STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:729–740
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.1781

Simple improved confidence intervals for comparing matched proportions

Alan Agresti*,† and Yongyi Min

Department of Statistics, University of Florida, Gainesville, FL 32611-8545, U.S.A.

SUMMARY

For binary matched-pairs data, this article discusses interval estimation of the difference of probabilities and an odds ratio for comparing 'success' probabilities. We present simple improvements of the commonly used Wald confidence intervals for these parameters. The improvement of the interval for the difference of probabilities is to add two observations to each sample before applying it. The improvement for estimating an odds ratio transforms a confidence interval for a single proportion. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: binomial distribution; difference of proportions; logit model; odds ratio; score confidence interval; Wald confidence interval

1. INTRODUCTION

Matched-pairs data are common in biomedical studies, such as studies that focus on changes in subjects' responses over time, cross-over experiments comparing drugs, observations at pairs of body locations such as measurements relating to eyes or ears, and retrospective case-control studies. For binary responses, McNemar's test is the commonly applied significance test for comparing the two response distributions. The most common parameters are then the difference of proportions and an odds ratio. The difference of proportions is a natural parameter for randomized experiments and in longitudinal studies, whereas the odds ratio is natural for retrospective case-control studies. This article proposes simple ways of improving standard methods used to form confidence intervals for these parameters.

For interval estimation of the difference of proportions, textbooks present the Wald large-sample interval [1, 2]. This is the maximum likelihood (ML) estimate plus and minus a normal z-score times the estimated standard error (apart from possibly a continuity correction). Its coverage probabilities tend to be too low. Sections 2 and 3 show that a simple adjustment based on adding two observations to each sample provides substantial improvement, giving

*Correspondence to: Alan Agresti, Department of Statistics, University of Florida, Gainesville, FL 32611-8545, U.S.A.
†E-mail: aa@stat.ufl.edu

Received May 2003
Accepted November 2003

performance not much different from the interval based on inverting a score test. Section 4 shows that transforming the score confidence interval for a single proportion yields an improved interval for an odds ratio parameter.

2. IMPROVED CONFIDENCE INTERVAL FOR THE DIFFERENCE OF PROPORTIONS

For n matched pairs on a binary response, denote the probability of outcome i for the first observation and outcome j for the second observation by π_{ij}, where outcome 1 = 'success' and 2 = 'failure'. Denote the four corresponding sample proportions by $p_{11} = a/n$, $p_{12} = b/n$, $p_{21} = c/n$, and $p_{22} = d/n$. For instance, a is the number of pairs that are 'successes' for both observations. Table I summarizes the notation. The multinomial distribution is typically assumed for the cell counts. Let $\pi_1 = \pi_{11} + \pi_{12}$ and $\pi_2 = \pi_{11} + \pi_{21}$. The marginal totals are dependent binomials, with index n and parameters π_1 and π_2.

Denote the sample proportions of successes by p_1 and p_2. Their difference equals

$$p_2 - p_1 = p_{21} - p_{12} = (c - b)/n$$

With multinomial sampling, the sample estimate of $\mathrm{Var}(p_2 - p_1)$ is

$$\widehat{\mathrm{Var}}(p_2 - p_1) = \frac{[(p_{12} + p_{21}) - (p_{21} - p_{12})^2]}{n} = \frac{(b + c) - (c - b)^2/n}{n^2}$$

The Wald $100(1 - \alpha)$ per cent confidence interval for $(\pi_2 - \pi_1)$ is

$$(p_2 - p_1) \pm z_{\alpha/2}\sqrt{[(p_{12} + p_{21}) - (p_{21} - p_{12})^2]/n} \tag{1}$$

This uses the sample estimate of the exact standard error of $p_2 - p_1 = p_{21} - p_{12}$.

For a single proportion or a difference of independent proportions, the Wald method behaves poorly. Its true coverage probabilities are often well below the nominal level [3, 4]. In interval (1), three disadvantages analogous to ones that occur in those cases are readily apparent. First, the interval degenerates to (0, 0) when $p_{12} = p_{21} = 0$. Normally $\pi_{21} > 0$ and $\pi_{12} > 0$, so the true standard error is positive. Second, the interval degenerates to the Wald interval for π_2 if $p_1 = 0$ and to the Wald interval for $-\pi_1$ if $p_2 = 0$. Third, the interval centres

Table I. Notation for counts in 2×2 table for matched pairs, with sample proportions in parentheses.

	Column		
Row	Success	Failure	Total
Success	a (p_{11})	b (p_{12})	$a + b$ (p_1)
Failure	c (p_{21})	d (p_{22})	$c + d$
Total	$a + c$ (p_2)	$b + d$	n

at $(p_2 - p_1)$, not desirable when the distribution can be highly skewed. Not surprisingly, the Wald interval (1) for matched pairs also behaves poorly [5, 6], as does the corresponding non-null Wald test [7]. Its true coverage probabilities tend to fall below the nominal level [6], far below when both π_1 and π_2 are close to the parameter space boundary. With a continuity correction [2, p. 117] the performance improves, but having centre at $(p_2 - p_1)$ can still result in low coverage probabilities [6].

For a single proportion p with n observations, the Wald interval is $p \pm z_{\alpha/2}\sqrt{p(1-p)/n}$. The interval obtained by inverting the score test [8], which is the set of values π for which $|p - \pi|/\sqrt{\pi(1-\pi)/n} < z_{\alpha/2}$, performs much better [3]. A score test uses the standard error at parameter values that satisfy the null hypothesis rather than at the sample estimate. Adding $z^2_{\alpha/2}/2$ observations of each type before using the Wald formula yields an interval that contains the score interval, with the same midpoint, but with much better coverage probabilities than the ordinary Wald interval [9]. For 95 per cent confidence, $z_{0.025} \approx 2.0$ and this corresponds roughly to adding 2 outcomes of each type. This adjustment is now used in some introductory statistics texts [e.g. Reference [10]] in place of the Wald interval.

A similar simple adjustment improves the Wald interval for comparing two independent proportions [11]. With sample proportion p_i in sample i based on n_i observations, the Wald interval is

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

Adding 2 outcomes of each type, 1 of each type to each sample, improves the Wald interval substantially [11]. This adjusted interval is also now used in some introductory texts [e.g. Reference [10]].

An obvious question is whether a similar simple adjustment improves the Wald confidence interval (1) for comparing proportions with matched pairs. For instance, if we again add 1 outcome of each type to each sample, by adding $\frac{1}{2}$ to each of a, b, c, and d before constructing (1), does this improve the performance? Denote this interval by Wald+2, and more generally denote the Wald interval formed after adding $N/4$ to each cell of the table cross-classifying the two responses by Wald $+ N$. This interval is

$$(c^* - b^*)/n^* \pm z_{\alpha/2}\sqrt{(b^* + c^*) - [(c^* - b^*)^2/n^*]}/n^* \tag{2}$$

with $b^* = b + N/4$, $c^* = c + N/4$, $n^* = n + N$. When the interval overshoots the boundary at ± 1, it is truncated.

We studied the performance of the adjusted Wald $+ N$ interval for various N, focusing mainly on $N = 1, 2, 3, 4$. We express $\{\pi_{ij}\}$ in terms of the equivalent parameters $\{\pi_1, \pi_2, \theta\}$, where the odds ratio $\theta = (\pi_{11}\pi_{22}/\pi_{12}\pi_{21})$. We viewed the coverage probabilities as a function of π_2, for various combinations of π_1, θ, and the confidence coefficient. Results were qualitatively similar for all cases. The actual coverage probabilities for the Wald interval tend to fluctuate around a level somewhat below the nominal level, much below when $|\pi_2 - \pi_1|$ is very large. Of the Wald $+ N$ intervals, the one with $N = 2$ (i.e. adding 2 total to each sample, 0.5 to each cell) strikes a balance of tending to have coverage probabilities close to the nominal level while having a relatively small proportion of cases in which the coverage probability is well below the nominal level.

Table II. With $n=25$ and joint odds ratio $\theta=3$, table compares nominal 95 per cent confidence intervals for $\pi_2-\pi_1$ over $0<\pi_2<1$ for $\pi_1=0.5$ and 0.1, on the mean and minimum coverage probabilities, the proportion of coverage probabilities between 0.94 and 0.96 and between 0.93 and 0.97, the proportion below 0.93, the mean absolute distance between the actual coverage probability and 0.95, and the mean of the expected interval lengths.

π_1	Method	Mean	Min	(0.94, 0.96)	(0.93, 0.97)	<0.93	Distance	Length
0.5	Wald	0.933	0.919	0.091	0.671	0.329	0.017	0.436
	Wald + 1	0.942	0.921	0.669	0.975	0.025	0.008	0.436
	Wald + 2	0.947	0.927	0.919	0.992	0.008	0.005	0.434
	Wald + 3	0.950	0.917	0.864	0.975	0.025	0.006	0.432
	Wald + 4	0.950	0.910	0.594	0.924	0.076	0.010	0.430
	Score	0.952	0.938	0.891	0.991	0.000	0.006	0.448
0.1	Wald	0.923	0.800	0.096	0.466	0.534	0.027	0.354
	Wald + 1	0.948	0.891	0.579	0.864	0.057	0.011	0.366
	Wald + 2	0.954	0.909	0.673	0.849	0.028	0.010	0.375
	Wald + 3	0.951	0.912	0.563	0.805	0.060	0.012	0.381
	Wald + 4	0.942	0.846	0.226	0.466	0.345	0.024	0.385
	Score	0.962	0.940	0.370	0.809	0.000	0.013	0.389

Table II illustrates typical results. It summarizes some characteristics of the Wald and Wald + N intervals for the case $\theta=3$ with $\pi_1=0.1$ and 0.5 for $n=25$ and 95 per cent confidence, viewed over the space of possible π_2. Figure 1 shows actual coverage probabilities of the ordinary Wald interval (1) and the Wald + 2 interval. Probabilities are plotted as a function of π_2 for the cases summarized in Table II. The adjustment provides substantial improvement.

In practice, relatively small values of $\pi_2-\pi_1$ usually have particular relevance. Figure 2 compares coverage probabilities of the Wald and Wald + 2 intervals for the fixed differences 0.0 and 0.1, again when $n=25$ and $\theta=1$ and 3. Table III summarizes characteristics for these cases. When $\pi_2-\pi_1=0$ for π_1 and π_2 near 0 or near 1, the Wald + 2 interval can be quite conservative, more so as θ increases. This is not surprising for such a small n, as then a tends to be nearly equal to n when π_1 and π_2 are close to 1 and d tends to be nearly equal to n when π_1 and π_2 are close to 0.

Similar results occurred for other cases considered. From the form of the estimated variance in the Wald interval, the same result applies if the added constant is redistributed in any way between a and d. In fact, quite similar performance results if no adjustment is made to those cells. That is, the interval with $b^*=b+0.5$, $c^*=c+0.5$, $n^*=n+1$, which makes no adjustment to cells a and d and which we denote by Wald + 1bc, has similar performance as Wald + 2. The interval Wald + 2bc with $b^*=b+1$, $c^*=c+1$, $n^*=n+2$, also performs much better than the Wald interval but tends to be very conservative, especially when $|\pi_2-\pi_1|$ is small and $|\log\theta|$ is large.

We also considered 90 per cent and 99 per cent confidence intervals. Similar results occurred. In summary, adding 0.5 at least to b and c before calculating the popular Wald interval (1) is a simple way of providing great improvement. Interestingly, Haldane [12] noted that the 0.5 correction to each cell works well for reducing first-order bias in estimating the log odds ratio.

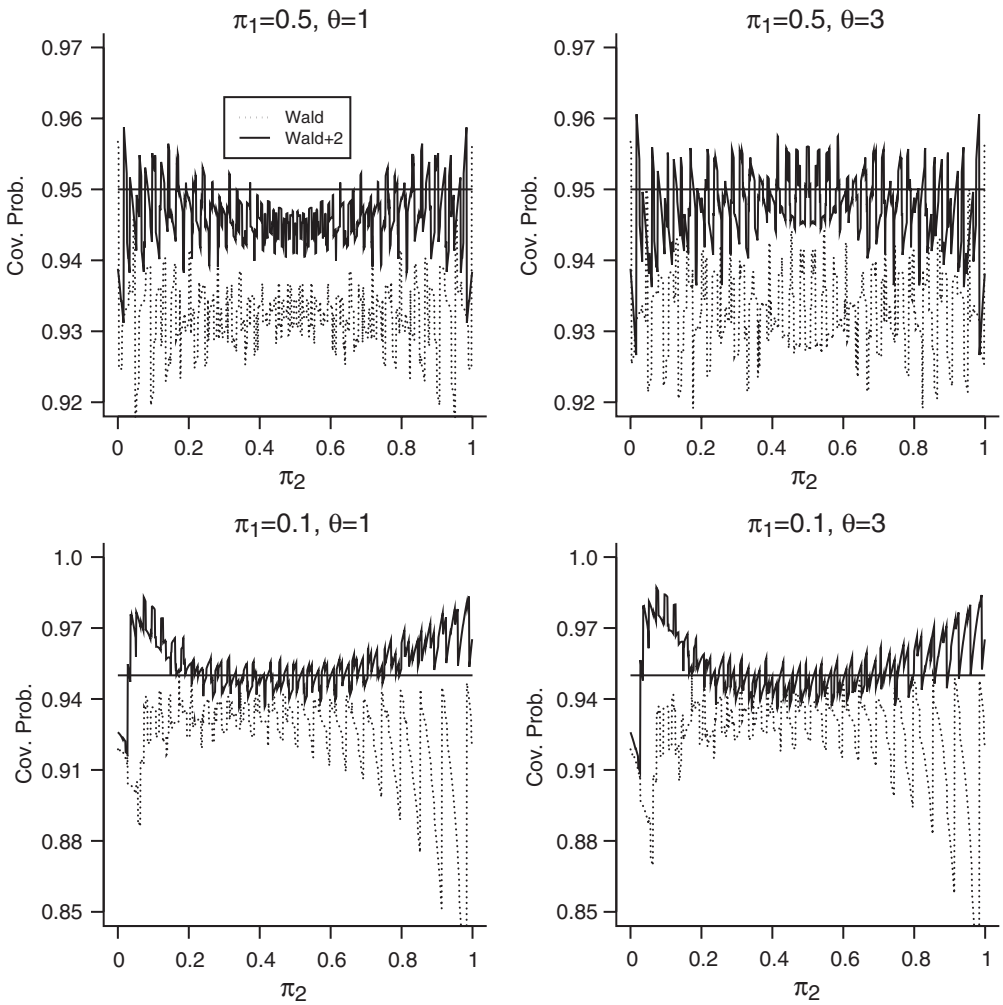

Figure 1. Coverage probabilities of 95 per cent confidence intervals for $\pi_2 - \pi_1$ based on dependent binomials with odds ratio θ and $n = 25$.

3. JUSTIFICATION FOR ADJUSTED CONFIDENCE INTERVAL?

Although the Wald + 2 interval provides improved coverage performance over the ordinary Wald interval, an obvious disadvantage is its ad hoc nature. This gives it inherent disadvantages, such as the possibility of overshooting the boundary. Is there any justification for such an interval other than that it seems to work well? In the single proportion case, the adjustment of adding 2 successes and 2 failures occurred as a simple approximation to the score interval, which was a method known to perform well in that case. For matched pairs, the score interval

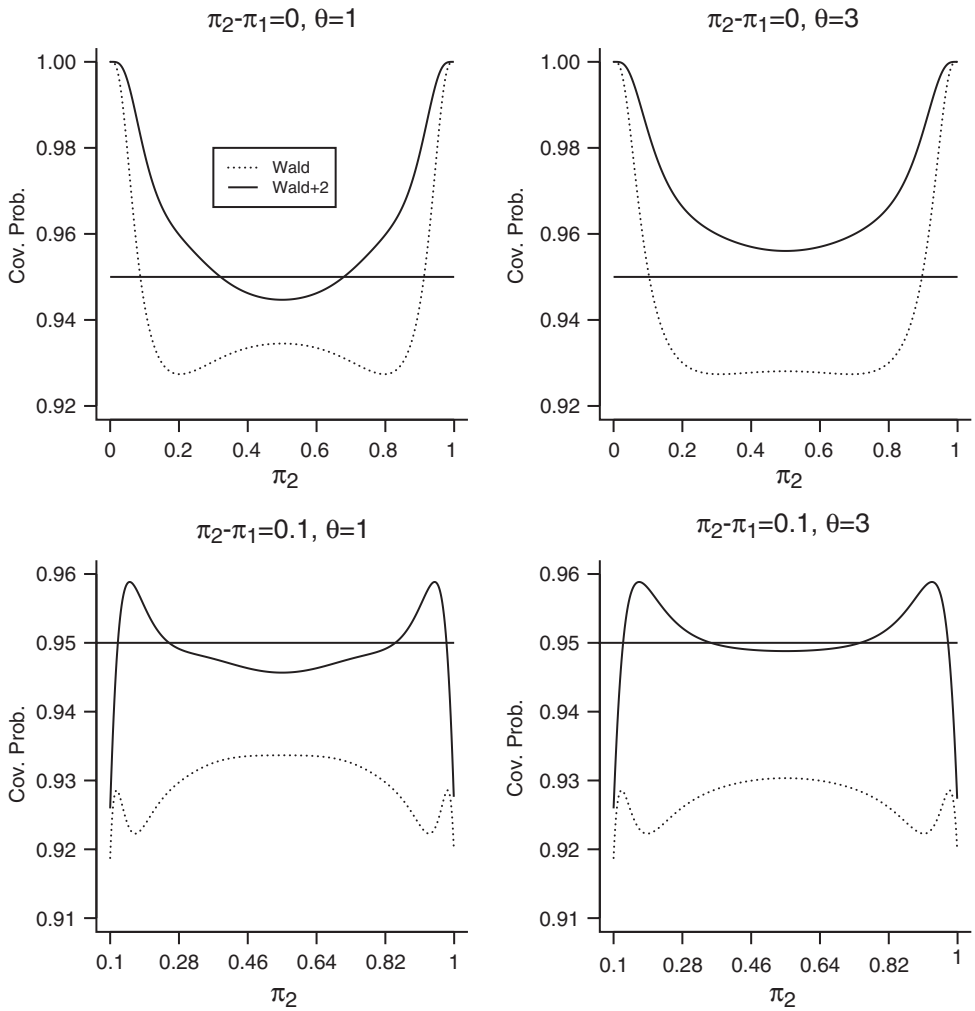

Figure 2. Coverage probabilities of 95 per cent confidence intervals for $\pi_2 - \pi_1$ based on dependent bionomials with odds ratio θ and $n = 25$.

consists of the set of $\Delta = \pi_2 - \pi_1$ values for which

$$\frac{|(p_2 - p_1) - \Delta|}{\sqrt{[(\hat{\pi}_{12}(\Delta) + \hat{\pi}_{21}(\Delta)) - \Delta^2]/n}} < z_{\alpha/2} \tag{3}$$

where $\hat{\pi}_{jk}(\Delta)$ denotes the ML estimate of π_{jk} under the constraint that $(\pi_2 - \pi_1) = \Delta$. There is not a closed-form expression for the resulting interval, but it can be obtained using iterative methods. Surprisingly, this conceptually simple method does not seem to have been proposed until Tango [13].

Table III. With $n = 25$ and joint odds ratio θ, table compares nominal 95 per cent confidence intervals for $\pi_2 - \pi_1$ over the parameter space for which $\pi_2 - \pi_1 = 0$ or 0.1.

θ	$\pi_2 - \pi_1$	Method	Mean	Min	(0.94, 0.96)	(0.93, 0.97)	< 0.93	Distance	Length
1	0	Wald	0.941	0.927	0.070	0.595	0.284	0.020	0.419
		Wald + 2	0.962	0.945	0.603	0.737	0.000	0.015	0.421
		Score	0.961	0.951	0.841	0.875	0.000	0.011	0.457
	0.1	Wald	0.930	0.919	0.000	0.586	0.414	0.020	0.448
		Wald + 2	0.949	0.926	0.979	0.994	0.006	0.004	0.442
		Score	0.954	0.952	0.908	0.966	0.000	0.004	0.470
3	0	Wald	0.940	0.927	0.098	0.264	0.599	0.022	0.368
		Wald + 2	0.969	0.956	0.403	0.663	0.000	0.019	0.377
		Score	0.961	0.951	0.817	0.859	0.000	0.011	0.421
	0.1	Wald	0.927	0.919	0.000	0.192	0.808	0.023	0.395
		Wald + 2	0.951	0.926	0.974	0.994	0.006	0.003	0.396
		Score	0.956	0.953	0.881	0.957	0.000	0.006	0.432

Studies by Tango [13, 14] and Newcombe [15] as well as our own evaluations showed that the score method (3) performs very well for the difference of proportions with matched pairs, even for fairly small sample sizes. Tables II and III also summarize results for it. Although the score method tends to be a bit conservative (especially when either parameter is near the boundary), it has the advantages that it cannot overshoot the $[-1, 1]$ bounds and its coverage probabilities rarely fall much below the nominal level. We did note, however, that this method tends to have coverage probabilities farther from the nominal level than either Wald + 2 or Wald + 1*bc* for 90 per cent intervals. Also, the Wald + 2 interval tends to be shorter than the score interval, especially when the success probabilities are similar (see Table III), although all the intervals tend to be very wide for the small n reported here.

Is the score interval related in any way to a simple adjustment of the Wald interval? Suppose we replace the nuisance parameter $(\pi_{12} + \pi_{21})$ in (3) by $(p_{12} + p_{21})$. Inverting the test then results in a closed-form interval,

$$(p_2 - p_1)(n/n^*) \pm z_{\alpha/2}\sqrt{[n^*(p_{12} + p_{21}) - n(p_{21} - p_{12})^2]}/n^* \tag{4}$$

where $n^* = n + z^2_{\alpha/2}$. This shrinks the midpoint $(p_2 - p_1)$ of the Wald interval but, like the Wald interval, is degenerate when $p_{12} = p_{21} = 0$. May and Johnson [5] proposed this interval. Lloyd [16] took the same approach but used a continuity correction.

Interval (4) has the same midpoint as the Wald + N interval with $N = z^2_{\alpha/2}$. That adjusted Wald interval corresponds to replacing the nuisance parameter $\pi = (\pi_{12} + \pi_{21})$ in the score approach by $(b + c + z^2_{\alpha/2}/2)/(n + z^2_{\alpha/2})$, which is the midpoint of the score confidence interval for a binomial parameter with a sample of size n and $b + c$ successes. For 95 per cent confidence, this interval is essentially Wald + 4. It performs somewhat better than (4) and much better than the ordinary Wald interval (1). However, we found that this interval tends to be too conservative when $|\pi_2 - \pi_1|$ is small, more so as $|\log\theta|$ increases, yet it tends to be too liberal when $|\pi_2 - \pi_1|$ is very large.

By contrast, the related interval Wald + 2 does not tend to overly shrink. Although the score-test does not provide as clear motivation as in the single-proportion case for adding a specific number of observations to the data before forming the Wald interval, it does motivate

Table IV. Example of binary matched pairs, from cross-over study reported by Jones and Kenward [17].

	High dose		
Low dose	Success	Failure	Total
Success	53	8	61
Failure	16	9	25
Total	69	17	86

Table V. Ninety five per cent confidence intervals for difference of proportions and for odds ratio for model (5) with Table IV.

Parameter	Confidence interval	
Difference of proportions	Wald	$(-0.017, 0.203)$
	Wald + 2	$(-0.019, 0.201)$
	Score	$(-0.020, 0.207)$
Odds ratio	Wald	$(0.86, 4.67)$
	Transformed score	$(0.88, 4.56)$
	Transformed score-cc	$(0.81, 5.09)$
	Transformed Blaker 'exact'	$(0.84, 4.91)$

the usefulness of shrinkage. In our evaluations, the coverage performance of the Wald + 2 interval was comparable to the score test-based interval (as Tables II and III suggest), but the score interval was usually somewhat wider and had fewer cases with low coverage probability.

Viewing the Wald and Wald + 2 intervals for a variety of data sets reveals that they are usually quite similar. To illustrate, consider Table IV, based on data from a cross-over study described by Jones and Kenward [17]. This table summarizes results of a comparison of low- and high-dose analgesics for relief of primary dysmenorrhea. Table V shows the Wald and Wald + 2 intervals for Table IV, for which the difference in the success proportions between the high- and low dose is 0.093. The Wald + 2 interval usually has similar length as the Wald interval, but the slight shrinkage of the midpoint toward 0 results in much improved coverage probabilities. For these data, the score interval is somewhat wider. Here, this is partly reflective of the conservatism that method tends to exhibit when the true difference is small, but this is typical of what we observed for many data sets. Although it tends to be wider, the score interval rarely has converage probability much below the nominal level.

Many statisticians will view the improved performance of Wald + 2 over the Wald interval as support for a Bayesian approach. One can regard the Wald + 2 interval as an ordinary Wald interval applied with Bayesian estimates. For the multinomial distribution, the sample proportions formed after adding 0.5 to each cell are the Bayes estimates of $\{\pi_{ij}\}$ when those parameters have a Dirichlet distribution with all its parameters equal to 0.5; in fact, that is the Jeffreys prior for the multinomial. See Reference [18] for an introductory text that takes the

approach of conducting approximate Bayesian inference by using standard frequentist formulas with Bayes estimates in place of the usual ML estimates.

Of course, other less ad hoc intervals may also perform well, such as ones based directly on the likelihood function. It is not the purpose of this paper to conduct a comparison of various methods. See Reference [6] for such a study. Our goal is merely to provide a simple improvement over the interval that is most commonly used in practice for this problem.

4. IMPROVED CONFIDENCE INTERVALS FOR THE ODDS RATIO WITH MATCHED PAIRS

Denote the two observations for pair i by (y_{i1}, y_{i2}), $i = 1, \ldots, n$, where $y_{it} = 1$ is a success and $y_{it} = 2$ is a failure, $t = 1, 2$. A common model for matched pairs is

$$\text{logit}[P(y_{i1} = 1)] = \alpha_i, \quad \text{logit}[P(y_{i2} = 1)] = \alpha_i + \beta \tag{5}$$

where β is a pair-specific log odds ratio. Ordinary ML estimation of β fails because the number of $\{\alpha_i\}$ parameters is proportional to n [19, pp. 244–245]. A popular alternative approach eliminates $\{\alpha_i\}$ by conditioning on their sufficient statistics [20]. The resulting conditional ML estimator equals $\hat{\beta} = \log(c/b)$. The estimated asymptotic variance of $\hat{\beta}$ is $(b^{-1} + c^{-1})$. This estimate is often used with retrospective case-control studies, for which one cannot estimate the difference of proportions.

The Wald interval for the log odds ratio applies the delta method to $\hat{\beta} = \log(c/b)$, yielding

$$\log(c/b) \pm z_{\alpha/2}\sqrt{b^{-1} + c^{-1}} \tag{6}$$

Wald intervals for log odds ratios with independent binomial samples tend to be conservative, sometimes overly so [21]. Our investigations showed that the same is true with samples generated from model (5). Also, a disadvantage of (6) is that it requires adjustment if b or $c = 0$. The interval remains conservative when applied after adding a constant to b and to c, although the interval always then exists and the estimated log odds ratio has finite bias. A correction of 1 to b and c works well for reducing bias in estimating the conditional odds ratio [22].

For model (5), $\exp(\beta) = \pi_{21}/\pi_{12}$. This suggests that one can obtain an ordinary binomial interval (L, U) for $\pi^*_{21} = \pi_{21}/(\pi_{21} + \pi_{12})$ and then use $(L/(1-L), U/(1-U))$ as the interval for the odds ratio [23]. In fact, this works well when one uses the score confidence interval for π^*_{21}, which is the set of π^*_{21} for which

$$\frac{|c/(b+c) - \pi^*_{21}|}{\sqrt{\pi^*_{21}(1-\pi^*_{21})/(b+c)}} < z_{\alpha/2}$$

Table VI illustrates results for model (5), when $\{\alpha_i\}$ are *iid* $N(\mu, \sigma^2)$, for the eight combinations of μ, σ, and β for which each equal 0 or 1, for 95 per cent and 99 per cent intervals. This is a simple way of improving performance over the Wald interval (6).

Table VI. For the conditional model (5), with $\alpha_i \sim N(\mu, \sigma^2)$, table compares coverage probabilities for confidence intervals for log odds ratio β, where score, score-cc (score with continuity correction) and Blaker refer to transforming binomial confidence intervals for $\pi_{21}/(\pi_{12}+\pi_{21})$.

			95 % confidence				99 % confidence			
β	σ	μ	Wald	score	score-cc	Blaker	Wald	score	score-cc	Blaker
0	0	0	0.971	0.957	0.976	0.957	1.000	0.993	0.996	0.993
		1	0.981	0.958	0.979	0.958	1.000	0.993	0.998	0.994
	1	0	0.979	0.957	0.978	0.958	1.000	0.993	0.997	0.993
		1	0.985	0.957	0.978	0.960	1.000	0.994	0.998	0.993
1	0	0	0.976	0.956	0.980	0.966	0.995	0.992	0.996	0.995
		1	0.981	0.966	0.986	0.971	0.998	0.989	0.996	0.996
	1	0	0.978	0.960	0.982	0.968	0.996	0.991	0.996	0.995
		1	0.981	0.966	0.986	0.972	0.998	0.989	0.996	0.996
Mean			0.979	0.960	0.981	0.964	0.998	0.992	0.997	0.994

Likewise, one could apply the $(L/(1-L), U/(1-U))$ transform to 'exact' confidence intervals for π_{21}^* that use the binomial instead of the normal distribution and guarantee that the coverage probability is at least the nominal level [23]. Best known is the Clopper–Pearson interval that inverts two separate one-sided binomial tests, but this tends to be overly conservative. Blaker [24] gave a less conservative interval for a binomial proportion. Table VI also shows the coverage probabilities with this approach.

Breslow and Day [23, p. 166] suggested estimating the odds ratio by transforming a binomial confidence interval. However, they suggested using the score confidence interval for π_{21}^* with a continuity correction. Results with this approach are overly conservative, about as much so as the Clopper–Pearson interval. Table VI also shows results for it. We do not recommend it, as it tends to be more conservative than a method (Blaker's) that guarantees achieving at least the nominal confidence level. Yet another approach would instead transform the Agresti and Coull [9] adjusted Wald interval for π_{21}^*. Letting $\delta = z_{\alpha/2}^2/2$, this interval for π_{21}^* is $p^* \pm z_{\alpha/2}\sqrt{p^*(1-p^*)/(b+c+2\delta)}$ with $p^* = (c+\delta)/(b+c+2\delta)$. Our evaluations showed that this is less conservative, but it does not tend to work as well as the score-test-based interval (results not shown here).

Table VI shows results for 95 per cent and 99 per cent confidence levels, but similar results occur for 90 per cent confidence. Over the eight settings listed, the mean coverage probability for transforming the (Wald interval, score interval, score interval with continuity correction, Blaker interval) is (0.934, 0.896, 0.956, 0.924). Table V shows the results of the 95 per cent confidence intervals for Table IV, for which the estimated odds ratio is $\frac{16}{8} = 2.0$.

In summary, we recommend transforming the score interval if the goal is simplicity and obtaining true coverage probability near the nominal level. To obtain an interval that guarantees achieving *at least* the nominal confidence level, we recommend transforming Blaker's interval. We also recommend transforming Blaker's interval instead of the score interval if one expects the odds ratio to be extremely large or extremely small, because the score interval for a binomial proportion can have poor coverage probability [9] when the proportion is very close to 0 or very close to 1.

5. CONCLUSION

For interval estimation of probabilities and of parameters comparing probabilities, inverting the score test provides a good, general-purpose method. The Wald confidence interval is computationally simple and commonly used, but it has inadequate coverage probabilities. For interval estimation of the difference of probabilities with matched-pairs data, a simple modification of the Wald interval based on adding observations to each sample provides great improvement. One can view the Wald+2 interval as a computationally simple alternative to the score interval. For interval estimation of the pair-specific odds ratio, a simple transformation of the score interval for a single proportion performs well. In future research it would be of interest to develop confidence intervals for related parameters that apply for data sets stratified by a covariate.

Unfortunately, score confidence intervals are currently unavailable in standard software packages, even for comparisons of proportions with independent samples. The website http://www.stat.ufl.edu/~aa/cda/software.html contains code for using the free software R to construct many confidence intervals, including the Tango score confidence interval and our adjusted Wald confidence interval for a difference of proportions and the transformed score confidence interval for the odds ratio.

ACKNOWLEDGEMENTS

This research was partially supported by grants from NIH and NSF. The authors thank Dr Sander Greenland for helpful comments about an earlier draft.

REFERENCES

1. Agresti A. *Categorical Data Analysis* (2nd edn). Wiley: New York, 2002.
2. Fleiss JL. *Statistical Methods for Rates and Proportions* (2nd edn). Wiley: New York, 1981.
3. Ghosh BK. A comparison of some approximate confidence intervals for the binomial parameter. *Journal of the American Statistical Association* 1979; **74**:894–899.
4. Brown LD, Cai TT, Das Gupta A. Interval estimation for a binomial proportion. *Statistical Science* 2001; **16**:101–117.
5. May WL, Johnson WD. Confidence intervals for differences in correlated binary proportions. *Statistics in Medicine* 1997; **16**:2127–2136.
6. Newcombe R. Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine* 1998; **17**:2635–2650.
7. Liu J, Hsueh H, Hsieh E, Chen JJ. Tests for equivalence or non-inferiority for paired binary data. *Statistics in Medicine* 2002; **21**:231–245.
8. Wilson EB. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 1927; **22**:209–212.
9. Agresti A, Coull BA. Approximate is better than 'exact' for interval estimation of binomial proportions. *American Statistician* 1998; **52**:119–126.
10. Moore D, McCabe G. *Introduction to the Practice of Statistics* (4th edn). Freeman: New York, 2003.
11. Agresti A, Caffo B. Simple and effective confidence intervals for proportions and difference of proportions result from adding two successes and two failures. *American Statistician* 2000; **54**:280–288.
12. Haldane JBS. The estimation and significance of the logarithm of a ratio of frequencies. *Annals of Human Genetics* 1956; **20**:309–311.
13. Tango T. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine* 1998; **17**:891–908.
14. Tango T. Letter to the editor. *Statistics in Medicine* 1999; **18**:3511–3513.
15. Newcombe RG. Interval estimation for the mean of a variable taking the values 0, 1 and 2. *Statistics in Medicine* 2003; **22**:2737–2750.

16. Lloyd CJ. Confidence intervals from the difference between two correlated proportions. *Journal of the American Statistical Association* 1990; **85**:1154–1158.
17. Jones B, Kenward MG. Modelling binary data from a three-period cross-over trial. *Statistics in Medicine* 1987; **6**:555–564.
18. Berry DA. *Statistics*: *A Bayesian Perspective*. Wadsworth: Belmont, CA, 1996.
19. Andersen EB. *Discrete Statistical Models with Social Science Applications*. North-Holland: Amsterdam, 1980.
20. Cox DR. Two further applications of a model for binary regression. *Biometrika* 1958; **45**:562–565.
21. Agresti A. On logit confidence intervals for the odds ratio with small samples. *Biometrics* 1999; **55**:597–602.
22. Greenland S. Small-sample bias and corrections for conditional maximum-likelihood odds-ratio estimators. *Biostatistics* 2000; **1**:113–122.
23. Breslow N, Day NE. *Statistical Methods in Cancer Research*: *The Analysis of Case-Control Studies*. IARC: Lyon, 1980; 165–166.
24. Blaker H. Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics* 2000; **28**:783–798.

