}\pi_{21}^{x_{21}}\pi_{22}^{x_{22}}. \tag{8.1}$$

TABLE 8.5
The joint probabilities of a paired 2×2 table

	Event B		
Event A	**Success**	**Failure**	**Total**
Success	π_{11}	π_{12}	π_{1+}
Failure	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	1

Additional notation: $\boldsymbol{\pi} = \{\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}\}$

In problems with paired 2×2 data, we are usually interested in the marginal success probabilities π_{1+} and π_{+1}, that is the success probabilities for Event A and Event B. To study π_{1+} and π_{+1} is equivalent to studying π_{12} and π_{21}. The joint distribution of $\{x_{11}, x_{12}, x_{21}, x_{22}\}$ is given in Equation 8.1. As in Section 1.6, we can use the conditional approach to eliminate the nuisance parameters by conditioning on the sufficient statistics for them. When we condition on n_{11} and the total number of discordant pairs, $n_{\mathrm{d}} = n_{12} + n_{21}$, only x_{12} and $n_d - x_{12}$ remain as variables. The conditional distribution is the binomial probability distribution given by

$$f(x_{12} \,|\, \mu, n_{11}, n_{\mathrm{d}}) = \binom{n_{\mathrm{d}}}{x_{12}}\mu^{x_{12}}(1-\mu)^{n_{\mathrm{d}}-x_{12}}, \tag{8.2}$$

where $\mu = \pi_{12}/(\pi_{12}+\pi_{21})$. As we shall see in Section 8.5.3, the distribution in Equation 8.2, under the null hypothesis, will be free of unknown parameters.

The unconditional approach is to consider all possible tables with N pairs of observations. The full likelihood for the unknown parameters is given in Equation 8.1. This likelihood can be factorized into three binomial probabilities, see Lloyd (2008). One of these factors depends solely on the binomial distribution of n_{11}; however, n_{11} is sufficient for π_{11}/π_{22}. It contains no information about the discordant pairs, and we can ignore it without losing information about the association. The distribution of the discordant pairs x_{12} and x_{21} is given by the trinomial probability distribution

$$\begin{aligned} f(x_{12}, x_{21} \,|\, \pi_{12}, \pi_{21}, N) &= \\ &\frac{N!}{x_{12}!x_{21}!(N - x_{12} - x_{21})!}\pi_{12}^{x_{12}}\pi_{21}^{x_{21}}(1 - \pi_{12} - \pi_{21})^{N-x_{12}-x_{21}}. \end{aligned} \tag{8.3}$$

8.4 Statistical Models

8.4.1 Marginal Models

If we assume that the probability of a specific realization of the kth pair, $k = 1, 2, \ldots, N$ is independent of k (the subject), we have a marginal (or population-averaged) model. The probability of success for Event A (π_{1+}) and the probability of success for Event B (π_{+1}) are the marginal probabilities that $Y_1 = 1$ and $Y_2 = 1$, respectively. A marginal probability model for the relationship between the success probabilities and the events can be formulated as the generalized linear model

$$\text{link}\big[\Pr(Y_t = 1 \,|\, x_t)\big] = \alpha + \beta x_t,$$

where $t = 1, 2$ indexes the events, with $x_1 = 1$ for Event A and $x_2 = 0$ for Event B. Interest is on the parameter β, and the choice of link function determines how β is interpreted. We use the identity link to study the difference between probabilities (Section 8.6), the log link for the ratio of probabilities (Section 8.8), and the logit link for the odds ratio (Section 8.9).

8.4.2 Subject-Specific Models

In the previous section, we assumed that the probabilities were independent of the subject. When we have matched pairs data, it is often more realistic to assume that the π_{ij} vary by subject, such that the probabilities are subject specific. Interest is then on the association within the pair, conditional on the subject. We may view the data from N matched pairs as N 2×2 tables, one for each pair (Table 8.6). Collapsing over the subjects results in Table 8.1. A subject-specific model includes a subject-specific parameter (α_k):

$$\text{link}\big[\Pr(Y_t = 1\,|\,x_{kt})\big] = \alpha_k + \beta x_{kt}, \tag{8.4}$$

where $t = 1, 2$ indexes the events and $k = 1, 2, \ldots, N$ indexes the subjects. For subject number k, we have that $x_{k1} = 1$ for Event A and $x_{k2} = 0$ for Event B. The effect of event (β) on the probability of success is now conditional on the subject. Equation 8.4 is a *conditional model*, and β is a measure of the within-subject association, which is generally of greater interest than the marginal association. The practical consequences of assuming either a marginal or a subject-specific model will be explained when we consider tests for association (Section 8.5), confidence intervals for the difference between probabilities (Section 8.6), confidence intervals for the ratio of probabilities (Section 8.8), and confidence intervals for the odds ratio (Section 8.9). The subject-specific model is of special interest for the odds ratio.

TABLE 8.6
Matched pairs data displayed as N 2 × 2 tables, where the first four subjects (matched pairs) represent each of the four possible outcomes

	Event B		
Event A	**Success**	**Failure**	**Subject (pair)**
Success	1	0	1
Failure	0	0	
Success	0	1	2
Failure	0	0	
Success	0	0	3
Failure	1	0	
Success	0	0	4
Failure	0	1	
⋮	⋮		⋮
Success	n_{11k}	n_{12k}	k
Failure	n_{21k}	n_{22k}	
⋮	⋮		⋮
Success	n_{11N}	n_{12N}	N
Failure	n_{21N}	n_{22N}	

8.5 Tests for Association

8.5.1 The Null and Alternative Hypotheses

In studies of paired binomial probabilities, interest is on the marginal success probabilities π_{1+} and π_{+1}. When $\pi_{1+} = \pi_{+1}$, we also have that $\pi_{2+} = \pi_{+2}$. A test for H_0: $\pi_{1+} = \pi_{+1}$ is thus a test for *marginal homogeneity*. If we assume a subject-specific model, interest is on the *conditional independence* between Y_1 and Y_2 in the three-way $2 \times 2 \times N$ table. Testing for conditional independence (controlling for subject) is equivalent to testing for marginal homogeneity, and we shall treat the two situations as one. The following sets of hypotheses are equivalent:

$$\begin{array}{ccc} H_0: \pi_{1+} = \pi_{+1} & \text{versus} & H_{\text{A}}: \pi_{1+} \neq \pi_{+1} \\ & \Updownarrow & \\ H_0: \pi_{2+} = \pi_{+2} & \text{versus} & H_{\text{A}}: \pi_{2+} \neq \pi_{+2} \\ & \Updownarrow & \\ H_0: \pi_{12} = \pi_{21} & \text{versus} & H_{\text{A}}: \pi_{12} \neq \pi_{21} \end{array}$$

8.5.2 The McNemar Asymptotic Test

Under the null hypothesis, the expected number of success-failure pairs is equal to the expected number of failure-success pairs. Conditional on n_{11} and the total number of discordant pairs ($n_{\text{d}} = n_{12} + n_{21}$), n_{12} is binomially distributed with parameters n_{d} and μ, see Section 8.3.

Under H_0, $\mu = 1/2$, and the standard error estimate of n_{12} is

$$\widehat{\text{SE}}_0(n_{12}) = \sqrt{n_{\text{d}}\mu(1-\mu)} = \frac{1}{2}\sqrt{n_{12} + n_{21}}.$$

The McNemar asymptotic test is based on the McNemar (1947) test statistic:

$$Z_{\text{McNemar}}(\mathbf{n}) = \frac{n_{12} - \frac{1}{2}(n_{12} + n_{21})}{\widehat{\text{SE}}_0(n_{12})} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}, \tag{8.5}$$

which, under H_0, has an asymptotic standard normal distribution. Because we have estimated the standard error under the null hypothesis, Z_{McNemar} is a score statistic (see Section 1.7). We obtain P-values for the McNemar asymptotic test as

$$P\text{-value} = \Pr\Big[Z \geq \big|Z_{\text{McNemar}}(\mathbf{n})\big|\Big],$$

where Z is a standard normal variable.

The concordant pairs of observations (n_{11} and n_{22}) do not contribute to the test statistic in Equation 8.5 because the statistic is derived under the condition that the total number of discordant pairs is fixed. This might seem like a disadvantage of the method because, intuitively, the evidence of a true difference between the events should decrease when the number of identical outcomes (success-success and failure-failure) increases. It turns out, however, that the concordant pairs have negligible effect on tests of association, but they may affect measures of effect size, both in terms of estimates and precision (Agresti and Min, 2004).

Edwards (1948) proposed a continuity corrected version of the McNemar asymptotic test. The purpose of the continuity correction was to approximate the McNemar exact conditional test (see Section 8.5.3). The continuity corrected test statistic is

$$Z_{\mathrm{McNemarCC}}(\mathbf{n}) = \frac{|n_{12} - n_{21}| - 1}{\sqrt{n_{12} + n_{21}}},$$

and its approximate distribution is the standard normal distribution.

Both versions of the McNemar asymptotic test (with and without continuity correction) are undefined when $n_{12} = n_{21} = 0$.

8.5.3 The McNemar Exact Conditional Test

Recall from Section 1.9 that an exact test derives P-values by summing the (exact) probabilities of all possible tables ($\mathbf{x}$) that agree less than or equally with the null hypothesis than does the observed table ($\mathbf{n}$):

$$\text{exact } P\text{-value} = \Pr\bigl[T(\mathbf{x}) \geq T(\mathbf{n}) \,|\, H_0\bigr].$$

Here, $T()$ denotes an arbitrary test statistic, defined such that large values indicate less agreement with the null hypothesis than do small values. Under H_0, we have an unknown common success probability $\pi = \pi_{1+} = \pi_{+1}$, and this is a nuisance parameter. As explained in Section 8.3, we can eliminate the nuisance parameter by conditioning on n_{11} and the total number of discordant pairs, $n_{\mathrm{d}} = n_{12} + n_{21}$. The McNemar test statistic in Equation 8.5 can then be reduced to

$$T_{\mathrm{McNemar}}(\mathbf{n} \,|\, n_{\mathrm{d}}) = n_{12},$$

and the probability of observing x_{12}, which now completely characterizes the entire 2 × 2 table, is given by the binomial probability distribution in Equation 8.2. Under the null hypothesis, we have that $\mu = \pi_{12}/(\pi_{12} + \pi_{21}) = 1/2$, and we may simplify Equation 8.2 to

$$f(x_{12} \,|\, n_{11}, n_{\mathrm{d}}) = \binom{n_{\mathrm{d}}}{x_{12}} \left(\frac{1}{2}\right)^{n_{\mathrm{d}}}.$$

The one-sided McNemar exact conditional P-value is

$$\text{one-sided } P\text{-value} = \sum_{x_{12}=0}^{\min(n_{12}, n_{21})} f(x_{12} \,|\, n_{11}, n_{\mathrm{d}}), \tag{8.6}$$

which we multiply by two to obtain the two-sided P-value. If $n_{12} = n_{21}$, let the two-sided P-value be 1.0. The McNemar exact conditional test is sometimes called the *exact conditional binomial test*.

The McNemar exact conditional test is the uniformly most powerful unbiased test for testing H_0, see Section 1.6.

8.5.4 The McNemar Mid-*P* Test

Section 1.10 presented the mid-P approach as a way to reduce the conservatism of exact conditional methods. Here, we use the mid-P approach on the McNemar exact conditional test. To obtain the mid-P value, we subtract half the probability of the observed outcome (n_{12}) from the one-sided exact conditional P-value in Equation 8.6 and double the results:

$$\begin{aligned}\text{mid-}P\text{ value} &= 2\cdot\left[\text{one-sided }P\text{-value} - \frac{1}{2}f(n_{12}\,|\,n_{11},n_{\text{d}})\right]\\ &= \text{two-sided }P\text{-value} - f(n_{12}\,|\,n_{11},n_{\text{d}}). \qquad (8.7)\end{aligned}$$

When $n_{12} = n_{21}$, the McNemar mid-P value is

$$\text{mid-}P\text{ value} = 1 - \frac{1}{2}f(n_{12}\,|\,n_{11},n_{\text{d}}).$$

8.5.5 The McNemar Exact Unconditional Test

In Section 8.5.3, we eliminated the nuisance parameter by conditioning on n_{11} and n_d to obtain an exact conditional test. The unconditional test, on the other hand, uses information from both types of discordant pairs, x_{12} and x_{21}. The distribution of the discordant pairs is given by the trinomial probability distribution in Equation 8.3, which under the null hypothesis $\pi_{12} = \pi_{21}$ reduces to

$$\begin{aligned}f(x_{12},x_{21}\,|\,\pi,N) &=\\ &\frac{N!}{x_{12}!x_{21}!(N-x_{12}-x_{21})!}\left(\frac{\pi}{2}\right)^{x_{12}+x_{21}}(1-\pi)^{N-x_{12}-x_{21}},\end{aligned}$$

where $\pi = \pi_{12} + \pi_{21}$ is the probability of a discordant pair (the nuisance parameter). The exact unconditional approach is to eliminate the nuisance parameter by maximization over the domain of π:

$$P\text{-value} = \max_{0\leq\pi\leq 1}\left\{\sum_{\Omega(\mathbf{x}|N)} I\big[T(\mathbf{x}) \geq T(\mathbf{n})\big]\cdot f(x_{12},x_{21}\,|\,\pi,N)\right\}, \qquad (8.8)$$

where $\Omega(\mathbf{x}|N)$ denotes the set of all tables with N observations, $I()$ is the indicator function, and $T()$ is a test statistic, defined such that tables with large values of T agree less with the null hypothesis than do tables with small values of T.

The Berger and Boos procedure (see Section 4.4.7) can be used to reduce the nuisance parameter space:

$$P\text{-value} = \max_{\pi \in C_\gamma} \left\{ \sum_{\Omega(\mathbf{x}|N)} I\big[T(\mathbf{x}) \geq T(\mathbf{n})\big] \cdot f(x_{12}, x_{21} \,|\, \pi, N) \right\} + \gamma,$$

where C_γ is a $100(1-\gamma)\%$ confidence interval for π, and γ is a small value, for instance $\gamma = 0.0001$.

The first to propose an exact unconditional test for the paired 2×2 table was Suissa and Shuster (1991) who used the McNemar test statistic

$$T_{\text{McNemar}}(\mathbf{n}) = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}$$

in Equation 8.8. We shall refer to this test as the McNemar exact unconditional test.

A reasonable alternative to the McNemar statistic is the likelihood ratio statistic. Lloyd (2008) compared exact unconditional tests based on the two statistics and found no practical differences between them.

8.5.6 Examples

Airway Hyper-Responsiveness Status before and after Stem Cell Transplantation (Table 8.2)

The null hypothesis of interest is that the probability of AHR before SCT is equal to the probability of AHR after SCT: H_0: $\pi_{1+} = \pi_{+1}$. We test this against the alternative hypothesis: H_A: $\pi_{1+} \neq \pi_{+1}$. The observed proportions of AHR are $\hat{\pi}_{1+} = 2/21 = 0.095$ (before SCT) and $\hat{\pi}_{1+} = 8/21 = 0.38$ (after SCT). The observed value of the McNemar test statistic (Equation 8.5) is

$$Z_{\text{McNemar}}(\mathbf{n}) = \frac{1-7}{\sqrt{1+7}} = -2.12.$$

To obtain the P-value for the asymptotic McNemar test, we can refer -2.12 to the standard normal distribution or we may take the square of the observed value, $-2.12^2 = 4.50$, and refer that to the chi-squared distribution with one degree of freedom. The resulting P-value will be the same no matter the method. Here, we use the chi-squared distribution, which is consistent with the way we calculated the Pearson chi-squared test for the unpaired 2×2 table in Section 4.4. Thus, the P-value for the asymptotic McNemar test is

$$P\text{-value} = \Pr(\chi^2_1 \geq 4.50) = 0.0339.$$

If we use the Edwards continuity correction, the test statistic is

$Z_{\text{McNemarCC}}(\mathbf{n}) = 1.77$. The P-value then is $P = 0.0771$, which is quite a bit higher than the P-value of the uncorrected test.

The sample size is small in this example, and we may question whether it is appropriate to use an asymptotic test. For the unpaired 2×2 table in Chapter 4, we used Cochran's criterion (see page 100) as a rule of thumb to decide if it was safe to use asymptotic tests. There is no Cochran's criterion for the paired 2×2 table, and there are no other obvious criteria for deciding when the sample size is sufficiently large to allow for asymptotic tests. An evaluation of the tests will be carried out in Section 8.5.7, and we shall gain more insight into the performances and scopes of application of the tests. Here, we proceed with the calculation of the tests for the data in Table 8.2 and leave the recommendations of which test to use in which situation to Section 8.10.

The McNemar exact conditional test reduces the sample space to tables that have the same number of discordant pairs as the observed table, $n_{\text{d}} = 1 + 7 = 8$. Thus, nine tables are possible; however, because the minimum value of n_{12} and n_{21} is 1, only two probabilities are needed to calculate the one-sided P-value according to Equation 8.6. The calculations are shown in Table 8.7. To obtain the two-sided P-value, we double the one-sided P-value and get $P = 0.0703$. This value is similar to the P-value for the asymptotic McNemar test with continuity correction. As this example illustrates, the McNemar exact conditional test is afflicted by discreteness; only two probabilities went into the calculations of the P-value. As with other exact conditional methods, the result is conservative inference.

TABLE 8.7
Quantities involved in the calculation of the one-sided P-value of the McNemar exact conditional test on the data in Table 8.2

x_{12}	$f(x_{12}\,\vert\,n_{\text{d}})$	Cumulative probability
0	0.0039	0.0039
1	0.0313	0.0352

We now turn to the McNemar mid-P test. To calculate it (see Equation 8.7), we need the P-value from the McNemar exact conditional test ($P = 0.0703$) and the probability of the observed outcome. The latter is shown in the second column of the last row in Table 8.7. The McNemar mid-P value then is mid-$P = 0.0703 - 0.0313 = 0.0391$.

The McNemar exact unconditional test includes all possible tables with $N = 21$ pairs. There are 253 ways of distributing the 21 pairs to the cell counts x_{12} and x_{21}, and for 112 of these tables, the McNemar test statistic is equal to or greater than that for the observed table. So, for each value of π (the nuisance parameter), a P-value is obtained as the sum of 112 probabilities (Equation 8.8). Figure 8.1 shows the P-value as a function of π. The exact unconditional P-value without the Berger and Boos procedure is taken as the maximum of this function across the entire nuisance parameter space, which results in $P = 0.0353$. To apply the Berger and Boos procedure with $\gamma = 0.0001$, we calculate a 99.99% confidence interval for π with the Clopper-Pearson exact interval: $C_\gamma = (0.070, 0.79)$. The exact unconditional P-value is now the maximum P-value over C_γ, to which we add the value of γ. The C_γ interval is indicated as the shaded area in Figure 8.8. The maximum P-value over C_γ is the same as the maximum P-value over $(0, 1)$. The exact unconditional P-value with Berger and Boos procedure is therefore $P = 0.0353 + 0.0001 = 0.0354$.

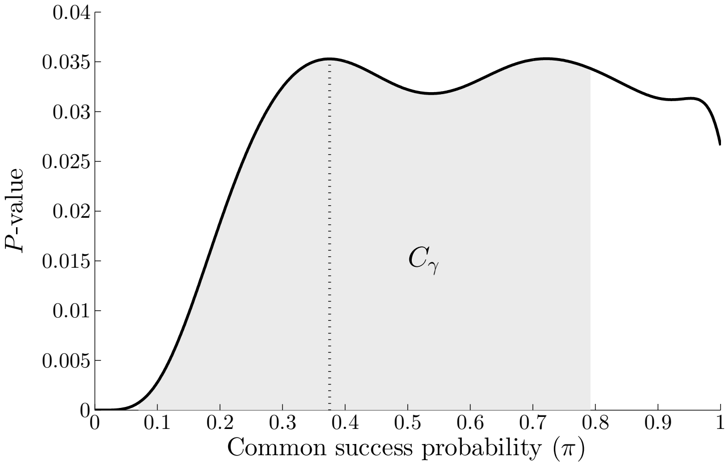

FIGURE 8.1
P-value as a function of the common success probability (π) for the McNemar exact unconditional test on the data in Table 8.2. The dotted vertical line shows the maximum P-value and its corresponding value of π. The shaded area indicates the C_γ interval.

Table 8.8 summarizes the results.

TABLE 8.8
Results of six tests for association on the data in Table 8.2

Test	***P*-value**
McNemar asymptotic	0.0339
McNemar asymptotic with continuity correction	0.0771
McNemar exact conditional	0.0703
McNemar mid-P	0.0391
McNemar exact unconditional	0.0353
McNemar exact unconditional*	0.0354

*Calculated with Berger and Boos procedure ($\gamma = 0.0001$)

Complete Response before and after Consolidation Therapy (Table 8.3)

The previous example showed that quite different results were obtained for the six tests on a table with 21 pairs of observations. The study by Cavo et al. (2012)—with results shown in Table 8.3—is similar, in that each patient in the study is measured before and after treatment; however, the sample size is considerably larger, with 161 pairs of observations (patients). The null hypothesis is that the probability of complete response is the same before (π_{1+}) and after (π_{+1}) consolidation therapy: H_0: $\pi_{1+} = \pi_{+1}$. The two-sided alternative is H_A: $\pi_{1+} \neq \pi_{+1}$. We do not give the details of the computations of the tests but show the results in Table 8.9. Even with this medium-to-large sample size, we obtain noticeable different results. The P-values of the asymptotic, mid-P, and exact unconditional tests are similar, whereas the P-values of the asymptotic test with continuity correction and the exact conditional test are considerably higher. The exact conditional test is still a victim of discreteness: only seven ($= \min(n_{12}, n_{21}) + 1$, see Equation 8.6) probabilities are used to compute the P-value. In contrast, the P-value of the exact unconditional test is a sum of 10 290 probabilities.

TABLE 8.9
Results of six tests for association on the data in Table 8.3

Test	***P*-value**
McNemar asymptotic	0.0330
McNemar asymptotic with continuity correction	0.0550
McNemar exact conditional	0.0525
McNemar mid-P	0.0347
McNemar exact unconditional	0.0342
McNemar exact unconditional*	0.0341

*Calculated with Berger and Boos procedure ($\gamma = 0.0001$)

The Association between Floppy Eyelid Syndrome and Obstructive Sleep Apnea-Hypopnea Syndrome (Table 8.4)

The study by Ezra et al. (2010), summarized in Table 8.4, is a matched case-control study. Each pair of outcomes consists of the exposure status (OSAHS) of one case (a patient with FES) and the exposure status of one matching control (a patient without FES). The null hypothesis is that the proportion of exposed cases is equal to the proportion of exposed controls: H_0: $\pi_{1+} = \pi_{+1}$ versus H_A: $\pi_{1+} \neq \pi_{+1}$. This hypothesis setup is the same as in the other two examples. Testing for association in matched case-control studies is thus identical to testing for association in cohort studies where each participant is measured twice.

The observed proportion of exposed cases is $\hat{\pi}_{1+} = 32/102 = 0.31$, and the observed proportion of exposed controls is $\hat{\pi}_{1+} = 9/102 = 0.088$. All the six tests for association give $P < 0.00011$. A strong association between FES and OSAHS is indicated; however, in a matched case-control study, it is more appropriate to study the within-subject association, for which the subject-specific model in Equation 8.4 can be used. In Section 8.9.5, we estimate the conditional odds ratio and its confidence interval to quantify the within-subject association.

8.5.7 Evaluation of Tests

Evaluation Criteria

We evaluate tests for association by calculating their actual significance levels and power. The actual significance level and power depend on the probabilities π_{11}, π_{12}, π_{21}, and π_{22}, the number of pairs (N), and the nominal significance level α. Because the parameters of interest are the probabilities of success for Event A (π_{1+}) and Event B (π_{+1}), we reparameterize $\{\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}\}$ into the equivalent parameter set $\{\pi_{1+}, \pi_{+1}, \theta\}$, where $\theta = \pi_{11}\pi_{22}/\pi_{12}\pi_{21}$. For each parameter space point—any realization of $\{\pi_{1+}, \pi_{+1}, \theta, N, \alpha\}$—we use complete enumeration to calculate the actual significance level (ASL) if $\pi_{1+} = \pi_{+1} = \pi$, or power if $\pi_{1+} \neq \pi_{+1}$. That is, we perform the tests on all possible tables with N pairs and add the probability of all tables with P-values less than the nominal significance level:

$$\begin{aligned}\mathrm{ASL}(\pi, \theta, N, \alpha) \ &= \\ &\sum_{x_{11}=0}^{N} \sum_{x_{12}=0}^{N-x_{11}} \sum_{x_{21}=0}^{N-x_{11}-x_{12}} I\big[P(\mathbf{x}) \leq \alpha\big] \cdot f(\mathbf{x}\,|\,\pi, \theta, N)\end{aligned}$$

and

$$\begin{aligned}\mathrm{Power}(\pi_{1+}, \pi_{+1}, \theta, N, \alpha) \ &= \\ &\sum_{x_{11}=0}^{N} \sum_{x_{12}=0}^{N-x_{11}} \sum_{x_{21}=0}^{N-x_{11}-x_{12}} I\big[P(\mathbf{x}) \leq \alpha\big] \cdot f(\mathbf{x}\,|\,\pi_{1+}, \pi_{+1}, \theta, N),\end{aligned}$$

where $I()$ is the indicator function, $P(\mathbf{x})$ is the P-value for a test on $\mathbf{x} = \{x_{11}, x_{12}, x_{21}, N - x_{11} - x_{12} - x_{21}\}$, and $f()$ is the multinomial probability distribution (Equation 8.1).

Evaluation of Actual Significance Level

By fixing the number of matched pairs (N) and the parameter θ, and setting $\alpha = 0.05$, we can plot the actual significance level as a function of the common success probability (π). Figure 8.2 shows the three non-exact McNemar tests for $N = 50$ and $\theta = 2.0$. The McNemar asymptotic test violates the nominal significance level for nearly half the range of π; however, the violations are small: the maximum actual significance level is 5.3%. The McNemar mid-P test has actual significance levels close to but below the nominal level. The McNemar asymptotic test with continuity correction, on the other hand, is very conservative: it has significance levels below 3% for all the shown parameter space points. Later in this section, we shall see that the McNemar exact conditional test performs similarly.

The results in Figure 8.2 are typical for a wide range of situations. In an evaluation study covering almost 10 000 scenarios (Fagerland et al., 2013), the McNemar asymptotic test frequently violated the nominal level, but its actual significance level was never above 5.37%. The McNemar mid-P test did not violate the nominal level in any of the almost 10 000 scenarios. This latter result is unusual: mid-P tests (and confidence intervals) usually exhibit occasional but small infringements on the nominal level.

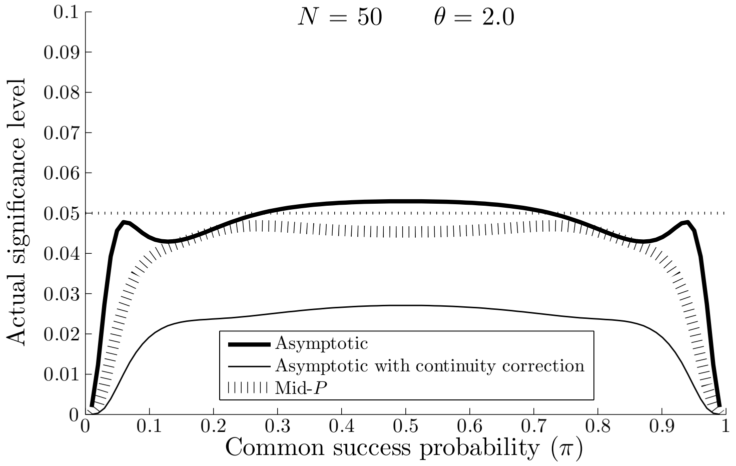

FIGURE 8.2
Actual significance levels of three McNemar tests

We now turn to three exact tests: the McNemar exact conditional test and the McNemar exact unconditional tests with ($\gamma = 0.0001$) and without ($\gamma = 0$) the Berger and Boos procedure. The situation in Figure 8.3, which shows the actual significance levels of the exact tests for $N = 20$ and $\theta = 2.0$, is both typical and atypical. The typical results are that the exact conditional test is overly conservative, here with an actual significance level below 2%, and that the exact unconditional tests perform much better. The atypical result is the large difference between the exact unconditional tests with and without the Berger and Boos procedure. In most of the situations we consider in this book, the Berger and Boos procedure may have a noticeable impact on P-values and confidence intervals for particular data (see, for instance, Table 4.15); however, we rarely see such an obvious improvement as in Figure 8.3. This large improvement in performance for the McNemar exact unconditional test seems to be confined to small sample sizes ($N < 25$). For larger sample sizes, there is no noticeable difference in actual significance levels between the tests with and without the Berger and Boos procedure. Figure 8.4 shows an example with $N = 50$. Note that the exact conditional test is still very conservative. It performs similarly to the McNemar asymptotic test with continuity correction, which can be seen in Figure 8.2.

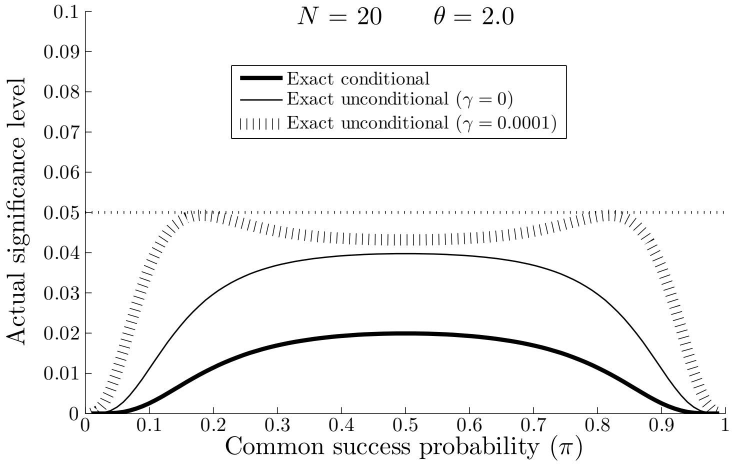

FIGURE 8.3
Actual significance levels of three exact McNemar tests

The McNemar asymptotic test (without continuity correction) also performs well for small sample sizes. Figure 8.5 shows the actual significance levels of the asymptotic test, the mid-P test, and the exact unconditional test with $\gamma = 0.0001$ for a total of only 15 matched pairs. The maximum actual significance level of the asymptotic test in this case is 5.03%. This performance of the standard asymptotic test is excellent and surprising: we are used to the fact that simple asymptotic tests (and confidence intervals) produce substantial violations of the nominal level in small samples. This is certainly the case with the Pearson chi-squared test for the unpaired 2 × 2 table; see, for instance, Figure 4.4 and the discussions in Section 4.4.9.

Evaluation of Power

In the preceding evaluations of actual significance level, we plotted the actual significance level as a function of the common success probability and kept the sample size fixed. To evaluate power, it is more instructive to treat the success probabilities π_{1+} and π_{+1} as fixed, and consider power as a function of the sample size. In Figure 8.6, we have fixed $\pi_{1+} = 0.25$, $\pi_{+1} = 0.5$, and $\theta = 2.0$.

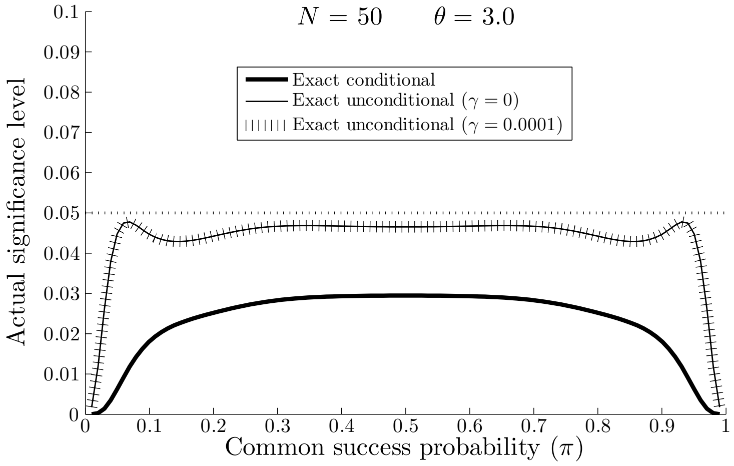

FIGURE 8.4
Actual significance levels of three exact McNemar tests

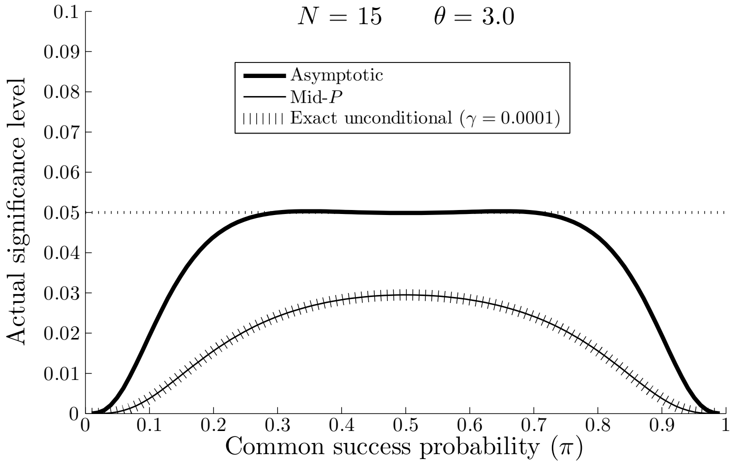

FIGURE 8.5
Actual significance levels of three McNemar tests

The plot shows how the probability (power) to detect a difference in success probabilities of 25% versus 50% depends on the number of matched pairs (N). We have restricted N to values between 40 and 80 so that the power of most of the tests is between 65% and 95%, which should be the most interesting range of power for most practical situations. We observe several interesting differences between the tests. The asymptotic test is clearly the most powerful test, followed by the mid-P test and the exact unconditional test (without the Berger and Boos procedure). The powers of the exact conditional test and the asymptotic test with continuity correction trail that of the other tests considerably. If we were to design a study with an 80% chance of detecting $\pi_{1+} = 0.25$ versus $\pi_{+1} = 0.5$, a plan to use the asymptotic test would require nine or ten fewer matched pairs than a plan to use the exact conditional test.

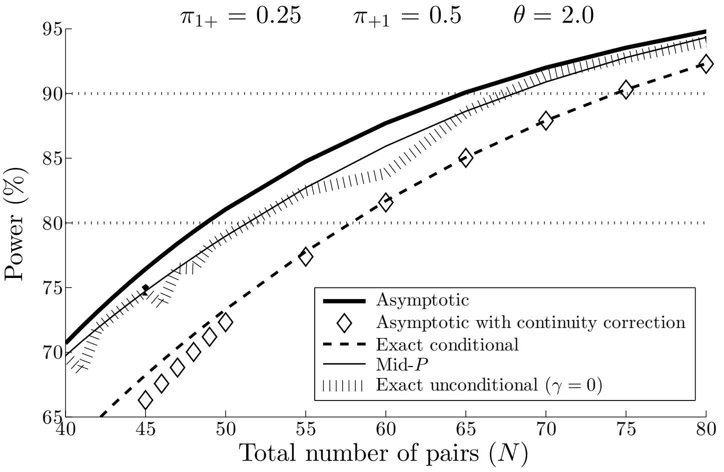

FIGURE 8.6
Power of five McNemar tests

We observed a noticeable improvement in actual significance levels for the McNemar exact unconditional test when the Berger and Boos procedure with $\gamma = 0.0001$ was used. Figure 8.7 shows that the Berger and Boos procedure also evokes a small benefit in power. This benefit is related to the paradoxical result that the exact unconditional test without Berger and Boos procedure sometimes loses power when the number of matched pairs is increased by one. We have yet to experience this unwanted behavior when the Berger and Boos procedure is in use.

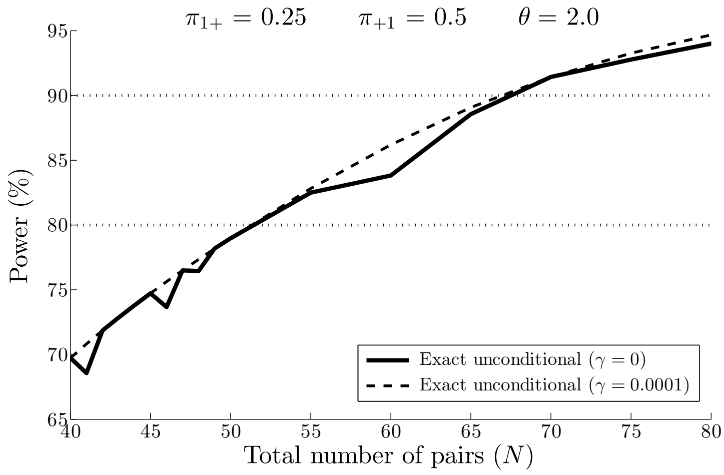

FIGURE 8.7
Power of the McNemar exact unconditional test with ($\gamma = 0.0001$) and without ($\gamma = 0$) the Berger and Boos procedure

8.6 Confidence Intervals for the Difference between Probabilities

8.6.1 Introduction and Estimation

The difference between the marginal probabilities (the success probabilities) is a natural effect measure for paired randomized trials and paired longitudinal studies. The canonical link function for the generalized linear models in Section 8.4 is the linear link. The subject-specific model is given by

$$\Pr(Y_t = 1 \,|\, x_{kt}) = \alpha_k + \beta x_{kt},$$

for $t = 1, 2$ and $k = 1, 2, \ldots, N$. For the kth subject, $x_{k1} = 1$ for Event A and $x_{k2} = 0$ for Event B. We have that

$$\Pr(Y_1 = 1 \,|\, x_{k1}) - \Pr(Y_2 = 1 \,|\, x_{k2}) = \beta.$$

For each subject, β is the difference between the probabilities of Event A and Event B. By summation, we see that β is the difference between the marginal probabilities. If we assume a marginal model instead of a subject-specific model, we drop the subscript k from x and α in the preceding equations and obtain the same result; thus, the marginal association is the same as the within-subject association.

We define the difference between probabilities as

$$\Delta = \pi_{1+} - \pi_{+1}.$$

The maximum likelihood estimate of Δ is given by the sample proportions:

$$\hat{\Delta} = \hat{\pi}_{1+} - \hat{\pi}_{+1} = \frac{n_{1+} - n_{+1}}{N} = \frac{n_{12} - n_{21}}{N}.$$

Sections 8.6.2–8.6.5 present different confidence interval methods for Δ. In Section 8.6.6, we apply the methods to the examples presented in Section 8.2. The methods are evaluated in Section 8.6.7, and Section 8.10 provides recommendations.

8.6.2 Wald Intervals

The (asymptotic) *Wald interval* for Δ is the most used interval for paired binomial probabilities. It is defined as:

$$\hat{\Delta} \pm \frac{z_{\alpha/2}}{N}\sqrt{n_{12} + n_{21} - \frac{(n_{12} - n_{21})^2}{N}}.$$

When $n_{12} = n_{21} = 0$, the zero-width interval $(0, 0)$ is produced.

A continuity correction—similar to the one for the asymptotic McNemar test in Section 8.5.2—can be applied to the Wald interval. We call the resulting interval the *Wald interval with continuity correction*:

$$\hat{\Delta} \pm \frac{z_{\alpha/2}}{N}\sqrt{n_{12} + n_{21} - \frac{(|n_{12} - n_{21}| - 1)^2}{N}}.$$

As with the Wald interval, the Wald interval with continuity correction gives the interval $(0, 0)$ when $n_{12} = n_{21} = 0$.

Agresti and Min (2005b) investigate the effects of adding pseudo-frequencies to the observed cells in Table 8.1 before calculating the Wald interval. They find that adding 1/2 to each cell improves performance:

$$\frac{\tilde{n}_{12} - \tilde{n}_{21}}{\tilde{N}} \pm \frac{z_{\alpha/2}}{\tilde{N}}\sqrt{\tilde{n}_{12} + \tilde{n}_{21} - \frac{(\tilde{n}_{12} - \tilde{n}_{21})^2}{\tilde{N}}},$$

where $\tilde{n}_{12} = n_{12} + 1/2$, $\tilde{n}_{21} = n_{12} + 1/2$, and $\tilde{N} = N + 2$. We refer to this interval as the *Wald interval with Agresti-Min adjustment.*

Another simple adjustment to the Wald interval was proposed by Bonett and Price (2012). First, calculate the *Laplace estimates* $\tilde{\pi}_{12} = (n_{12}{+}1)/(N{+}2)$ and $\tilde{\pi}_{21} = (n_{21} + 1)/(N + 2)$. Then, calculate a confidence interval for Δ as

$$\tilde{\pi}_{12} - \tilde{\pi}_{21} \pm z_{\alpha/2}\sqrt{\frac{\tilde{\pi}_{12} + \tilde{\pi}_{21} - (\tilde{\pi}_{12} - \tilde{\pi}_{21})^2}{N + 2}}.$$

We refer to this interval as the *Wald interval with Bonett-Price adjustment.*

Neither of the four versions of the Wald interval is guaranteed to respect the $[-1, 1]$ boundary of Δ. When overshoot happens, the usual approach is to truncate the overshooting limit to 1 or -1. The disadvantage of this approach is that the interval can be artificially narrow, thus underestimating the uncertainty in the data.

8.6.3 The Newcombe Square-And-Add Interval (MOVER Wilson Score)

In Chapter 4, we encountered several applications of the square-and-add approach—also called the *method of variance estimates recovery (MOVER)*—for the construction of confidence intervals for different effect measures for the unpaired 2×2 table. Recall that MOVER is a general method that constructs a confidence interval for the difference of two parameters, $\theta_1 - \theta_2$, by combining two separate confidence intervals for θ_1 and θ_2. Let (l_1, u_1) denote the interval for θ_1, and let (l_2, u_2) denote the interval for θ_2. For paired binomial data, the confidence limits for $\theta_1 - \theta_2$ are

$$L^* = \hat{\theta}_1 - \hat{\theta}_2 - \sqrt{\big(\hat{\theta}_1 - l_1\big)^2 + \big(u_2 - \hat{\theta}_2\big)^2 - 2\psi\big(\hat{\theta}_1 - l_1\big)\big(u_2 - \hat{\theta}_2\big)} \tag{8.9}$$

and

$$U^* = \hat{\theta}_1 - \hat{\theta}_2 + \sqrt{\big(u_1 - \hat{\theta}_1\big)^2 + \big(\hat{\theta}_2 - l_2\big)^2 - 2\psi\big(u_1 - \hat{\theta}_1\big)\big(\hat{\theta}_2 - l_2\big)}, \tag{8.10}$$

where $\hat{\theta}_1$ and $\hat{\theta}_2$ are estimates of θ_1 and θ_2, and $\psi = \widehat{\text{corr}}\big(\hat{\theta}_1, \hat{\theta}_2\big)$ is an estimate of the correlation coefficient between $\hat{\theta}_1$ and $\hat{\theta}_2$. A derivation of and motivation for Equations 8.9 and 8.10 can be found in Newcombe (1998a) and Tang et al. (2010). Tang et al. also provide examples of early applications of the method.

Equations 8.9 and 8.10 give rise to many different confidence intervals. Each choice of confidence interval method for the binomial parameter—to calculate (l_1, u_1) and (l_2, u_2)—and each choice of estimate for ψ leads to a distinct method. Newcombe (1998a) proposed and evaluated several different square-and-add intervals for $\Delta = \pi_{1+} - \pi_{+1}$. Here, we consider the best performing of these intervals, which is based on Wilson score intervals (see Section 2.4.3) for the binomial probability π_{1+}:

$$(l_1, u_1) = \frac{2n_{1+} + z_{\alpha/2}^2 \mp z_{\alpha/2}\sqrt{z_{\alpha/2}^2 + 4n_{1+}\Big(1 - \frac{n_{1+}}{N}\Big)}}{2\Big(N + z_{\alpha/2}^2\Big)} \tag{8.11}$$

and π_{+1}:

$$(l_2, u_2) = \frac{2n_{+1} + z_{\alpha/2}^2 \mp z_{\alpha/2}\sqrt{z_{\alpha/2}^2 + 4n_{+1}\Big(1 - \frac{n_{+1}}{N}\Big)}}{2\Big(N + z_{\alpha/2}^2\Big)}. \tag{8.12}$$

If any of the marginal sums $(n_{1+}, n_{2+}, n_{+1}, n_{+2})$ is zero, set $\psi = 0$. Otherwise, let $A = n_{11}n_{22} - n_{12}n_{21}$ and compute ψ as

$$\psi = \begin{cases} (A - N/2)/\sqrt{n_{1+}n_{2+}n_{+1}n_{+2}} & \text{if} \quad A > N/2, \\ 0 & \text{if} \quad 0 \le A \le N/2, \\ A/\sqrt{n_{1+}n_{2+}n_{+1}n_{+2}} & \text{if} \quad A < 0. \end{cases}$$

The lower (L) and upper (U) limits of the Newcombe square-and-add interval for Δ are given by

$$L = \hat{\Delta} - \sqrt{\big(\hat{\pi}_{1+} - l_1\big)^2 + \big(u_2 - \hat{\pi}_{+1}\big)^2 - 2\psi\big(\hat{\pi}_{1+} - l_1\big)\big(u_2 - \hat{\pi}_{+1}\big)} \qquad (8.13)$$

and

$$U = \hat{\Delta} + \sqrt{\big(\hat{\pi}_{+1} - l_2\big)^2 + \big(u_1 - \hat{\pi}_{1+}\big)^2 - 2\psi\big(\hat{\pi}_{+1} - l_2\big)\big(u_1 - \hat{\pi}_{1+}\big)}, \qquad (8.14)$$

where $\hat{\pi}_{1+} = n_{1+}/N$ and $\hat{\pi}_{+1} = n_{+1}/N$.

8.6.4 The Tango Asymptotic Score Interval

Tango (1998) developed an asymptotic score interval for the difference between paired probabilities based on inverting two asymptotic $\alpha/2$ level score tests (the tail method). For a specified value $\Delta_0 \in [-1, 1]$, the score statistic is

$$T_{\text{score}}(\mathbf{n}\,|\,\Delta_0) = \frac{n_{12} - n_{21} - N\Delta_0}{\sqrt{N\big[2\tilde{p}_{21} + \Delta_0(1 - \Delta_0)\big]}}, \qquad (8.15)$$

where $\mathbf{n} = \{n_{11}, n_{12}, n_{21}, n_{22}\}$, as usual, denotes the observed table and $\tilde{p}_{21}$ is the maximum likelihood estimate of π_{21}, constrained to $\pi_{1+} - \pi_{+1} = \Delta_0$, given as

$$\tilde{p}_{21} = \frac{\sqrt{B^2 - 4AC} - B}{2A},$$

where $A = 2N$, $B = -n_{12} - n_{21} + (2N - n_{12} + n_{21})\Delta_0$, and $C = -n_{21}\Delta_0(1 - \Delta_0)$. The Tango asymptotic score interval (L, U) for Δ is obtained by solving

$$T_{\text{score}}(\mathbf{n}\,|\,L) = z_{\alpha/2}$$

and

$$T_{\text{score}}(\mathbf{n}\,|\,U) = -z_{\alpha/2}$$

iteratively, for instance, with the secant or bisection method. It is possible—although tricky—to derive closed-form expressions for L and U (Newcombe, 2013, Chapter 8). An Excel implementation is given as web-based supplementary material to Newcombe (2013).

8.6.5 The Sidik Exact Unconditional Interval

The score statistic in Equation 8.15 can also be used to derive exact unconditional tests, which in turn may be inverted to obtain exact unconditional confidence intervals for Δ. There are two main approaches: we can invert two one-sided $\alpha/2$ level tests or one two-sided α level test. The first approach (the tail method) ensures that the non-coverage in each tail does not exceed $\alpha/2$. The limits from such an interval are thereby consistent with the results of the corresponding exact unconditional one-sided test. An interval based on inverting one two-sided test, on the other hand, guarantees that the overall non-coverage does not exceed α but makes no claims about the left and right tails. It is consistent with the results of the corresponding exact unconditional two-sided test.

Here, we consider an interval first proposed by Hsueh et al. (2001), which inverts two one-sided exact score tests. We have two nuisance parameters: π_{12} and π_{21}. The version described in the following is due to Sidik (2003), who showed how to simplify the computations of the interval by reducing the dimensions of the nuisance parameter space from two to one.

Let $\mathbf{x} = \{x_{11}, x_{12}, x_{21}, x_{22}\}$ denote an arbitrary outcome with N pairs. The probability of observing $\mathbf{x}$ is given by the trinomial probability distribution:

$$f(x_{12}, x_{21} \,|\, \pi_{12}, \Delta_0, N) \;=\; \frac{N!}{x_{12}!x_{21}!(N - x_{12} - x_{21})!}\pi_{12}^{x_{12}}(\pi_{12} - \Delta_0)^{x_{21}}(1 - 2\pi_{12} + \Delta_0)^{N - x_{12} - x_{21}},$$

where $\Delta_0 = \pi_{12} - \pi_{21}$. This is a reparameterized version of Equation 8.3.

As shown in Sidik (2003), we can eliminate the remaining nuisance parameter (π_{12}) by taking the maximum value over the domain $D(\Delta_0) : \{0 \leq \pi_{12} \leq (1 + \Delta_0)/2\}$. The lower ($L$) and upper ($U$) confidence limits of the Sidik exact unconditional interval for Δ are the solutions—calculated iteratively—of the two equations:

$$\max_{\pi_{12} \in D(\Delta_0)} \left\{ \sum_{\Omega(\mathbf{x}|\Delta_0, N)} I\big[T(\mathbf{x} \,|\, L) \geq T(\mathbf{n} \,|\, L)\big] \cdot f(x_{12}, x_{21} \,|\, \pi_{12}, L, N) \right\} = \alpha/2 \tag{8.16}$$

and

$$\max_{\pi_{12} \in D(\Delta_0)} \left\{ \sum_{\Omega(\mathbf{x}|\Delta_0, N)} I\big[T(\mathbf{x} \,|\, U) \leq T(\mathbf{n} \,|\, U)\big] \cdot f(x_{12}, x_{21} \,|\, \pi_{12}, U, N) \right\} = \alpha/2, \tag{8.17}$$

where $T()$ is the score statistic in Equation 8.15.

The Berger and Boos procedure (Section 4.4.7) may be used to reduce the domain of π_{12} for the maximizations in Equations 8.16 and 8.17. This is not as straightforward as in the previous cases in this book, because the Berger and Boos procedure must be applied to the two-dimensional nuisance parameter space defined by π_{12} and π_{21}. Sidik (2003) has shown how to define a confidence interval, C_γ, for π_{12}, and that taking the maximum value over C_γ is equivalent to taking the maximum value over the two-dimensional confidence set for π_{12} and π_{21}. Let L_{CP} and U_{CP} denote a $100(1-\gamma)\%$ Clopper-Pearson exact interval (see Section 2.4.7) for $2\pi_{12} - \Delta_0$ based on the assumption that $x_{12}+x_{21}$ is binomially distributed with parameters N and $2\pi_{12}-\Delta_0$. Then, the lower limit of C_γ is $(L_{\mathrm{CP}} + \Delta_0)/2$, and the upper limit of C_γ is $(U_{\mathrm{CP}} + \Delta_0)/2$. The Sidik exact unconditional interval for Δ with Berger and Boos procedure is obtained by substituting Equations 8.16 and 8.17 with

$$\max_{\pi_{12}\in C_\gamma}\left\{\sum_{\Omega(\mathbf{x}|\Delta_0,N)} I\big[T(\mathbf{x}\,|\,L) \geq T(\mathbf{n}\,|\,L)\big]\cdot f(x_{12},x_{21}\,|\,\pi_{12},L,N)\right\} + \gamma = \alpha/2$$

and

$$\max_{\pi_{12}\in C_\gamma}\left\{\sum_{\Omega(\mathbf{x}|\Delta_0,N)} I\big[T(\mathbf{x}\,|\,U) \leq T(\mathbf{n}\,|\,U)\big]\cdot f(x_{12},x_{21}\,|\,\pi_{12},U,N)\right\} + \gamma = \alpha/2.$$

We suggest that $\gamma = 0.0001$ is used.

8.6.6 Examples

Airway Hyper-Responsiveness Status before and after Stem Cell Transplantation (Table 8.2)

The two parameters of interest are the probability of AHR before SCT (π_{1+}) and the probability of AHR after SCT (π_{+1}). The estimated probabilities are $\hat{\pi}_{1+} = 1/21 = 0.095$ and $\hat{\pi}_{+1} = 7/21 = 0.38$. The maximum likelihood estimate of the difference between the probabilities is

$$\hat{\Delta} = \frac{n_{12} - n_{21}}{N} = \frac{1-7}{21} = -0.286.$$

Table 8.10 gives eight different 95% confidence intervals for Δ. We do not go into the computational details of the methods here but refer the reader to Section 4.5.7, where we show how to calculate some similar confidence intervals for the difference between independent probabilities. The sample size is small in this example—the total number of pairs is only 21—and we would expect the different interval methods to vary considerably, as we observed with the tests for association in Table 8.8. The intervals in Table 8.10 are, however, quite similar, although the Sidik exact unconditional interval is slightly wider than the others. Interestingly, neither of the intervals contains zero, the null value. There is thus no interval for the difference between probabilities that gives results that agree with the McNemar exact conditional test ($P = 0.070$) or the McNemar asymptotic test with continuity correction ($P = 0.077$) for these data. All the intervals in Table 8.10 agree well with the McNemar asymptotic ($P = 0.034$), McNemar mid-P ($P = 0.039$), and McNemar exact unconditional ($P = 0.035$) tests.

TABLE 8.10
95% confidence intervals for the difference between probabilities ($\hat{\Delta} = -0.286$) based on the data in Table 8.2

	Confidence limits		
Interval	**Lower**	**Upper**	**Width**
Wald	-0.520	-0.052	0.468
Wald with continuity correction	-0.529	-0.042	0.487
Wald with Agresti-Min adjustment	-0.493	-0.029	0.465
Wald with Bonett-Price adjustment	-0.508	-0.013	0.495
Newcombe square-and-add	-0.507	-0.026	0.481
Tango asymptotic score	-0.517	-0.026	0.491
Sidik exact unconditional	-0.537	-0.020	0.517
Sidik exact unconditional*	-0.532	-0.020	0.512

*Calculated with Berger and Boos procedure ($\gamma = 0.0001$)

Complete Response before and after Consolidation Therapy (Table 8.3)

The aim of this example is to estimate the difference between the probabilities of complete response before and after consolidation therapy for patients with multiple myeloma. The sample proportion of patients with complete response before consolidation therapy is $\hat{\pi}_{1+} = 65/161 = 0.404$. After consolidation therapy, the sample proportion is $\hat{\pi}_{1+} = 75/161 = 0.466$. We estimate the difference between probabilities as

$$\hat{\Delta} = \frac{n_{12} - n_{21}}{N} = \frac{6 - 16}{161} = -0.0621.$$

Table 8.11 shows eight different 95% confidence intervals for Δ. Only minor differences between the methods can be observed, whereas the tests for association in Table 8.9 gave considerably larger variation in results for these data.

8.6.7 Evaluation of Intervals

Evaluation Criteria

We use three indices of performance to evaluate confidence intervals: coverage probability, width, and location (see Section 1.4). In the following, we show how coverage, width, and location for the difference between paired probabilities can be calculated exactly with complete enumeration.

The coverage probability, width, and location depend on the probabilities π_{11}, π_{12}, π_{21}, and π_{22}, and the number of pairs (N). Because the parameters of interest are the probabilities of success for Event A (π_{1+}) and Event B (π_{+1}), we reparameterize $\{\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}\}$ into the equivalent parameter set $\{\pi_{1+}, \pi_{+1}, \theta\}$, where $\theta = \pi_{11}\pi_{22}/\pi_{12}\pi_{21}$. The exact coverage probability

TABLE 8.11
95% confidence intervals for the difference between probabilities ($\hat{\Delta} = -0.0621$) based on the data in Table 8.3

	Confidence limits		
Interval	**Lower**	**Upper**	**Width**
Wald	-0.118	-0.006	0.113
Wald with continuity correction	-0.119	-0.006	0.113
Wald with Agresti-Min adjustment	-0.118	-0.005	0.114
Wald with Bonett-Price adjustment	-0.120	-0.003	0.116
Newcombe square-and-add	-0.119	-0.005	0.114
Tango asymptotic score	-0.124	-0.005	0.119
Sidik exact unconditional	-0.126	-0.005	0.121
Sidik exact unconditional*	-0.124	-0.005	0.118

*Calculated with Berger and Boos procedure ($\gamma = 0.0001$)

for the difference between probabilities is

$$\mathrm{CP}(\pi_{1+}, \pi_{+1}, \theta, N, \alpha) = \sum_{x_{11}=0}^{N} \sum_{x_{12}=0}^{N-x_{11}} \sum_{x_{21}=0}^{N-x_{11}-x_{12}} I(L \leq \Delta \leq U) \cdot f(\mathbf{x} \,|\, \pi_{1+}, \pi_{+1}, \theta, N), \quad (8.18)$$

where $I()$ is the indicator function, $L = L(\mathbf{x}, \alpha)$ and $U = U(\mathbf{x}, \alpha)$ are the lower and upper $100(1-\alpha)\%$ confidence limits of an interval for the table $\mathbf{x} = \{x_{11}, x_{12}, x_{21}, N - x_{11} - x_{12} - x_{21}\}$, and $f()$ is the multinomial probability distribution (Equation 8.1). The exact expected interval width is defined as

$$\mathrm{Width}(\pi_{1+}, \pi_{+1}, \theta, N, \alpha) = \sum_{x_{11}=0}^{N} \sum_{x_{12}=0}^{N-x_{11}} \sum_{x_{21}=0}^{N-x_{11}-x_{12}} (U - L) \cdot f(\mathbf{x} \,|\, \pi_{1+}, \pi_{+1}, \theta, N).$$

Location is measured by the MNCP/NCP index. The total non-coverage probability (NCP) is computed as $1 - \mathrm{CP}$, where CP is defined in Equation 8.18. The mesial non-coverage probability (MNCP) is defined as

$$\mathrm{MNCP}(\pi_{1+}, \pi_{+1}, \theta, N, \alpha) = \sum_{x_{11}=0}^{N} \sum_{x_{12}=0}^{N-x_{11}} \sum_{x_{21}=0}^{N-x_{11}-x_{12}} I(L > \Delta \geq 0 \text{ or } U < \Delta \leq 0) \cdot f(\mathbf{x} \,|\, \pi_{1+}, \pi_{+1}, \theta, N).$$

Evaluation of Coverage Probability

Figure 8.8 illustrates the coverage probability of the four Wald intervals. Here, $\alpha = 0.05$, such that 95% confidence intervals are calculated. We have a small sample size (25 pairs of observations) and the intervals perform quite differently. The standard Wald interval has unacceptable low coverage. An improvement is obtained with the Wald interval with continuity correction, although its coverage is still quite low. The Wald interval with Agresti-Min adjustment is a greater improvement with coverage probabilities mostly between 94% and 95%. The only interval with coverage above 95% in Figure 8.8 is the Wald interval with Bonett-Price adjustment. It is conservative in almost all situations and performs much like an exact interval, although it cannot guarantee coverage at least to the nominal level. As with other conservative intervals, it may produce too wide intervals.

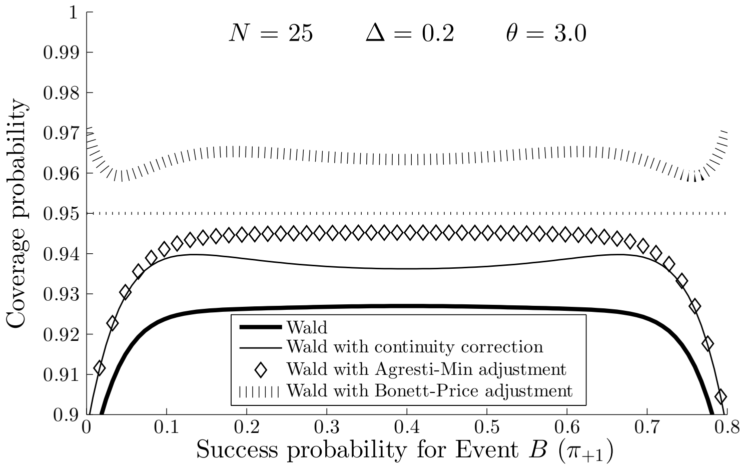

FIGURE 8.8
Coverage probabilities of four Wald intervals for the difference between probabilities

Figure 8.9 shows the same four Wald intervals for a sample size of $N = 100$ matched pairs. All intervals now have coverage closer to the nominal level compared with Figure 8.8, although the low coverage of the standard Wald interval may still cause concern. Note the excellent performance of the Wald interval with Bonett-Price adjustment: it has coverage slightly above the nominal level for all values of π_{+1}.

An example of the coverage properties of the Newcombe square-and-add, Tango asymptotic score, and Sidik exact unconditional intervals is shown in Figure 8.10. We include two versions of the Sidik exact unconditional interval: one with ($\gamma = 0.0001$) and one without ($\gamma = 0$) the Berger and Boos procedure. As noted in Section 8.5.7, when we evaluated the McNemar tests, we rarely see large effects of the Berger and Boos procedure on actual significance levels (tests) or coverage probabilities (confidence intervals). The paired 2×2 table seems to be an exception: the McNemar exact unconditional test with

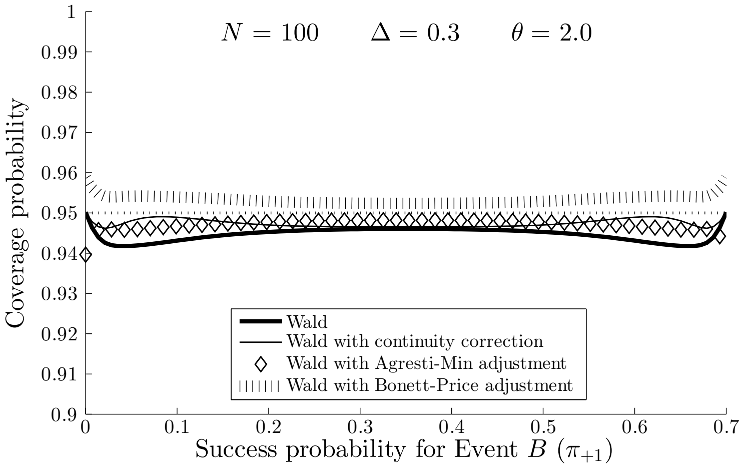

FIGURE 8.9
Coverage probabilities of four Wald intervals for the difference between probabilities

$\gamma = 0.0001$ has actual significance levels closer to the nominal level than the test with $\gamma = 0$, and the Sidik exact unconditional interval with $\gamma = 0.0001$ has coverage probabilities closer to the nominal level than the interval with $\gamma = 0$. For the exact unconditional test, this benefit is confined to $N < 25$, whereas for the exact unconditional interval, the benefit persists for many other combinations of N-, Δ-, and θ-values. The two other intervals in Figure 8.10 have coverage probabilities closer to the nominal level than the exact unconditional intervals. The Tango asymptotic score interval is particularly good in this example with only minor deviations from the nominal 95% coverage for all values of π_{+1}.

Unfortunately, the excellent performance of the Tango asymptotic score interval in Figure 8.10 does not continue for all choices of parameter values. Figure 8.11 shows that the coverage probability of the Tango asymptotic score interval can be quite low, even with as much as 40 matched pairs. In this example, the Newcombe square-and-add interval has the best coverage properties, although the Wald interval with Bonett-Price adjustment and the Sidik exact unconditional interval also perform quite well.

Evaluation of Width

Figure 8.12 shows an example of the expected widths of six confidence intervals for the difference between probabilities. The Wald interval and the Wald interval with continuity correction are not included because of their poor coverage properties. The situation in Figure 8.12 is representative for most other

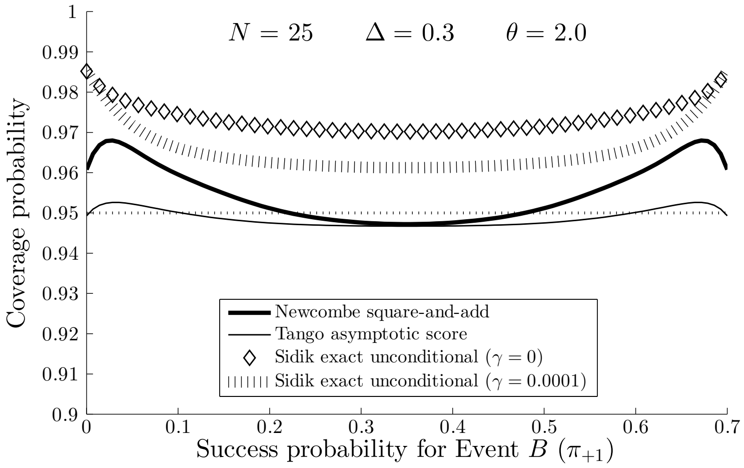

FIGURE 8.10
Coverage probabilities of four confidence intervals for the difference between probabilities

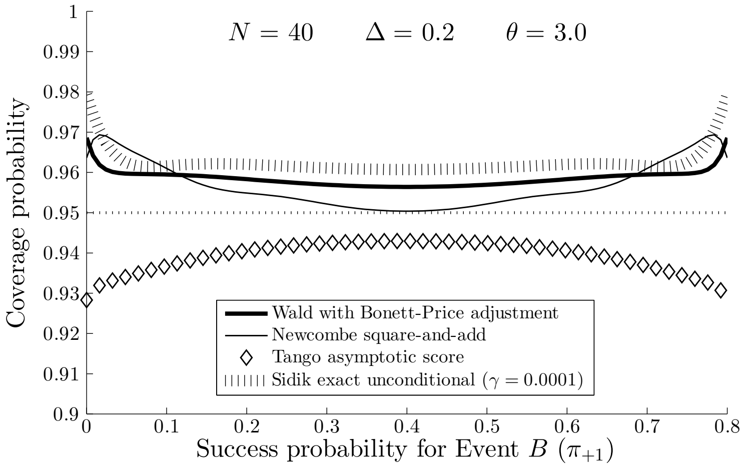

FIGURE 8.11
Coverage probabilities of four confidence intervals for the difference between probabilities

choices of parameters: the Wald interval with Agresti-Min adjustment, the Newcombe square-and-add interval, and the Tango asymptotic score interval are the shortest intervals followed by the Wald interval with Bonett-Price adjustment. The exact unconditional intervals are wider than the other intervals, and the interval with Berger and Boos procedure ($\gamma = 0.0001$) is slightly more narrow than the interval without Berger and Boos procedure ($\gamma = 0$). It may seem from Figure 8.12 that the differences in interval widths are considerable; however, the range of the y-axis (the width) is limited to 0.1, which may trick the eye and exaggerate the differences. Tables 8.10 and 8.11 show two examples where the practical differences in interval widths are mostly small.

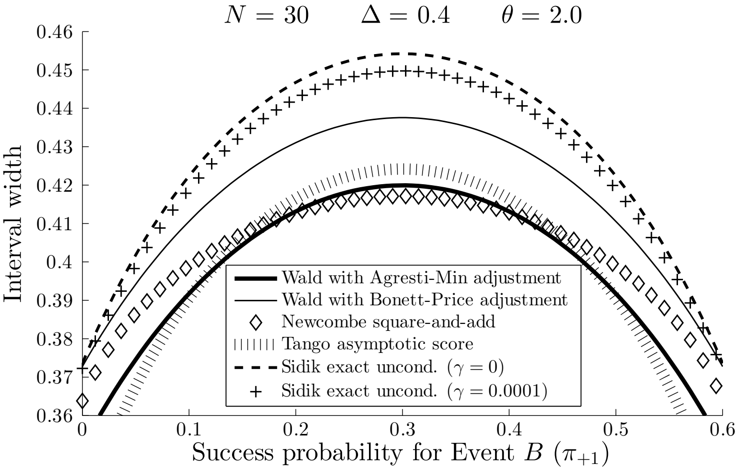

FIGURE 8.12
Expected width of six confidence intervals for the difference between probabilities

Evaluation of Location

Figure 8.13 shows a typical example of the location index MNCP/NCP for four of the confidence intervals for the difference between probabilities. The location of the Sidik exact unconditional interval (with and without Berger and Boos procedure) is usually in the satisfactory range ($0.4 \leq \text{MNCP/NCP} \leq 0.6$), as in Figure 8.13, although it can be slightly mesially located for other parameter values. The Wald interval with Bonett-Price adjustment and the Newcombe square-and-add interval are either slightly too mesially located (Figure 8.13) or with location just inside the satisfactory range. Four intervals are not shown: the Wald interval, the Wald interval with continuity correction, and the Tango asymptotic score interval have mostly satisfactory location, whereas the Wald interval with Agresti-Min adjustment has location similar to the Wald interval with Bonett-Price adjustment. Neither of the eight intervals is too distally located.

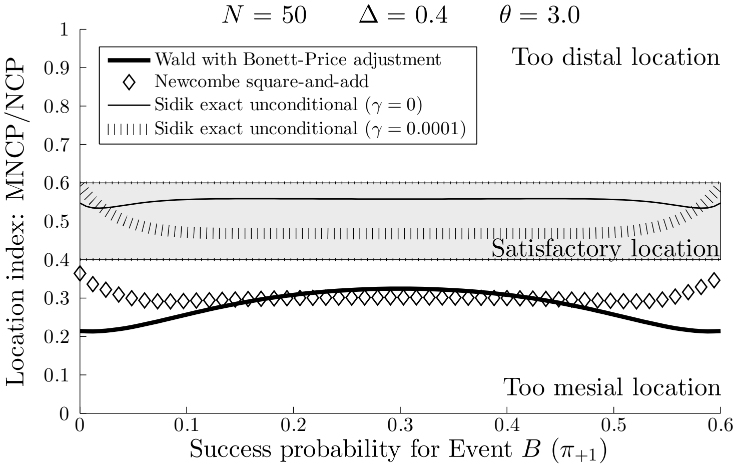

FIGURE 8.13
Location, as measured by the MNCP/NCP index, of four confidence intervals for the difference between probabilities

8.7 Confidence Intervals for the Number Needed to Treat

8.7.1 Introduction and Estimation

The number needed to treat was introduced in Section 4.6 for the unpaired 2×2 table. We can also calculate a number needed to treat for the paired 2×2 table, and the underlying concepts and ideas are the same. We therefore refer the reader to Section 4.6 for a general description and background material for the number needed to treat, including a brief discussion of the practical utility of the effect measure, and references to opposing views on whether and how the number needed to treat should be used.

As with the number needed to treat for unpaired data in Section 4.6.2, we estimate the number needed to treat for paired data as the reciprocal of the difference between probabilities (Walter, 2001):

$$\mathrm{NNT} = \frac{1}{\hat{\pi}_{1+} - \hat{\pi}_{+1}}.$$

With this notation, we assume that the "treatment" in the number needed to treat is associated with the binary event with success probability π_{+1} (Event B in Table 8.1), and that success indicates an unfavorable outcome, such as the presence of a certain disease. That is, when the difference between probabilities is positive, NNT is also positive, and Event B represents a beneficial event compared with Event A. If instead success represents a favorable outcome, a positive value of the difference between probabilities—and thereby a positive value of NNT—indicates a benefit for Event A compared with Event B. One may simply reverse the order of π_{1+} and π_{+1}, and define the difference between probabilities as $\pi_{+1} - \pi_{1+}$, to obtain the desired sign of NNT, if necessary.

As explained in Section 4.6.2, we may also—as suggested by Altman (1998)—denote positive values of NNT by NNTB: the number of patients needed to be treated for one additional patient to benefit; and negative values of NNT can be made positive and denoted by NNTH: the number of patients needed to be treated for one additional patient to be harmed. A proper interpretation of the number needed to treat thus is dependent on a careful definition of the items involved and their direction.

8.7.2 Confidence Intervals

A confidence interval for the number needed to treat is obtained by first calculating a confidence interval for the associated difference between probabilities. One of the methods in Section 8.6 should be used, and we denote the lower and upper confidence limits by L and U, respectively. If the confidence interval for the difference between probabilities does not include zero, the confidence interval for NNTB and NNTH can be obtained by taking the reciprocals of the absolute values of L and U and reversing their order:

$$1/|U| \;\text{ to }\; 1/|L|. \tag{8.19}$$

If, on the other hand, the interval (L, U) contains zero, the confidence interval for the number needed to treat should be denoted by (Altman, 1998):

$$\text{NNTH } 1/|L| \;\text{ to }\; \infty \;\text{ to }\; \text{NNTB } 1/U.$$

Figure 4.24 on page 133 illustrates the correspondence between the scales of the difference between probabilities and the number needed to treat, which may help deciphering the above expression.

8.7.3 Examples

Airway Hyper-Responsiveness Status before and after Stem Cell Transplantation (Table 8.2)

In this example, interest is on the probabilities of AHR before (π_{1+}) and after (π_{+1}) SCT. Here, the treatment is SCT, and the outcome denoted by "success" is an unfavorable event. A positive value of the difference between probabilities ($\Delta = \pi_{1+} - \pi_{+1}$) thereby indicates a beneficial effect of SCT, and vice versa for a negative value of Δ. With the data in Table 8.2, we get that

$$\hat{\Delta} = \frac{n_{12} - n_{21}}{N} = \frac{1 - 7}{21} = -0.286.$$

We estimate that SCT increases the probability of AHR by about 29 percentage points. Because AHR is a harmful event, we change the sign of $\hat{\Delta}$ and rephrase the number needed to treat in terms of the number needed to harm:

$$\text{NNTH} = \frac{1}{0.286} = 3.5.$$

We estimate that for every 3.5th patient treated with SCT, one additional patient will experience AHR.

To estimate a 95% confidence interval for the NNTH, we first calculate a 95% confidence interval for the corresponding Δ. This was done in Section 8.6.6 (see Table 8.10), where we observed quite similar results for the eight different interval methods. Here, we use the Wald interval with Bonett-Price adjustment, which is very easy to calculate and performs well in most situations. For the data in Table 8.2 (with Δ defined as $\pi_{+1} - \pi_{1+}$), the Wald interval with Bonett-Price adjustment is (0.013 to 0.508). Because this interval does not include zero, we use Equation 8.19 to find the corresponding 95% confidence interval for NNTH:

$$\left(\frac{1}{0.508} \text{ to } \frac{1}{0.013}\right) = (1.97 \text{ to } 76.9).$$

The frequency with which one additional patient will experience AHR may be as high as every 2nd patient or as low as every 77th patient treated with SCT.

Complete Response before and after Consolidation Therapy (Table 8.3)

When we defined the number needed to treat in Section 8.7.1, we assumed that “success” indicated an unfavorable outcome. Now, the outcome of interest is a beneficial one: complete response. To obtain a proper interpretation of the number needed to treat, we therefore reverse the sign of the estimate of the difference between probabilities. The sample proportions of patients with complete response before and after consolidation therapy are $\hat{\pi}_{1+} = 65/161 = 0.404$ and $\hat{\pi}_{+1} = 75/161 = 0.466$, respectively. The (reversed) estimate of the difference between probabilities then is $\hat{\Delta} = \hat{\pi}_{+1} - \hat{\pi}_{1+} = 0.0621$. The consolidation treatment seems to increase the probability of complete response, and we rephrase the number needed to treat as the number needed to benefit:

$$\text{NNTB} = \frac{1}{0.0621} = 16.1.$$

We estimate that for every 16 patients treated with consolidation therapy, one additional patient will have complete response.

A 95% Wald interval with Bonett-Price adjustment for Δ is (0.003 to 0.120), see Table 8.11. Because this interval does not contain zero, we use Equation 8.19 to find the corresponding 95% confidence interval for NNTB:

$$\left(\frac{1}{0.120} \text{ to } \frac{1}{0.003}\right) = (8.33 \text{ to } 333).$$

We state, with 95% confidence, that as few as 8.3 or as many as 333 patients need to be treated for one additional patient to benefit.

8.8 Confidence Intervals for the Ratio of Probabilities

8.8.1 Introduction and Estimation

In this section, we assume that the link function for the generalized linear models in Section 8.4 is the log link. The subject-specific model is given by

$$\log\big[\Pr(Y_t = 1\,|\,x_{kt})\big] = \alpha_k + \beta x_{kt},$$

for $t = 1, 2$ and $k = 1, 2, \ldots, N$. For the kth subject, $x_{k1} = 1$ for Event A and $x_{k2} = 0$ for Event B. We have that

$$\Pr(Y_t = 1\,|\,x_{kt}) = \exp(\alpha_k + \beta x_{kt})$$

and

$$\frac{\Pr(Y_1 = 1\,|\,x_{k1})}{\Pr(Y_2 = 1\,|\,x_{k2})} = \frac{\exp(\alpha_k + \beta \cdot 1)}{\exp(\alpha_k + \beta \cdot 0)} = \exp(\beta).$$

For each subject, the probability of Event A is $\exp(\beta)$ times the probability of Event B. By summation, we see that $\exp(\beta)$ is the ratio of the marginal probabilities. If we assume a marginal model, we drop the subject-specific subscript k from x and α in the preceding equations and obtain the same result. As for the difference between probabilities in Section 8.6.1, the marginal and the within-subject associations are the same.

We define the ratio of paired probabilities as the probability of success for Event A divided by the probability of success for Event B:

$$\phi = \frac{\pi_{1+}}{\pi_{+1}}.$$

The ratio of probabilities may be a more informative effect measure than the difference between probabilities in several situations, particularly when one or both probabilities are close to zero. We use the sample proportions to estimate ϕ:

$$\hat{\phi} = \frac{\hat{\pi}_{1+}}{\hat{\pi}_{+1}} = \frac{n_{1+}/N}{n_{+1}/N} = \frac{n_{11} + n_{12}}{n_{11} + n_{21}}.$$

Sections 8.8.2–8.8.5 present different confidence interval methods for ϕ. In Section 8.8.6, we apply the methods to the examples presented in Section 8.2. The methods are evaluated in Section 8.8.7, and Section 8.10 provides recommendations.

8.8.2 The Wald Interval

The Wald confidence interval for ϕ (Desu and Raghavarao, 2004, pp. 184–185) is obtained by exponentiating the endpoints of

$$\log \hat{\phi} \pm z_{\alpha/2}\sqrt{\frac{n_{12}+n_{21}}{n_{1+}\cdot n_{+1}}}. \tag{8.20}$$

When $n_{12} = n_{21} = 0$, the standard error estimate in (8.20) is zero, and the Wald interval produces the zero-width interval $(1, 1)$. If $n_{1+} = 0$, the estimate is $\hat{\phi} = 0$ and no upper limit is calculated. Similarly, if $n_{+1} = 0$, the estimate is infinite and no lower limit is calculated.

8.8.3 The Tang Asymptotic Score Interval

Under the constraint $\phi = \phi_0$, the score statistic for the ratio of paired binomial probabilities is (Tang et al., 2003, 2012)

$$T_{\text{score}}(\mathbf{n}\,|\,\phi_0) = \frac{n_{1+} - n_{+1}\phi_0}{\sqrt{N\bigl(1+\phi_0\bigr)\tilde{p}_{21} + \bigl(n_{11}+n_{12}+n_{21}\bigr)\bigl(\phi_0 - 1\bigr)}},$$

where

$$\tilde{p}_{21} = \frac{-B + \sqrt{B^2 - 4AC}}{2A},$$

and

$$\begin{aligned} A &= N(1+\phi_0), \\ B &= (n_{11}+n_{21})\phi_0^2 - (n_{11}+n_{12}+2n_{21}), \\ C &= n_{21}(1-\phi_0)(n_{11}+n_{12}+n_{21})/N. \end{aligned}$$

The Tang asymptotic score interval (L, U) for ϕ is obtained by solving the equations

$$T_{\text{score}}(\mathbf{n}\,|\,L) = z_{\alpha/2}$$

and

$$T_{\text{score}}(\mathbf{n}\,|\,U) = -z_{\alpha/2}.$$

An iterative algorithm is needed to solve the equations.

8.8.4 The Bonett-Price Hybrid Wilson Score Interval

Bonett and Price (2006) proposed a closed-form confidence interval for ϕ based on combining two Wilson score intervals (see Section 2.4.3) for the binomial parameters π_{1+} and π_{+1}. Let $n^* = n_{11} + n_{12} + n_{21}$, and define

$$A = \sqrt{\frac{n_{12}+n_{21}+2}{(n_{1+}+1)(n_{+1}+1)}}, \quad B = \sqrt{\frac{1-\frac{n_{1+}+1}{n^*+2}}{n_{1+}+1}}, \quad C = \sqrt{\frac{1-\frac{n_{+1}+1}{n^*+2}}{n_{+1}+1}},$$

and

$$z = \frac{A}{B + C}\, z_{\alpha/2}.$$

The Wilson score interval for π_{1+} is

$$(l_1, u_1) = \frac{2n_{1+} + z^2 \mp z\sqrt{z^2 + 4n_{1+}\Big(1 - \frac{n_{1+}}{n^*}\Big)}}{2(n^* + z^2)}, \tag{8.21}$$

and for π_{+1}, it is

$$(l_2, u_2) = \frac{2n_{+1} + z^2 \mp z\sqrt{z^2 + 4n_{+1}\Big(1 - \frac{n_{+1}}{n^*}\Big)}}{2(n^* + z^2)}. \tag{8.22}$$

The Bonett-Price hybrid Wilson score interval for ϕ is

$$\left(\frac{l_1}{u_2} \text{ to } \frac{u_1}{l_2}\right).$$

A continuity corrected version is obtained with the following adjustments to Equations 8.21 and 8.22:

$$\begin{aligned} l_1 &= \frac{2n_{1+} + z^2 - 1 - z\sqrt{z^2 - 2 - \frac{1}{n^*} + 4n_{1+}\Big(1 - \frac{n_{1+}+1}{n^*}\Big)}}{2(n^* + z^2)}, \\ u_1 &= \frac{2n_{1+} + z^2 + 1 + z\sqrt{z^2 + 2 - \frac{1}{n^*} + 4n_{1+}\Big(1 - \frac{n_{1+}-1}{n^*}\Big)}}{2(n^* + z^2)}, \end{aligned}$$

and

$$\begin{aligned} l_2 &= \frac{2n_{+1} + z^2 - 1 - z\sqrt{z^2 - 2 - \frac{1}{n^*} + 4n_{1+}\Big(1 - \frac{n_{1+}+1}{n^*}\Big)}}{2(n^* + z^2)}, \\ u_2 &= \frac{2n_{+1} + z^2 + 1 + z\sqrt{z^2 + 2 - \frac{1}{n^*} + 4n_{1+}\Big(1 - \frac{n_{1+}-1}{n^*}\Big)}}{2(n^* + z^2)}. \end{aligned}$$

If $n_{1+} = 0$ ($n_{+1} = 0$), set $l_1 = 0$ ($l_2 = 0$). If $n_{1+} = n^*$ ($n_{+1} = n^*$), set $u_1 = 1$ ($u_2 = 1$). The *Bonett-Price hybrid Wilson score interval with continuity correction* provides more conservative confidence intervals than the uncorrected interval.

8.8.5 The MOVER Wilson Score Interval

Section 8.6.3 introduced a MOVER confidence interval for the difference between paired binomial probabilities. That approach can also be used to construct confidence intervals for the ratio of paired probabilities. To find the lower confidence limit L for $\phi = \pi_{1+}/\pi_{+1}$, let $\theta_1 = \pi_{1+}$ and $\theta_2 = L\pi_{+1}$. As shown in Tang et al. (2012), we can use Equation 8.9 and the fact that

$$\Pr\bigl(\pi_{1+}/\pi_{+1} \leq L\bigr) = \Pr\bigl(\pi_{1+} - L\pi_{+1} \leq 0\bigr) = \alpha/2 \quad \Rightarrow \quad L^* = 0$$

to obtain

$$L = \frac{A - \hat{\pi}_{1+}\hat{\pi}_{+1} + \sqrt{\bigl(A - \hat{\pi}_{1+}\hat{\pi}_{+1}\bigr)^2 - l_1\bigl(2\hat{\pi}_{1+} - l_1\bigr)u_2\bigl(2\hat{\pi}_{+1} - u_2\bigr)}}{u_2\bigl(u_2 - 2\hat{\pi}_{+1}\bigr)}, \quad (8.23)$$

where $A = (\hat{\pi}_{1+} - l_1)(u_2 - \hat{\pi}_{+1})\widehat{\text{corr}}(\hat{\pi}_{1+}, \hat{\pi}_{+1})$. The upper confidence limit U for ϕ is found in a similar manner:

$$U = \frac{B - \hat{\pi}_{1+}\hat{\pi}_{+1} - \sqrt{\bigl(B - \hat{\pi}_{1+}\hat{\pi}_{+1}\bigr)^2 - u_1\bigl(2\hat{\pi}_{1+} - u_1\bigr)l_2\bigl(2\hat{\pi}_{+1} - l_2\bigr)}}{l_2\bigl(l_2 - 2\hat{\pi}_{+1}\bigr)}, \quad (8.24)$$

where $B = (u_1 - \hat{\pi}_{1+})(\hat{\pi}_{+1} - l_2)\widehat{\text{corr}}(\hat{\pi}_{1+}, \hat{\pi}_{+1})$. As in Newcombe (1998a) and Tang et al. (2012), we can use the phi coefficient, which in this case, also is the Pearson correlation coefficient, given by

$$\widehat{\text{corr}}(\hat{\pi}_{1+}, \hat{\pi}_{+1}) = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1+}n_{2+}n_{+1}n_{+2}}}.$$

If the denominator is 0, set $\widehat{\text{corr}}(\hat{\pi}_{1+}, \hat{\pi}_{+1}) = 0$.

The confidence limits L and U in Equations 8.23 and 8.24 depend on the particular confidence interval used to obtain (l_1, u_1) and (l_2, u_2). Tang et al. (2012) consider several different interval methods for the binomial parameter and their corresponding MOVER intervals, and recommend using the Wilson score interval (see Section 2.4.3). In that case, the appropriate expressions for (l_1, u_1) and (l_2, u_2) are given in Equations 8.11 and 8.12.

The MOVER Wilson score interval produces the zero-width interval $(1, 1)$ when $n_{11} = n_{22}$ and $n_{12} = n_{21} = 0$.

8.8.6 Examples

Airway Hyper-Responsiveness Status before and after Stem Cell Transplantation (Table 8.2)

The proportion of patients with AHR before SCT is $\hat{\pi}_{1+} = 2/21 = 0.096$, and the proportion of patients with AHR after SCT is $\hat{\pi}_{+1} = 8/21 = 0.38$. We estimate the ratio of probabilities as

$$\hat{\phi} = \frac{n_{11} + n_{12}}{n_{11} + n_{21}} = \frac{1 + 1}{1 + 7} = 0.25.$$

The probability of AHR after SCT is estimated to be four times the probability of AHR before SCT. Table 8.12 shows five different 95% confidence intervals for ϕ. The MOVER Wilson score interval is the shortest interval, followed by the Bonett-Price hybrid Wilson score and Tang asymptotic score intervals. Neither of these three intervals contains the null value ($\phi = 1.0$). The Wald interval is slightly wider and has 1.0 as the upper limit. The Bonett-Price hybrid Wilson score interval with continuity correction is considerably wider than the other intervals. Overall, there is less agreement between the intervals for the ratio of probabilities than was the case for the difference between probabilities (Table 8.10), for which none of the seven intervals contained the null value ($\Delta = 0$).

TABLE 8.12
95% confidence intervals for the ratio of probabilities ($\hat{\phi} = 0.25$) based on the data in Table 8.2

	Confidence limits		
Interval	**Lower**	**Upper**	**Log width**
Wald	0.063	1.000	2.77
Tang asymptotic score	0.065	0.907	2.63
Bonett-Price hybrid Wilson score	0.068	0.923	2.61
Bonett-Price hybrid Wilson score CC*	0.042	1.127	3.29
MOVER Wilson score	0.069	0.869	2.54

*CC = continuity correction

Complete Response before and after Consolidation Therapy (Table 8.3)

An estimate of the ratio of the probabilities for the data in Table 8.3 is

$$\hat{\phi} = \frac{n_{11} + n_{12}}{n_{11} + n_{21}} = \frac{59 + 6}{59 + 16} = 0.867.$$

We estimate the probability of complete response before consolidation therapy to be 13% smaller than the probability of complete response after consolidation therapy. Table 8.13 provides 95% confidence intervals for ϕ. All five intervals give similar confidence limits, although the Bonett-Price hybrid score interval with continuity correction is slightly wider than the other intervals. It is the only interval that includes the null value ($\phi = 1.0$); however, the other four intervals have upper limits that are marginally below the null value ($U \approx 0.99$ for all four intervals).

TABLE 8.13
95% confidence intervals for the ratio of probabilities ($\hat{\phi} = 0.867$) based on the data in Table 8.3

	Confidence limits		
Interval	**Lower**	**Upper**	**Log width**
Wald	0.760	0.989	0.263
Tang asymptotic score	0.748	0.988	0.278
Bonett-Price hybrid Wilson score	0.758	0.991	0.268
Bonett-Price hybrid Wilson score CC*	0.747	1.006	0.297
MOVER Wilson score	0.759	0.987	0.262

*CC = continuity correction

8.8.7 Evaluation of Intervals

Evaluation Criteria

As usual, we use three indices of performance to evaluate confidence intervals: coverage probability, width, and location (see Section 1.4 for general descriptions). In the following, we show how coverage, width, and location for the ratio of probabilities can be calculated exactly with complete enumeration. The succeeding expressions are simple modifications of the formulas in Section 8.6.7. The exact coverage probability for the ratio of probabilities is defined as

$$\mathrm{CP}(\pi_{1+},\pi_{+1},\theta,N,\alpha) = \sum_{x_{11}=0}^{N}\sum_{x_{12}=0}^{N-x_{11}}\sum_{x_{21}=0}^{N-x_{11}-x_{12}} I(L \leq \phi \leq U)\cdot f(\mathbf{x}\,|\,\pi_{1+},\pi_{+1},\theta,N), \quad (8.25)$$

where $\theta = \pi_{11}\pi_{22}/\pi_{12}\pi_{21}$, $I()$ is the indicator function, $L = L(\mathbf{x},\alpha)$ and $U = U(\mathbf{x},\alpha)$ are the lower and upper $100(1-\alpha)\%$ confidence limits of an interval for the table $\mathbf{x} = \{x_{11}, x_{12}, x_{21}, N - x_{11} - x_{12} - x_{21}\}$, and $f()$ is the multinomial probability distribution (Equation 8.1). The exact expected interval width (on the logarithmic scale) is defined as

$$\mathrm{Width}(\pi_{1+},\pi_{+1},\theta,N,\alpha) = \sum_{x_{11}=0}^{N}\sum_{x_{12}=0}^{N-x_{11}}\sum_{x_{21}=0}^{N-x_{11}-x_{12}} \big[\log(U)-\log(L)\big]\cdot f(\mathbf{x}\,|\,\pi_{1+},\pi_{+1},\theta,N).$$

To calculate the location index MNCP/NCP, we compute NCP = 1 − CP, where CP is defined in Equation 8.25 and

$$\mathrm{MNCP}(\pi_{1+},\pi_{+1},\theta,N,\alpha) = \sum_{x_{11}=0}^{N}\sum_{x_{12}=0}^{N-x_{11}}\sum_{x_{21}=0}^{N-x_{11}-x_{12}} I(L,U,\phi)\cdot f(\mathbf{x}\,|\,\pi_{1+},\pi_{+1},\theta,N),$$

where $I(L,U,\phi) = I\big[\log(L) > \log(\phi) \geq 0 \text{ or } \log(U) < \log(\phi) \leq 0\big]$.

Evaluation of Coverage Probability

We illustrate the coverage properties of the five confidence intervals for the ratio of probabilities by plotting the coverage probability against the probability of success for Event A (π_{1+}). That means that we hold N, ϕ (and thereby π_{+1}), and θ fixed. Two examples with small sample sizes are shown in Figures 8.14 and 8.15. These figures show that each of the intervals is associated with drawbacks; neither interval always performs well. The standard Wald interval often performs adequately, such as in Figure 8.14; however, it can have coverage probabilities considerably lower than the nominal level, usually when $\pi_{1+} > 0.7$ and the number of matched pairs is fairly low, say, $N \leq 40$. An example can be seen in Figure 8.15. The Tang asymptotic score and Bonett-Price hybrid Wilson score intervals often have similar coverage probabilities. The coverage probabilities of the Bonett-Price interval are often slightly closer to the nominal level than those of the asymptotic score interval. Both intervals may have coverage considerably lower than the nominal level for small values of π_{1+} and moderately large values of ϕ (Figure 8.14). For an interval with closed-form expression, the Bonett-Price interval performs excellently. The MOVER Wilson score interval, also a closed-form method, performs well; however—although not shown here—it has lower and more frequent dips in coverage below the nominal level than do the asymptotic score and Bonett-Price intervals. The Bonett-Price interval with continuity correction is very conservative: it has coverage above 98% for more than half the parameter space points in Figures 8.14 and 8.15.

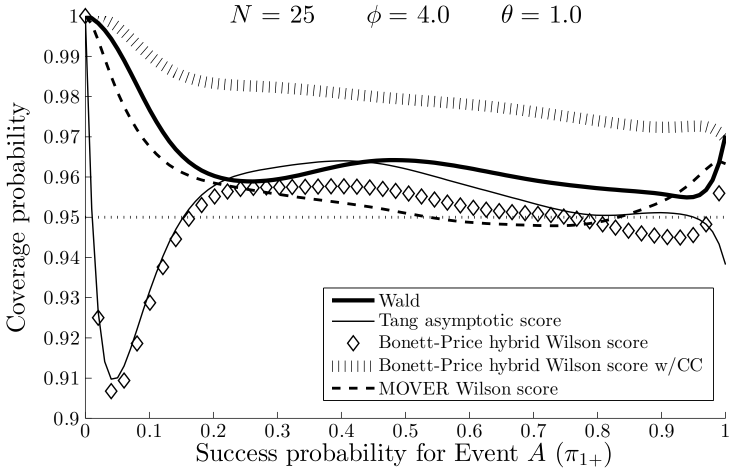

FIGURE 8.14
Coverage probabilities of five confidence intervals for the ratio of probabilities

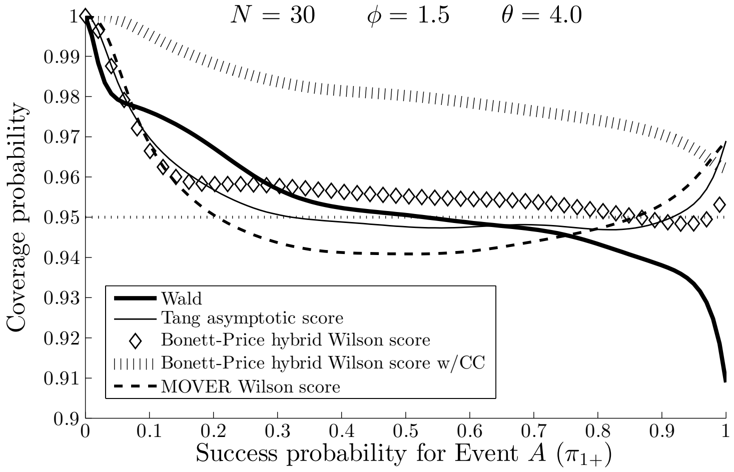

FIGURE 8.15
Coverage probabilities of five confidence intervals for the ratio of probabilities

Figure 8.16 illustrates how the intervals perform when we increase the sample size to 60 matched pairs. The Bonett-Price hybrid score interval with continuity correction is still very conservative; its minimum coverage probability is just below 97%. The Tang asymptotic score, Bonett-Price hybrid Wilson score, and MOVER Wilson score intervals all perform excellently, while the Wald interval is a bit too conservative.

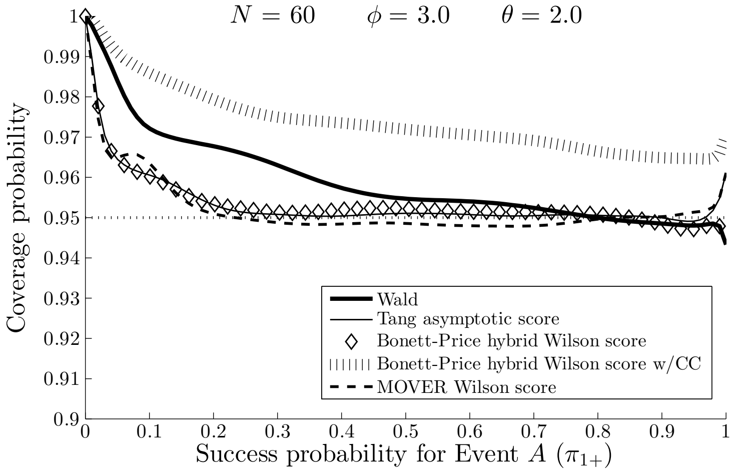

FIGURE 8.16
Coverage probabilities of five confidence intervals for the ratio of probabilities

Evaluation of Width

Figure 8.17 gives an example of the expected width of the intervals. The intervals can be ordered from the widest to the narrowest as follows: Bonett-Price hybrid score with continuity correction, Wald, Tang asymptotic score, Bonett-Price hybrid score, and MOVER Wilson score. In most cases, there is little to distinguish the widths of the latter three intervals.

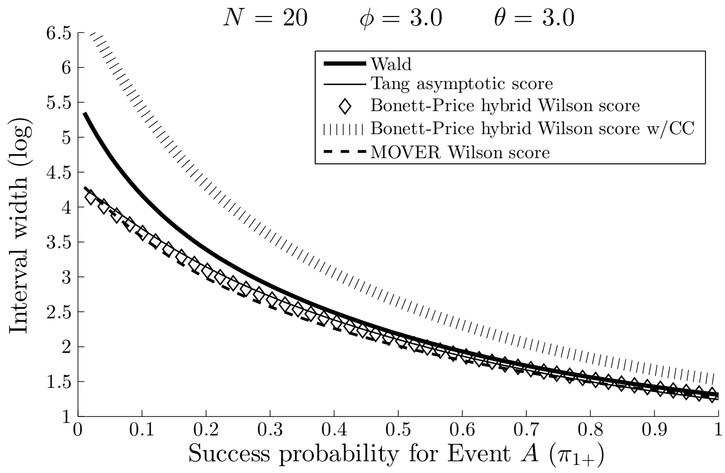

FIGURE 8.17
Expected width of six confidence intervals for the ratio of probabilities

Evaluation of Location

All intervals are too mesially located for most choices of parameter values (Figure 8.18). The MNCP/NCP values of the Tang asymptotic score and MOVER Wilson score intervals sometimes reach the satisfactory range (0.4, 0.6), but only for values of ϕ not too far from 1.0. The Wald interval and the Bonett-Price hybrid score interval with continuity correction have the worst location indices.

8.9 Confidence Intervals for the Odds Ratio

8.9.1 Introduction and Estimation

In matched cohort studies or clinical trials, we have exposure- (or treatment-) matching, in which exposed subjects are paired with unexposed subjects. In

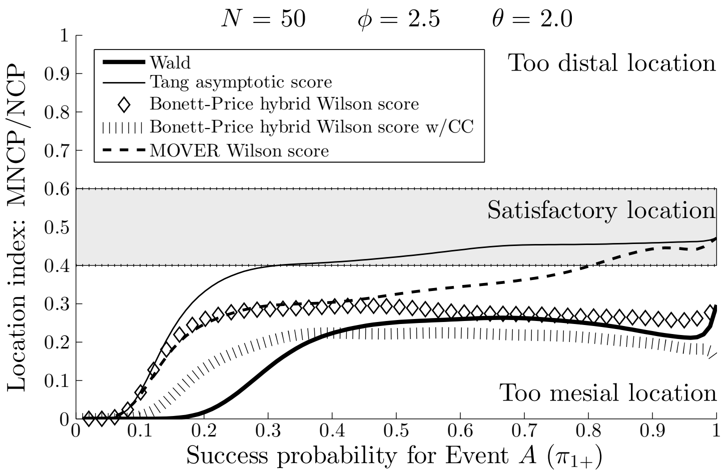

FIGURE 8.18
Location, as measured by the MNCP/NCP index, of four confidence intervals for the ratio of probabilities

matched case-control studies, on the other hand, matching is of diseased to non-diseased.

When the link function for the generalized linear models in Section 8.4 is the logit link, β is a log odds ratio. The subject-specific model is

$$\operatorname{logit}\bigl[\Pr(Y_t = 1 \,|\, x_{kt})\bigr] = \alpha_k + \beta x_{kt},$$

for $t = 1, 2$ and $k = 1, 2, \ldots, N$. For the kth subject, $x_{k1} = 1$ for Event A and $x_{k2} = 0$ for Event B. The odds for $Y_1 = 1$ is $\exp(\alpha_k + \beta)$ and the odds for $Y_2 = 1$ is $\exp(\alpha_k)$. Hence, for each subject, the odds of success for Event A is $\exp(\beta)$ times the odds for Event B. Averaging over the subjects will not give us the same interpretation of β as in the marginal model. For the marginal model, β equals the log odds ratio of the marginal probabilities in Table 8.1:

$$\beta_{\text{marginal}} = \log\left[\frac{\pi_{1+}/(1-\pi_{1+})}{\pi_{+1}/(1-\pi_{+1})}\right],$$

with maximum likelihood estimate $\hat{\beta}_{\text{marginal}} = \log[(n_{1+}/n_{2+})/(n_{+1}/n_{+2})]$.

For the subject-specific model, ordinary maximum likelihood estimation of β does not work because the number of α_k is proportional to N (Andersen, 1970; Agresti and Min, 2004). As shown in Andersen (1970), the unconditional maximum likelihood estimate of β converges to 2β. The solution is to use the conditional maximum likelihood estimate, obtained by conditioning on the number of discordant pairs ($n_{\text{d}} = n_{12} + n_{21}$), which is a sufficient statistic for

α_k. The conditional distribution is given in Equation 8.2, where

$$\mu = \frac{\pi_{12}}{\pi_{12} + \pi_{21}} = \frac{\theta_{\text{cond}}}{1 + \theta_{\text{cond}}}.$$

The conditional maximum likelihood estimate of θ_{cond} is

$$\hat{\beta}_{\text{cond}} = \log\left(\frac{n_{12}}{n_{21}}\right).$$

Note that $\hat{\beta}_{\text{cond}}$ equals the Mantel-Haenszel estimate (Breslow and Day, 1980, p. 165) of the common log odds ratio across N strata of matched case-control pairs (as in Table 8.6). We estimate the *conditional odds ratio* by

$$\hat{\theta}_{\text{cond}} = \frac{n_{12}}{n_{21}}.$$

We use the subscript "cond" to separate the paired-data conditional odds ratio ($\hat{\theta}_{\text{cond}}$) from the ordinary unconditional odds ratio ($\hat{\theta}$) used in several places throughout the book. The conditional odds ratio is the within pairs association, which generally is of more interest than the marginal association.

In a case-control study, the estimated within pairs association is the number of pairs with exposed cases and unexposed controls, divided by the number of pairs with unexposed cases and exposed controls.

Sections 8.9.2–8.9.4 present different confidence interval methods for θ_{cond}. In Section 8.9.5, we apply the methods to the examples presented in Section 8.2. The methods are evaluated in Section 8.9.6, and Section 8.10 provides recommendations.

8.9.2 The Wald Interval

An estimate of the asymptotic variance of $\hat{\beta}_{\text{cond}}$ is given by $1/n_{12}+1/n_{21}$. This is the standard Taylor series variance estimate of $\log(n_{12}/n_{21})$ and equals the Mantel-Haenszel variance estimate (Robins et al., 1986). To obtain the Wald interval for θ_{cond}, exponentiate the endpoints of

$$\log\big(\hat{\theta}_{\text{cond}}\big) \pm z_{\alpha/2}\sqrt{\frac{1}{n_{12}} + \frac{1}{n_{21}}}.$$

An equivalent expression is given by

$$\big(\hat{\theta}_{\text{cond}}/\text{EF to } \hat{\theta}_{\text{cond}} \cdot \text{EF}\big),$$

where EF is the error factor:

$$\text{EF} = \exp\left(z_{\alpha/2}\sqrt{\frac{1}{n_{12}} + \frac{1}{n_{21}}}\right).$$

The Wald interval is undefined if $n_{12} = 0$ or $n_{21} = 0$.

8.9.3 The Wald Interval with Laplace Adjustment

Greenland (2000) evaluated different bias-corrections for the odds ratio. We consider the simple *Laplace adjustment* obtained by adding 1 to each of n_{12} and n_{21} before calculating the Wald interval. The Wald interval with Laplace adjustment is given by exponentiating the endpoints of

$$\log\big(\tilde{\theta}_{\text{cond}}\big) \pm z_{\alpha/2}\sqrt{\frac{1}{\tilde{n}_{12}} + \frac{1}{\tilde{n}_{21}}},$$

where $\tilde{\theta}_{\text{cond}} = \tilde{n}_{12}/\tilde{n}_{21}$ and $\tilde{n}_{12} = n_{12} + 1$ and $\tilde{n}_{21} = n_{21} + 1$. The adjusted interval copes with $n_{12} = 0$ or $n_{21} = 0$ or both.

8.9.4 Intervals Obtained by Transforming Intervals for $\pi_{12}/(\pi_{12} + \pi_{21})$

In this section, we consider two asymptotic and two exact confidence intervals based on an approach described in Agresti and Min (2005b). Let (L_μ, U_μ) denote a confidence interval for the binomial parameter

$$\mu = \frac{\pi_{12}}{\pi_{12} + \pi_{21}}.$$

Because $\theta_{\text{cond}} = \mu/(1-\mu)$, a confidence interval for θ_{cond} is obtained as

$$(L \text{ to } U) = \left(\frac{L_\mu}{1 - L_\mu} \text{ to } \frac{U_\mu}{1 - U_\mu}\right). \tag{8.26}$$

In principle, any interval for μ can be used, and the confidence interval for θ_{cond} inherits the properties of the single binomial interval. In the following, let $n_{\text{d}} = n_{12} + n_{21}$.

Transforming the Wilson Score Interval

The Wilson (1927) score confidence interval for μ is given as

$$(L_\mu \text{ to } U_\mu) = \frac{2n_{12} + z_{\alpha/2}^2 \mp z_{\alpha/2}\sqrt{z_{\alpha/2}^2 + 4n_{12}\Big(1 - \frac{n_{12}}{n_{\text{d}}}\Big)}}{2\Big(n_{\text{d}} + z_{\alpha/2}^2\Big)}.$$

(See also Section 2.4.3). The transformation in Equation 8.26 gives the corresponding confidence interval for θ_{cond}.

The transformed Wilson score interval is equal to the approximate interval in Breslow and Day (1980, p. 166) without continuity correction. The interval with continuity correction is overly conservative (Agresti and Min, 2005b).

Transforming the Clopper-Pearson Exact Interval

Section 2.4.7 introduced the Clopper-Pearson exact interval for the binomial parameter and showed that the interval could be expressed with a beta distribution. Here, we repeat the expressions in terms of L_μ and U_μ, the lower and upper confidence limits for μ:

$$L_\mu = B(\alpha/2;\ n_{12},\ n_{21}+1)$$

and

$$U_\mu = B(1-\alpha/2;\ n_{12}+1,\ n_{21}).$$

$B(z; a, b)$ is the lower z-quantile of the beta distribution with parameters a and b. The transformation in Equation 8.26 yields an exact confidence interval for θ_{cond}.

Transforming the Clopper-Pearson Mid-P Interval

The Clopper-Pearson mid-P interval was introduced in Section 2.4.8. A mid-P interval (L_μ to U_μ) for μ can be obtained by iteratively solving

$$\sum_{i=n_{12}}^{n_{\mathrm{d}}} \binom{n_{\mathrm{d}}}{i} {L_\mu}^i (1-L_\mu)^{n_{\mathrm{d}}-i} - \frac{1}{2}\binom{n_{\mathrm{d}}}{n_{12}} {L_\mu}^{n_{12}} (1-L_\mu)^{n_{\mathrm{d}}-n_{12}} = \alpha/2$$

and

$$\sum_{i=0}^{n_{12}} \binom{n_{\mathrm{d}}}{i} {U_\mu}^i (1-U_\mu)^{n_{\mathrm{d}}-i} - \frac{1}{2}\binom{n_{\mathrm{d}}}{n_{12}} {U_\mu}^{n_{12}} (1-U_\mu)^{n_{\mathrm{d}}-n_{12}} = \alpha/2.$$

No simplification using the beta distribution is available for the Clopper-Pearson mid-P interval. An interval for θ_{cond} is obtained with the transformation in Equation 8.26.

Transforming the Blaker Exact Interval

Section 2.4.7 also included a description of the Blaker exact interval, for which the evaluations in Section 2.4.10 revealed some beneficial properties as compared with the Clopper-Pearson exact interval. For convenience, we repeat the expressions for the Blaker exact interval here, with notation appropriate for the problem of computing a confidence interval for θ_{cond}.

For $k = 0, 1, \ldots, n_{\mathrm{d}}$, define the function

$$\gamma(k, \pi_*) = \min\left[\sum_{i=k}^{n_{\mathrm{d}}} \binom{n_{\mathrm{d}}}{i} \pi_*^i (1-\pi_*)^{n_{\mathrm{d}}-i}, \sum_{i=0}^{k} \binom{n_{\mathrm{d}}}{i} \pi_*^i (1-\pi_*)^{n_{\mathrm{d}}-i}\right],$$

where π_* denotes an arbitrary confidence limit for μ. Let $\gamma(n_{12}, \pi_*)$ denote the value of γ for the observed data. The confidence limits of the Blaker exact interval for μ are the two solutions of π_* that satisfy the equation

$$\sum_{k=0}^{n_\mathrm{d}} I\big[\gamma(k,\pi_*) \leq \gamma(n_{12},\pi_*)\big] \cdot \binom{n_\mathrm{d}}{k} \pi_*^k (1-\pi_*)^{n_\mathrm{d}-k} = \alpha,$$

where $I()$ is the indicator function. The transformation in Equation 8.26 gives the corresponding exact interval for θ_cond.

8.9.5 Examples

The Association between Floppy Eyelid Syndrome and Obstructive Sleep Apnea-Hypopnea Syndrome (Table 8.4)

Previous sections in this chapter have shown how to estimate the difference between probabilities (Section 8.6.6) and the ratio of probabilities (Section 8.8.6)—with confidence intervals—for the data in Tables 8.2 and 8.3. Here, we do not estimate the odds ratio for these examples but turn our attention to the matched case-control study of the association between floppy eyelid syndrome (FES) and obstructive sleep apnea-hypopnea syndrome (OSAHS), for which the observed data is shown in Table 8.4.

Because this is a case-control study, we are unable to use the difference between probabilities and the ratio of probabilities as effect measures. In Section 8.5.6, we calculated five tests for association for these data and observed a strong association between FES and OSAHS ($P < 0.00011$ for all tests). Now, we use the odds ratio to estimate the size of this association:

$$\hat{\theta}_\mathrm{cond} = \frac{n_{12}}{n_{21}} = \frac{25}{2} = 12.5.$$

The odds of OSAHS among the patients with FES is estimated to be 12.5 times the odds of OSAHS among the patients without FES. Alternatively—because of the interchangeable nature of the odds ratio—the odds of FES for patients with OSAHS is estimated to be 12.5 times the odds of FES for patients without OSAHS.

Table 8.14 shows 95% confidence intervals for θ_cond. All six intervals have lower limits well above the null value ($\theta_\mathrm{cond} = 1.0$). Still, there is considerable variation in the upper limits and the interval widths. Note the close agreement between the transformed Clopper-Pearson mid-P and the transformed Blaker exact intervals.

8.9.6 Evaluation of Intervals

Evaluation Criteria

Again, we use three indices of performance to evaluate confidence intervals: coverage probability, width, and location (see Section 1.4 for general descriptions). The calculations of coverage probability, width, and location for the

TABLE 8.14
95% confidence intervals for the odds ratio ($\hat{\theta}_{\text{cond}} = 12.5$) based on the data in Table 8.4

	Confidence limits		
Interval	**Lower**	**Upper**	**Log width**
Wald	2.96	52.8	2.88
Wald with Laplace adjustment	2.62	28.6	2.39
Transformed Wilson score	3.28	47.7	2.68
Transformed Clopper-Pearson exact	3.12	109	3.55
Transformed Clopper-Pearson mid-*P*	3.47	78.3	3.12
Transformed Blaker exact	3.30	74.1	3.11

odds ratio differ from those for the difference between probabilities and the ratio of probabilities. As shown in Section 8.9.1, the odds ratio is defined conditional on the discordant pairs. Under this condition, the sample space is one-dimensional: any one possible table is completely characterized by the count of one cell (x_{12}). Because of the conditional nature of the odds ratio, the coverage probability, width, and location are also defined conditional on the discordant pairs. The exact coverage probability for the odds ratio is defined as

$$\text{CP}(\pi_{12}, n_{\text{d}}, \alpha) = \sum_{x_{12}=0}^{n_{\text{d}}} I(L \leq \theta_{\text{cond}} \leq U) \cdot f(x_{12} \,|\, n_{\text{d}}, \pi_{12}), \tag{8.27}$$

where $n_{\text{d}} = n_{12} + n_{21}$, $I()$ is the indicator function, $L = L(x_{12}, \alpha)$ and $U = U(x_{12}, \alpha)$ are the lower and upper $100(1-\alpha)\%$ confidence limits of an interval for any table with x_{12} and $x_{21} = n_{\text{d}} - x_{12}$ discordant pairs, and $f()$ is the binomial probability distribution with parameters n_{d} and π_{12} evaluated at x_{12}:

$$f(x_{12} \,|\, n_{\text{d}}, \pi_{12}) = \binom{n_{\text{d}}}{x_{12}} \pi_{12}^{x_{12}} (1 - \pi_{12})^{n_{\text{d}} - x_{12}}.$$

The exact expected interval width (on the logarithmic scale) is defined as

$$\text{Width}(\pi_{12}, n_{\text{d}}, \alpha) = \sum_{x_{12}=0}^{n_{\text{d}}} \big[\log(U) - \log(L)\big] \cdot f(x_{12} \,|\, n_{\text{d}}, \pi_{12}).$$

To calculate the location index MNCP/NCP, we compute NCP = 1 − CP, where CP is defined in Equation 8.27 and

$$\text{MNCP}(\pi_{12}, n_{\text{d}}, \alpha) = \sum_{x_{12}=0}^{n_{\text{d}}} I(L, U, \theta_{\text{cond}}) \cdot f(x_{12} \,|\, n_{\text{d}}, \pi_{12}).$$

where

$$I(L, U, \theta_{\text{cond}}) = I\big[\log(L) > \log(\theta_{\text{cond}}) \geq 0 \text{ or } \log(U) < \log(\theta_{\text{cond}}) \leq 0\big].$$

Evaluation of Coverage Probability

We fix the number of discordant pairs and plot the coverage probability as a function of π_{12}, the probability of success for Event A and failure for Event B. The coverage probability is a highly discontinuous function of π_{12}, as was the case for the confidence intervals for the binomial parameter in Chapter 2. Newcombe and Nurminen (2011) argue that in these cases, it is more informative to consider the moving average of the coverage probabilities, because this smoothed curve provides a realistic assessment of the coverage achieved in practice. An example with 30 discordant pairs is shown in Figure 8.19, where the moving average curves of the two Wald intervals are superimposed on their coverage probabilities. The Wald interval tends to be conservative, although it has coverage probabilities quite close to the nominal level for values of π_{12} close to 0.5. The Wald interval with Laplace adjustment has good average coverage for parts of the parameter space; however, coverage can be very low for small and large values of π_{12}.

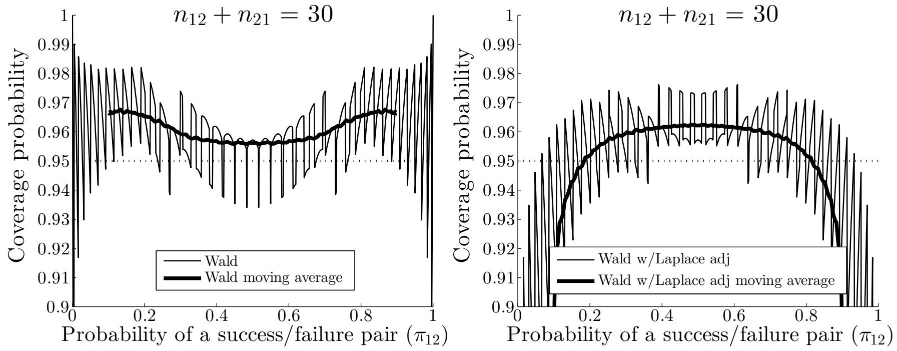

FIGURE 8.19
Coverage probabilities (with moving averages over the range $[\pi_{12} - 0.1, \pi_{12} + 0.1]$) of two Wald intervals for the odds ratio

Figure 8.20 shows an example of the coverage probabilities of the transformed Wilson score and transformed Clopper-Pearson mid-P intervals. Both intervals have excellent average coverage for most values of π_{12}. The Wilson score interval tends to fluctuate slightly more and dip slightly lower below the nominal level than do the mid-P interval. These performance traits persist for larger values of $n_{12} + n_{21}$, at least up to 100.

The coverage probabilities of the two exact intervals, the transformed Clopper-Pearson and Blaker intervals, are illustrated in Figure 8.21. Because these are exact intervals, their coverage probabilities are bounded below by the nominal level. The Blaker interval is considerably less conservative than the Clopper-Pearson interval. This difference is still clearly visible when the number of discordant pairs is increased to 100.

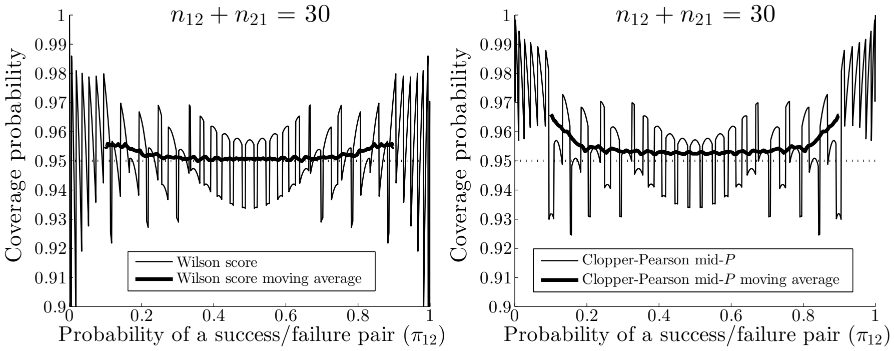

FIGURE 8.20
Coverage probabilities (with moving averages over the range $[\pi_{12} - 0.1, \pi_{12} + 0.1]$) of the transformed Wilson score and transformed Clopper-Pearson mid-P intervals for the odds ratio

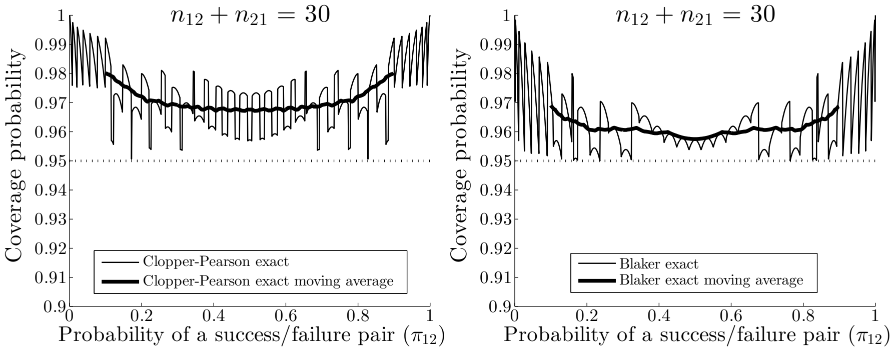

FIGURE 8.21
Coverage probabilities (with moving averages over the range $[\pi_{12} - 0.1, \pi_{12} + 0.1]$) of the transformed Clopper-Pearson exact and transformed Blaker exact intervals for the odds ratio

Evaluation of Width

The widths of the intervals can be ordered from the widest to the narrowest as follows: the transformed Clopper-Pearson exact interval, the transformed Blaker exact and the transformed Clopper-Pearson mid-P intervals (these two intervals have almost identical widths), the Wald interval, the transformed Wilson score interval, and the Wald interval with Laplace adjustment. Figure 8.22 gives an example for a small sample size ($n_{12} + n_{21} = 15$), where the between-interval differences are clearly seen. When the number of discordant pairs is greater than 50, the widths of all six intervals are similar.

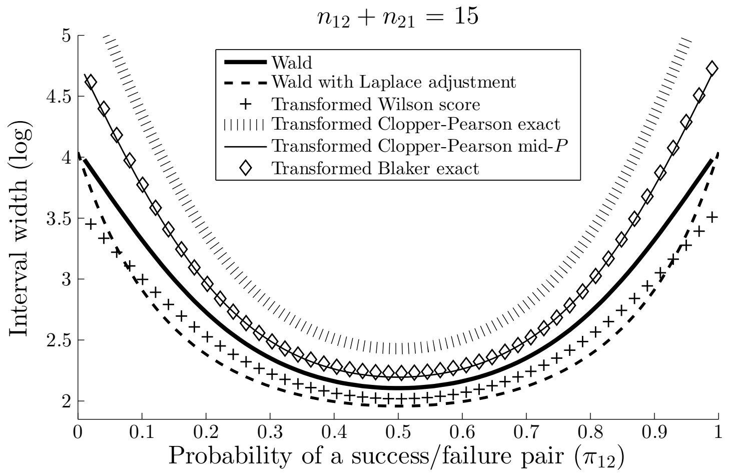

FIGURE 8.22
Expected width of six confidence intervals for the odds ratio

Evaluation of Location

The locations of the transformed Clopper-Pearson exact, transformed Clopper-Pearson mid-P, and transformed Blaker exact intervals, as measured by the MNCP/NCP index, are satisfactory for most combinations of parameters. An example is given in the right panel of Figure 8.23, which shows the location of the mid-P interval as a function of the probability of a success/failure pair (π_{12}) for a fixed total of 40 discordant pairs. In the left panel of Figure 8.23, the location of the transformed Wilson score interval is plotted. This interval has location mostly in the satisfactory range, except for small and large values of π_{12}, for which it is too mesially located. The Wald interval is slightly more mesially located than the transformed Wilson score interval, whereas the Wald interval with Laplace adjustment is too mesially located for all parameters, except when π_{12} is between 0.4 and 0.6.

8.10 Recommendations

8.10.1 Summary

Section 4.9 (recommendations for the unpaired 2×2 table) described the properties of the ideal method and made several general observations about how the different types of methods perform. Most of these observations also apply for the paired 2×2 table; however, we make two adaptations:

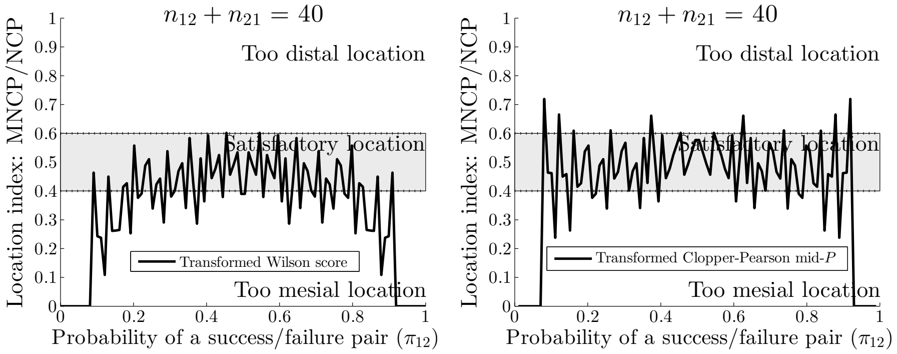

FIGURE 8.23
Location, as measured by the MNCP/NCP index, of the transformed Wilson score and transformed Clopper-Pearson mid-P intervals for the odds ratio

- The McNemar asymptotic test perform well even for small sample sizes
- Asymptotic score intervals do not perform as well for the paired 2 × 2 table as for the unpaired 2 × 2 table

Table 8.15 provides a summary of the recommended tests and confidence intervals, and gives the sample sizes for which the recommended methods are appropriate. The labels small, medium, and large cannot be given precise definitions, they will vary from one analysis to the other, and some subjectivity needs to be applied. As a rule of thumb, small may be taken as less than 50 number of pairs, medium as between 50 and 200 number of pairs, and large as more than 200 number of pairs. Sections 8.10.2–8.10.6 discuss the recommendations in more detail and summarize the merits of the different methods.

8.10.2 Tests for Association

Contrary to expectations, the simple McNemar asymptotic test performs well for all sample sizes. Its actual significance levels are close to the nominal level for almost any situation, except when the total number of matched pairs is very low ($N < 15$), in which case it is still better than the other tests in Section 8.5. The power of the McNemar asymptotic test is equal to or greater than that of the other tests for all situations. It frequently violates the nominal significance level, but not by much. The maximum actual significance level of the McNemar asymptotic test we have observed is 5.37% for a 5% nominal level. If that amount of infringement on the nominal level is acceptable—and we are of the opinion that it is—the asymptotic McNemar test can be considered the best test for association for the paired 2 × 2 table.

TABLE 8.15
Recommended tests and confidence intervals (CIs) for paired 2 × 2 tables

Analysis	Recommended methods	Sample sizes
Tests for association	McNemar asymptotic*	all
	McNemar mid-P*	all
	McNemar exact unconditional†	small/medium
CIs for difference between probabilities	Wald with Bonett-Price adjust.*	all
	Newcombe square-and-add*	small/medium
	Sidik exact unconditional†	small/medium
CIs for number needed to treat	The reciprocals of the limits of the recommended intervals for the difference between probabilities	
CIs for ratio of probabilities	Bonett-Price hybrid Wilson score*	all
	Tang asymptotic score	all
	MOVER Wilson score*	medium/large
	Wald*	large
CIs for odds ratio	Transformed Wilson score*	all
	Trans. Clopper-Pearson mid-P	all
	Transformed Blaker exact	small/medium
	Wald*	large

*These methods have closed-form expression
†Preferably with the Berger and Boos procedure ($\gamma = 0.0001$)

If an exact test is required, we recommend the McNemar exact unconditional test, preferable with the Berger and Boos procedure ($\gamma = 0.0001$), which is particularly beneficial in small sample sizes ($N < 25$). The commonly used McNemar exact conditional test is very conservative, and we do not recommend its use. Nor yet do we recommend use of the McNemar asymptotic test with continuity correction, which is at least as conservative as the exact conditional test. An easy-to-calculate alternative to the exact unconditional test is the McNemar mid-P test. Although the mid-P test cannot guarantee that the actual significance level does not exceed the nominal level, Fagerland et al. (2013) did not observe any violations of the nominal level in almost 10 000 considered scenarios. The supplementary materials to Fagerland et al. (2013) show how to calculate the mid-P test in eight commonly used software packages.

8.10.3 Confidence Intervals for the Difference between Probabilities

The Wald interval with Bonett-Price adjustment is a very good all-round interval, particularly considering how easy it is to calculate, and we recommend it for all sample sizes. Usually, we would recommend a standard (unadjusted)

Wald interval for large sample sizes; however, in this case, the Wald interval with Bonett-Price adjustment is so easy to calculate that there is no reason to resort to the standard Wald interval, which can have coverage slightly below the nominal level even for quite large sample sizes. The Wald interval with Bonett-Price adjustment can be a little bit conservative for small sample sizes, for which the Newcombe square-and-add interval often provides more narrow intervals. The Newcombe interval requires slightly more elaborate calculations than the Wald intervals; however, it has a closed-form expression and does not require dedicated software resources. The Tango asymptotic score interval usually performs quite well; however, it sometimes has coverage too far below the nominal level to recommend it for general use.

The Sidik exact unconditional interval also performs well. It is based on inverting two one-sided exact tests (the tail method) and sometimes provide overly conservative inference. One remedy, which seems to be far more effective for the paired 2 × 2 table than the unpaired 2 × 2 table, is to use the Berger and Boos procedure with $\gamma = 0.0001$. Another alternative is to use an exact unconditional interval that inverts one two-sided exact test (Tang et al., 2005). These intervals usually provide less conservative inference than the exact interval considered in Section 8.6.5; however, they have not yet found broad usage and are not available in standard software packages. The Sidik exact unconditional interval based on the tail method is available in the software StatXact (Cytel Inc., Cambridge, MA).

8.10.4 Confidence Intervals for the Number Needed to Treat

The calculation of a confidence interval for the number needed to treat is based on the confidence limits for the associated difference between probabilities. Hence, the recommended intervals for the difference between probabilities (Table 8.15) apply for the number needed to treat as well. The Wald interval with Bonett-Price adjustment deserves particular attention: it is very easy to calculate and performs well for most sample sizes and parameter values.

8.10.5 Confidence Intervals for the Ratio of Probabilities

The Bonett-Price hybrid Wilson score and the Tang asymptotic score intervals usually perform well, although low coverage can occur for small values of π_{1+} combined with moderately large values of ϕ. The Bonett-Price interval is particularly useful because it has a closed-form expression and thereby can be calculated without dedicated software resources. The Tang asymptotic score interval, on the other hand, requires iterative calculations. One advantage of the Tang interval is that it belongs to the family of score intervals, a well-known and general approach for constructing tests and confidence intervals for categorical data. According to our evaluations, the Bonett-Price hybrid Wilson score interval has coverage probabilities slightly closer to the nominal level than do the Tang asymptotic score interval. Both intervals can be recommended for general use. We also recommend the MOVER Wilson score interval, which, like the Bonett-Price hybrid Wilson score interval, can be calculated with simple arithmetics. Because of a high probability of low coverage, we do not recommend that the MOVER Wilson score interval is used for very small sample sizes, say, when the total number of matched pairs is less than 25.

No exact interval is available for the ratio of probabilities; however, the Bonett-Price hybrid Wilson score interval with continuity correction is conservative to the extent that it has a very low probability of coverage below the nominal level.

The standard Wald interval needs a large sample size to perform as well as the Bonett-Price hybrid Wilson score interval. When $N = 200$, the two intervals have similar coverage probabilities for most parameter values; however, the Bonett-Price interval has coverage probabilities slightly closer to the nominal level than the Wald interval for $\pi_{1+} < 0.3$. We believe the simplicity of the Wald interval makes up for this small difference in coverage probabilities and recommend the Wald interval when $N \geq 200$.

8.10.6 Confidence Intervals for the Odds Ratio

The transformed Wilson score and transformed Clopper-Pearson mid-P intervals have excellent average coverage probabilities for small as well as large sample sizes. The coverage probabilities of both intervals, however, fluctuate above and below the nominal level, and the smaller the sample size, the greater the fluctuations. Thus, for small sample sizes, the minimum coverage can be low. Nevertheless, we agree with Newcombe and Nurminen (2011), who argue for aligning the mean coverage—and not the minimum coverage—with the nominal $1 - \alpha$, and we recommend both intervals for all sample sizes. One advantage with the transformed Wilson score interval is that it has a closed-form expression, whereas the transformed Clopper-Pearson mid-P interval requires iterative calculations.

If the coverage probability is required to be at least to the nominal level, the transformed Blaker exact interval is superior to the transformed Clopper-Pearson exact interval. The Blaker interval has coverage closer to the nominal level, and it is shorter, than the Clopper-Pearson interval; however, it is also more complex to calculate and not well supported in software packages. The Clopper-Pearson exact interval (for the binomial parameter), on the other hand, is widely available in standard software packages.

When the sample size is large, say, with 100 or more discordant pairs, the Wald interval (without adjustment) performs about as well as the transformed Wilson score and transformed Clopper-Pearson mid-P intervals.