COMMUNICATIONS IN STATISTICS
Simulation and Computation®
Vol. 33, No. 4, pp. 1095–1113, 2004

Small Sample Confidence Intervals for the Odds Ratio

Raef Lawson*

School of Business, University at Albany, State University of New York, Albany, New York, USA

ABSTRACT

Numerous methods—based on exact and asymptotic distributions—can be used to obtain confidence intervals for the odds ratio in 2×2 tables. We examine ten methods for generating these intervals based on coverage probability, closeness of coverage probability to target, and length of confidence intervals. Based on these criteria, Cornfield's method, without the continuity correction, performed the best of the methods examined here. A drawback to use of this method is the significant possibility that the attained coverage probability will not meet the nominal confidence level. Use of a mid-*P* value greatly improves methods based on the "exact" distribution.

*Correspondence: Raef Lawson, School of Business, University at Albany, State University of New York, Albany, NY 12222, USA; Fax: (518) 442-3045; E-mail: lawson@albany.edu.

DOI: 10.1081/SAC-200040691 0361-0918 (Print); 1532-4141 (Online)

When combined with the Wilson rule for selection of a rejection set, the resulting method is a procedure that performed very well. Crow's method, with use of a mid-*P*, performed well, although it was only a slight improvement over the Wilson mid-*P* method. Its cumbersome calculations preclude its general acceptance. Woolf's (logit) method—with the Haldane–Anscombe correction— performed well, especially with regard to length of confidence intervals, and is recommended based on ease of computation.

Key Words: 2×2 tables; Confidence interval; Odds ratio; mid-*P*; Wilson; Woolf.

INTRODUCTION

Testing for the equality of two independent binomial proportions is an important statistical problem. Analysis of this problem is usually based on the odds ratio. In this paper we examine various methods for obtaining confidence intervals for the odds ratio using both approximate and exact (discrete) distributions. We look at the effect of using various methods for obtaining two sided confidence intervals, of the definition of *P*-value, and of the use of the continuity correction on those intervals, as measured using three criteria.

The usual assumption is that there are two independent binomial random variants X and Y from samples of size m and n with parameters of "success" of p_1 and p_2, respectively. The outcome of a trial can be presented in the following 2×2 table:

Group	Success	Failure	
I	X	$m - X$	m
II	Y	$n + Y$	n
Total	T	$m + n - T$	$m + n$

The likelihood of a given outcome (x, y) is

$$\binom{m}{x}\binom{n}{y} p_1^x q_1^{m-x} p_2^y q_2^{n-y},$$

which can be rewritten in terms of the odds ratio, $\psi = p_1q_2/p_2q_1$, where $q_i = 1 - p_i,\ i = 1, 2$:

$$\binom{m}{x}\binom{n}{y} q_1^m q_2^n \psi^x \left(\frac{p_2}{q_2}\right)^{x+y}.$$

Thus x and $x + y$ are the minimal sufficient statistics. Inference about ψ can be based on the conditional distribution of X given $X + Y = T$:

$$f(x|T;\psi) = \frac{\binom{m}{x}\binom{n}{T-x}\psi^x}{\sum_{j=0}^{T}\binom{m}{j}\binom{n}{T-j}\psi^j}$$

(Cornfield, 1956; Fisher, 1935; Lehmann, 1959) and then developed in terms of either an approximate or an exact (discrete) distribution. (The problem can also be approached based on the unconditional distribution. See Agresti, 1992; Agresti and Min, 2002; Barnard, 1945, 1947; Habaer, 1986; Rice, 1988.)

TWO-SIDED CONFIDENCE INTERVALS FOR DISCRETE DATA PROBLEMS

In general, two-sided confidence intervals can be obtaining in a variety of ways. Most commonly, they are constructed by inverting two separate one-sided tests, each with size at most $\alpha/2$. The interval consists of the set of θ_0 values for which the attained significance level, or P-value, exceeds $\alpha/2$ in conducting each test of the null hypothesis $\theta = \theta_0$. We will call this the Clopper–Pearson method (Clopper and Pearson, 1934).

An alternative to this method is the Wilson method (Wilson, 1942), described by Sterne (1954). This method defines the attained level of significance as $\hat{\alpha} = P\{x|f(x|\theta_0) \le f(x_0|\theta_0)\}$, which is the probability under H_0 of the set of values with probability as low, or lower, than that of the actual outcome x_0. The hypothesis is rejected whenever $\hat{\alpha} \le \alpha/2$. Use of this method for generating confidence intervals can result in narrower intervals. A disadvantage of this approach is that the separate limits that it generates have no direct meaning (Berry and Armitage, 1995).

The Crow method (Crow, 1956) for obtaining confidence intervals is a modification of the Wilson technique. For a discrete data problem, the Wilson technique gives an acceptance set $\{r, r+1, \ldots, s\}$ for each null hypothesis. Under the null hypothesis, this set of values will have a probability that is $\ge 1 - \alpha$. Crow considers replacing this set with either $\{r-1, r, \ldots, s-1\}$ or $\{r+1, \ldots, s, s+1\}$ provided that the replacement set has probability $\ge 1 - \alpha$ and *closer* to $1 - \alpha$. That is, Crow considers

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

moving the acceptance set leftward or rightward by one integer to bring the acceptance set closer to having a probability of $-\alpha$. This replacement is only occasionally possible, so the Crow method is only a slight improvement on the Wilson method.

The definition of *P*-value plays an important role in the construction of confidence intervals for discretely distributed random variates. *P*-values indicate how often sampling errors would lead to as great or greater discrepancy than that actually observed between a theoretical model and the actual observations. For a random variable X with a continuous cumulative distribution function $F(X|\theta_0)$, the determination of a *P*-value is clear. For discretely distributed random variates, various definitions of *P*-values have been proposed. These stem from the fact that a realization x of a discrete random variable X can be thought of as representing an underlying value in the range $(x-\frac{1}{2}, x+\frac{1}{2})$. The determination of a *P*-value depends on where in this range we view the observation as having come from. In a hypothesis testing situation, an upper-tail ordinary *P*-value views the observation as having come from the lower end of the possible range, i.e., it is equal to $\Pr\{X \geq x-\frac{1}{2}|\theta_0\}$.

An alternative definition of *P*-value, the mid-*P*, was introduced by Lancaster (1949, 1952). (See also the discussion by Barnard, 1989, 1990; Berry and Armitage, 1995; Hirji et al., 1991; Miettinen, 1974; Williams, 1988.) An upper-tail mid-*P* is given by

$$\Pr\{X > x|\theta_0\} + \frac{1}{2}\Pr\{X = x|\theta_0\}.$$

This definition of a *P*-value, based on the assumption that the observation came from the middle of the range of underlying values, can be viewed as making a *P*-value from a discrete distribution more comparable to one whose distribution is continuous (Hirji et al., 1991).

The confidence intervals discussed above were based on the ordinary definition of the *P*-value. Employing the mid-*P* definition, we can identify three additional tests (and related sets of confidence intervals):

- The Clopper–Pearson mid-*P* procedure. This is similar to the Clopper–Pearson procedure discussed above, except that $p[X \leq x_0]$ is replaced by $P[X < x_0] + 0.5\ P[X = x_0]$ and $P[X \geq x_0]$ is replaced by $0.5\ P[X = x_0] + P[X > x_0]$.
- The Wilson mid-*P* method, which defines the attained significance level as

$$\hat{\alpha} = P\{x|f(x|\theta_0) < f(x_0|\theta_0)\} + 0.5\ P\{x|f(x|\theta_0) = f(x_0|\theta_0)\}.$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

This represents the probability under H_0 of the set of values with lower probability than x_0 plus one half the probability of the set of values with probability equal to x_0. (Usually, there is no other x-value with the same probability as x_0, but sometimes there is exactly one other.) The null hypothesis is rejected if $\hat{\alpha} < \alpha$.

- The Crow mid-P method, based on the same procedure as the Crow method, but employing a mid-P value.

The Wilson and Crow methods (and their mid-P variants) can produce confidence "intervals" that are disconnected sets. Crow (1956) provides an example for the binomial case. When this situation occurred in the problems that we have examined, we adopted a conservative approach and enlarged the confidence set by filling in the gaps.

CONFIDENCE INTERVALS FOR THE ODDS RATIO

Using $f(x|T;\psi)$ for the probability law of the previous section, confidence intervals for the odds ratio can be obtained for each of the procedures—Clopper–Pearson, Wilson, and Crow—in both ordinary-P and mid-P versions. The roots, ψ_L and ψ_U, of the equations

$$\sum_{i=0}^{x} f(i|T;\psi_U) = \alpha/2$$

and

$$\sum_{i=0}^{x} f(i|T;\psi_L) = \alpha/2$$

form the Clopper–Pearson confidence intervals for ψ in the sense that

$$P[\psi_L \leq \psi \leq \psi_U] \geq 1 - \alpha$$

(Cornfield, 1956; Fisher, 1962). Confidence intervals can similarly be obtained corresponding to each of the test procedures described in the previous section.

Confidence intervals for ψ can also be obtained based on its approximate distribution. The conditional asymptotic distribution was stated without proof by Stevens (1951) and subsequently proven by Cornfield

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

(1956) and Hannan and Harkness (1963). If the four marginal frequencies are considered fixed, and if ψ is the true odds ratio, then we can get an asymptotic normal distribution for X. Let x be the unique permissible solution to the quadratic equation

$$\psi = \frac{x(n - T + x)}{(T - x)(m - x)}.$$

Then X is approximately normally distributed with

$$E(X|T;\psi) = x$$

and

$$\mathrm{Var}(X|T;\psi) = 1/W_{(X)}$$

where

$$W(x) = \frac{1}{x} + \frac{1}{T - x} + \frac{1}{m - x} + \frac{1}{n - T + x}.$$

The Cornfield and Woolf procedures are based on this approximate distribution:

- Cornfield's procedure (Cornfield, 1956). For a given observation X_0, let x_1 and x_2 denote the values of x (with $x_1 > x_2$) which provide a solution to

$$W(x)(X_0 - x)^2 = \chi^2_\alpha.$$

An approximate $1 - \alpha$ confidence interval is (ψ_L, ψ_U), where

$$\psi_L = \frac{(X_0 - x_1)(n - T_0 - x_1)}{(m - X_0 - x_1)(T_0 - x_1)}$$

and

$$\psi_U = \frac{(X_0 - x_2)(n - T_0 - x_2)}{(m - X_0 - x_2)(T_0 - x_2)}.$$

This procedure can be performed without a continuity correction as above, or with Yates's correction (Yates, 1934) for continuity, as

originally proposed. The appropriateness of this correction has been extensively debated (Conover, 1974; Cox, 1970; Grizzle, 1967; Haviland, 1990; Mantel and Greenhouse, 1968; Pearson, 1947; Plackett, 1964; Yates, 1984). To use the continuity correction, solve

$$W(x)(|X_0 - x| - 0.5)^2 = \chi^2_{\alpha,1}$$

- Woolf's (Logit) method (Woolf, 1955). An approximate $1-\alpha$ confidence interval for ψ is given by

  $$\exp\{L \pm z_\alpha \text{ s.e.}(L)\}$$

  where L is the logarithm of the sample odds ratio and s.e.(L) is the standard error of L given by

  $$\text{s.e.}(L) = \left(\frac{1}{x} + \frac{1}{m-x} + \frac{1}{y} + \frac{1}{n-y}\right)^{1/2}$$

This standard error will be infinite if any one of the four cells is zero. One can also perform this procedure using the Haldane (1940) and Anscombe (1956) correction. This adding of 1/2 to each cell value as a continuity correction results in an approximately unbiased estimator (Gart, 1962, 1971; Gart and Zweifel, 1967). We call this procedure the "adjusted Woolf" method. Fleiss (1979) indicates that the adjusted Woolf method yields confidence intervals appreciably narrower than those produced by the Clopper–Pearson method, and that the unadjusted Woolf method is an improvement. Agresti (1999) similarly notes that these methods perform "surprisingly well". Wilson and Langenberg (1999) investigated modifying the Woolf method so that the length of the confidence interval obtained was shortest while maintaining the desired level of confidence; they concluded that the shortest intervals did not appear to have important advantages over those obtained using the above methodology.

EVALUATION CRITERIA

Three criteria were used to evaluate the ten methods for generating confidence intervals discussed above. These included coverage probability, closeness of coverage probability to target, and length of confidence interval.

Coverage probability, $\mathrm{COV}(p_1, p_2)$. This is given by

$$\sum_{x=0}^{m}\sum_{y=0}^{n} P\{X = x\}P\{Y = y\}I\{\psi_L(x, y) \le \psi \le \psi_U(x, y)\}$$

where $I(\ )$ is the indicator functions and $(\psi_L(x, y), \psi_U(x, y))$ denotes the confidence interval for ψ computed from the outcome (x, y). The coverage probability $\mathrm{COV}(p_1, p_2)$ can be calculated numerically. We note that $p_2 = p_1/(p_1 + \psi(1 - p_1))$ so that $\mathrm{COV}(p_1, p_2)$can be regarded as a function of $\psi$and p_1. This is a continuous function of $p_1$but a discontinuous function of ψ.

''Average'' coverage was calculated in two ways. The first method consisted of a simple mean of the coverages for the 99 cases $p_1 = 0.01(0.01)0.99$, for fixed ψ. It may be argued that since the values of the parameters are fixed, there can be no rationale for averaging over them. However, use of a uniform prior distribution ''follows from the philosophical stance of the researcher that all potential values of the unknown parameter ... will be presumed equally likely until such time as sampling data provides evidence to the contrary'' (Rice, 1988, p. 21). The second method used to compute ''average'' coverage consisted of a weighted mean of the coverages for the 99 cases $\ln\psi = -2.45(0.05)2.45$ for fixed p_1, using weights corresponding to a uniform distribution on p_2. (That is, fixed p_1 plus a uniform distribution on p_2 implies a distribution on ψ.)

Closeness of coverage probabilities to target level. Given that the coverage, on average, of a method is close to $1-\alpha$, we would want the coverage for any given probability not to be too far from the target level. As a second criterion for evaluating the various methods for generating confidence intervals we therefore use the degree of closeness of the coverage probabilities to the target level, as operationalized by the mean square error (MSE).

Length of confidence intervals. As a final criterion, we compared all methods pair wise in terms of the probability that a given method produces shorter confidence intervals than a second method. The length of a confidence interval was calculated as the logarithm of the upper confidence limit minus the logarithm of the lower confidence limit. If one of the intervals being compared had an infinite upper or lower limit (on the log scale) while the other did not, then the latter was considered shorter. If both intervals have an infinite upper (lower) limit, then the one with the larger lower (smaller upper) limit was considered shorter.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

RESULTS

Confidence intervals with a 95% nominal level of significance were generated using each of the above methods for three problem sizes: $m = n = 10$; $m = 15, n = 20$; and $m = n = 20$. Coverage probabilities were computed for (a) $p_1 = 0.01(0.01)0.99$ for $\psi = 0.20$, (b) $p_1 = 0.01(0.01)0.99$ for $\psi = 1.00$, (c) $\ln \psi = -2.45(0.05)2.45$ for $p_1 = 0.2$, and (d) $\ln \psi = -2.45(0.05)2.45$ for $p_1 = 0.5$. The schemes with varying $\ln \psi$ yielded results consistent with those varying p_1. We will therefore not further discuss the former here, aside from the illustration that appears as Fig. 3, which we exhibit to show that the coverage is discontinuous in ψ.

Tables 1 and 2 present the mean coverage and mean square error for each of the methods and each of the three problem sizes when $\psi = 0.20$.

From Table 1 it can be seen that in each of the cases examined, the Clopper–Pearson rule produced confidence intervals that were extremely conservative, with noncoverage (=1 − coverage) averaging approximately 1% (vs. the nominal noncoverage level of 5%). This method also performed the worst based on the mean square error criterion (see Table 2). This came as no surprise, as the coverages generated by the various methods usually exceeded 95% and thus a conservative method (with a high mean coverage) would also have a larger MSE.

Figure 1 presents noncoverage plots for the Clopper–Pearson method for the case $m = n = 10$ and $\psi = 0.20$ using both ordinary-*P* and mid-*P* values. It can be seen that the noncoverage probabilities when using a ordinary-*P* value are very low over the whole range of p_1. The inadequacy of this method was noted for the binomial case by Angus

Table 1. Mean coverage, ψ=0.20, $p_1 = 0.01(0.01)0.99$.

	$m = n = 10$	$m = 15, n = 20$	$m = n = 20$
Clopper Pearson	0.9924	0.9866	0.9860
Clopper–Pearson mid-*P*	0.9796	0.9735	0.9707
Wilson	0.9796	0.9771	0.9740
Wilson mid-*P*	0.9645	0.9585	0.9599
Crow	0.9796	0.9745	0.9742
Crow mid-*P*	0.9645	0.9576	0.9586
Uncorrected Cornfield	0.9488	0.9558	0.9548
Corrected Cornfield	0.9899	0.9865	0.9858
Woolf	0.9705	0.9653	0.9627
Adjusted Woolf	0.9737	0.9709	0.9663

Table 2. Mean square error ($\times 1000$), ψ=0.20, $p_1 = 0.01(0.01)0.99$.

	$m = n = 10$	$m = 15, n = 20$	$m = n = 20$
Clopper–Pearson	1.818	1.397	1.333
Clopper–Pearson mid-P	1.006	0.651	0.577
Wilson	0.944	0.769	0.651
Wilson mid-P	0.380	0.300	0.206
Crow	0.944	0.659	0.662
Crow mid-P	0.380	0.297	0.183
Uncorrected Cornfield	0.152	0.299	0.128
Corrected Cornfield	1.598	1.376	1.298
Woolf	1.262	0.869	0.807
Adjusted Woolf	0.649	0.601	0.347

and Schafer (1984); they observed that the 5% level test often operates in only one tail so that the effective level is 2.5%. This result is not unexpected given the widespread criticism of the Fisher's exact test for its conservatism. Use of a mid-P value with the Clopper–Pearson method resulted in improved noncoverage, as shown in Fig. 1b, but this method is still quite conservative. Given the frequent calls for use of Fisher's test employing a mid-P value, the poor showing of this method is surprising.

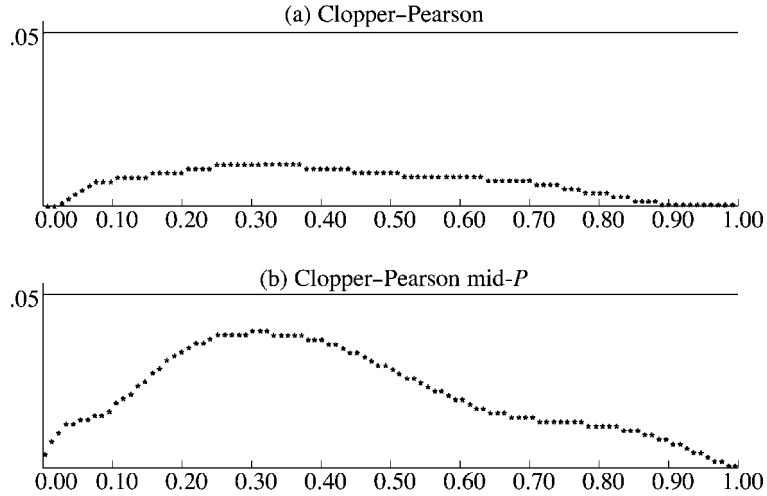

Figure 1. Noncoverage of (a) Clopper–Pearson and (b) Clopper–Pearson mid-P confidence intervals, $m = n = 10,\ \psi = 0.20,\ p_1 = 0.01(0.01)0.99$.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

The Wilson method produced confidence intervals whose coverage was much closer to the desired level than the Clopper–Pearson intervals, though it also is conservative. For the same situation as discussed above, average noncoverage ranged from 2 to 3% (see Fig. 2a and Table 1). The coverage of the mid-P version, shown in Fig. 2b, is much closer to target.

The Wilson methods require sorting outcomes by their probabilities. Occasionally, the probabilities of two events were equal. The algorithm we used was such that either both or neither of the outcomes were placed in the reject set. This tended to reduce the probability of the reject set and thus reduce the noncoverage probability. This situation is noticeable in plots of noncoverage by the Wilson mid-P method where $m = n$ and p_1 is fixed. The noncoverages when $\psi = 1.0$ ($\ln \psi = 0$) are dramatically lower than when ψ is near 1.0 (see Fig. 3). This is due to the symmetric distribution of the outcomes when $\psi = 1.0$, leading to many occurrences of the aforementioned situation.

For both the Clopper–Pearson and Wilson methods and for each size problem, use of a mid-P value rather than an ordinary P-value resulted in (a) coverages that were lower (and closer to the desired level) on average, (b) coverages that had smaller variability (as measured by MSE), and (c) confidence intervals that were shorter. However, as can be seen in

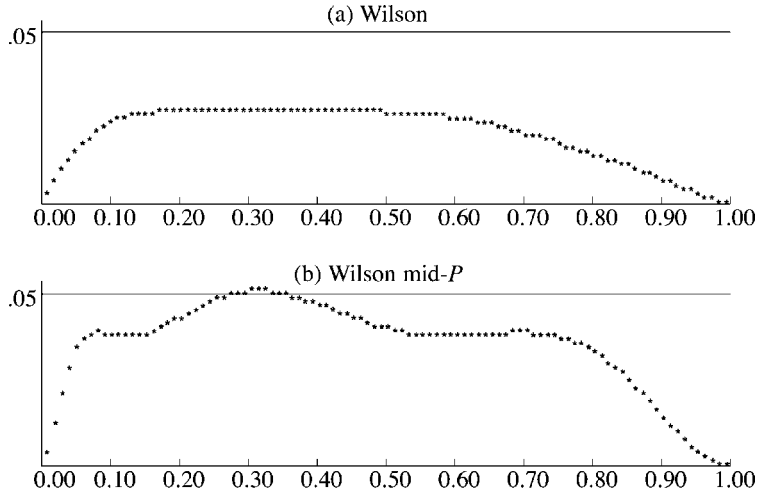

Figure 2. Noncoverage of (a) Wilson and (b) Wilson mid-P confidence intervals, $m = n = 10,\ \psi = 0.20,\ p_1 = 0.01(0.01)0.99$.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

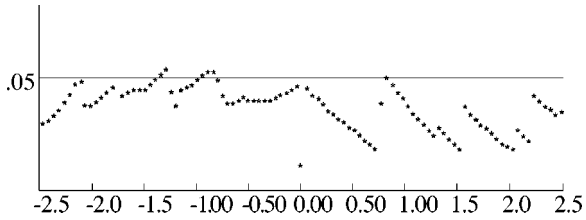

Figure 3. Noncoverage of Wilson mid-P confidence intervals, $m = n = 10$, $p_1 = 0.20$, $\ln\psi = -2.45(0.05)2.45$.

Fig. 2(b), use of a mid-P value with the Wilson method can result in a situation where the minimum coverage falls below the nominal confidence level. If it is critical that the nominal confidence level be met, then the Wilson mid-P method should not be used. However, if the nominal significance level is viewed instead as a desired level of significance, with actual levels of significance below this level being acceptable if not large, then this method performs well.

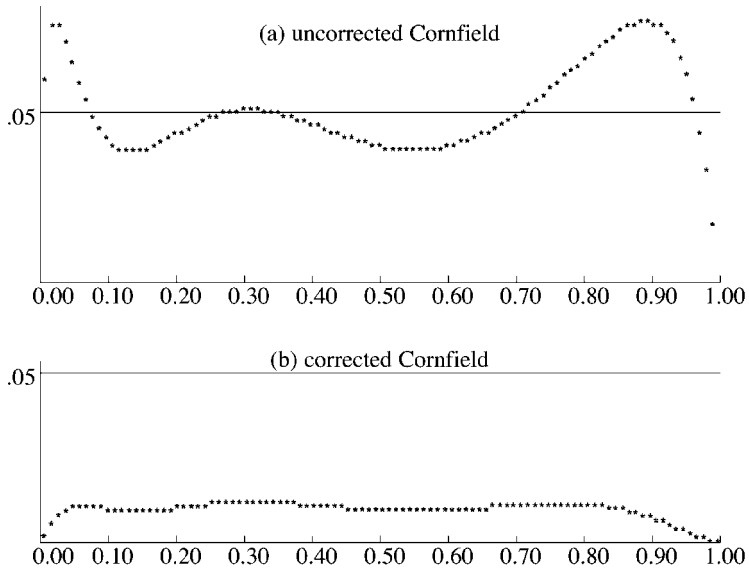

Figure 4. Noncoverage of (a) uncorrected Cornfield and (b) corrected Cornfield confidence intervals, $m = n = 10$, $\psi = 0.20$, $p_1 = 0.01(0.01)0.99$.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Of all of the "exact" method discussed here, the Crow methods performed the best, both in terms of mean coverage and mean square error. However, we will not consider them further as the methods are too complicated to prove popular. As Upton (1982) notes, "for a test to be useful ... it should be reasonably accurate, reasonably powerful and, above all, reasonably simple."

Use of the continuity correction greatly affected the confidence intervals generated by the Cornfield method. With the continuity correction this method had coverage plots similar to those of the extremely conservative Clopper–Pearson method (see Fig. 4b). Without the continuity correction, this method had a mean coverage in each instance nearer the target level than any other method (see Table 1). These results are consistent with those obtained by Gart and Thomas (1972). However, this method produces noncoverages that badly exceeded the target level on occasion, especially for extreme values of p_1 (less than 0.10 and greater than 0.90) and for smaller problem sizes (see Fig. 4a).

The Woolf (logit) method was generally less affected by the use or nonuse of the 1/2 adjustments (see Fig. 5). Over the range of problems examined, it produced mean coverage in the range of 2.6 to 3.7%. Based on this criterion, it performed worse than the Wilson mid-P and the uncorrected Crow methods, but better than either of the Clopper–Pearson methods. Similar results were obtained based on the MSE.

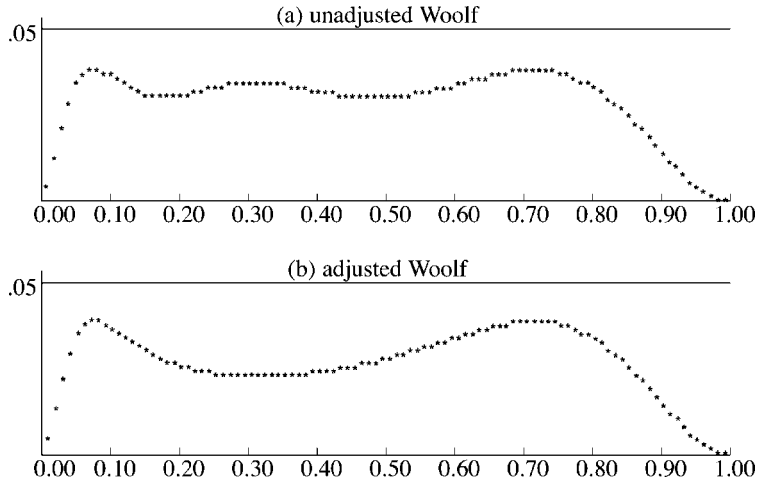

Figure 5. Noncoverage of (a) unadjusted and (b) adjusted Woolf confidence intervals, $m = n = 10,\ \psi = 0.20,\ p_1 = 0.01(0.01)0.99$.

Our findings with regard to probability of shorter intervals for the various methods were similar to those obtained using the other two methods of evaluation. Table 3 gives a display of the probabilities of shorter interval for each pair of methods. The Clopper–Pearson and corrected Cornfield methods, previously determined to be extremely conservative, generally had a zero probability of producing confidence intervals shorter than any other method, except when compared with each other. The only method against which the Clopper–Pearson had a positive probability of yielding shorter confidence intervals was the unadjusted Woolf method. However, at no point did this probability reach 0.50.

Table 3. Probability that ROW method gives shorter interval than COLUMN method.

	Cl–P	Cl–P mid-*P*	W	W mid-*P*	C	C Corrected	Wo	Adjusted Wo
(a) $m = 10, n = 10$								
Cl–P		0.0000	0.0000	0.0000	0.0000	0.2580	0.2881	0.0000
Cl–P, mid-*P*	0.9151		0.2688	0.0000	0.0000	0.8209	0.2881	0.0000
W	0.9151	0.4351		0.0375	0.0000	0.8209	0.2881	0.0000
W, mid-*P*	0.9151	0.8776	0.8776		0.2647	0.9151	0.8242	0.1667
C	0.9151	0.9151	0.9151	0.6504		0.9151	0.8964	0.0880
C, Corrected	0.6571	0.0942	0.0942	0.0000	0.0000		0.2881	0.0000
Wo	0.6083	0.6083	0.6083	0.0721	0.0000	0.6083		0.0000
Adjusted Wo	0.9151	0.9151	0.9151	0.7484	0.8271	0.9151	0.9151	
(b) $m = 15, n = 20$								
Cl–P		0.0000	0.0000	0.0000	0.0000	0.2641	0.1844	0.0000
Cl–P, mid-*P*	0.9499		0.3691	0.0000	0.0000	0.8731	0.1844	0.0000
W	0.9499	0.4336		0.0407	0.0000	0.8731	0.1844	0.0000
W, mid-*P*	0.9499	0.8970	0.9092		0.3306	0.9499	0.7829	0.1962
C	0.9499	0.9499	0.9499	0.6194		0.9499	0.9356	0.0965
C, Corrected	0.6858	0.0768	0.0769	0.0000	0.0000		0.1844	0.0000
Wo	0.7512	0.7512	0.7512	0.1527	0.0000	0.7512		0.0000
Adjusted Wo	0.9499	0.9499	0.9499	0.7537	0.8534	0.9499	0.9499	
(c) $m = 20, n = 20$								
Cl–P		0.0000	0.0000	0.0000	0.0000	0.2686	0.1796	0.0000
Cl–P, mid-*P*	0.9601		0.3653	0.0000	0.0000	0.8984	0.1796	0.0000
W	0.9601	0.4637		0.0308	0.0000	0.8984	0.1796	0.0000
W, mid-*P*	0.9601	0.9380	0.9294		0.3747	0.9601	0.8020	0.2946
C	0.9601	0.9601	0.9601	0.5854		0.9601	0.9506	0.0857
C, Corrected	0.6916	0.0617	0.0617	0.0000	0.0000		0.1796	0.0000
Wo	0.7710	0.7710	0.7710	0.1486	0.0000	0.7710		0.0000
Adjusted Wo	0.9601	0.9601	0.9601	0.6656	0.8745	0.9601	0.9601	

Note: CI–P = Clopper–Pearson, W = Wilson, C = Cornfield, and Wo = Woolf.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

DISCUSSION

Gart and Thomas (1972) analyzed four methods for obtaining confidence intervals for the odds ratio in 2×2 tables, including the adjusted Woolf and corrected Cornfield methods. The problem sizes they examined were the same as those used here. They concluded that the adjusted Woolf method gave limits which are much too narrow and that Cornfield's (corrected) method was the preferred method for computing 95% confidence intervals (when the expectation in each cell exceeded one.) Their conclusions were based on calculations of (a) the number of times each method gave a "legitimate" result (meaning $\psi_L \geq 0$) (b) the MSE of the probabilities computed over this set of tables, and (c) the mean of the sum of the two tails' p-values. Calculations (b) and (c) were done as ordinary averages and thus do not take into consideration the relative probability of each of the tables. Based on a calculation of coverage, in which each table is weighted by its probability, we determined that the adjusted Woolf method gives limits that are too wide rather than too narrow. Furthermore, as previously mentioned, the corrected Cornfield procedure yields extremely conservative confidence intervals.

In all instances, use of mid-P (vs. ordinary P-value) resulted in shorter confidence intervals. This is consistent with the results of prior studies which have found that use of a mid-P value helps correct for the conservatism of the exact test (Barnard, 1989; Berry and Armitage, 1995; Hirji, 1991; Hirji et al., 1991; Upton, 1992). Similarly, use of the Wilson method produced shorter confidence intervals than the Clopper–Pearson method, regardless of P-value definition. The probability that the Wilson (ordinary

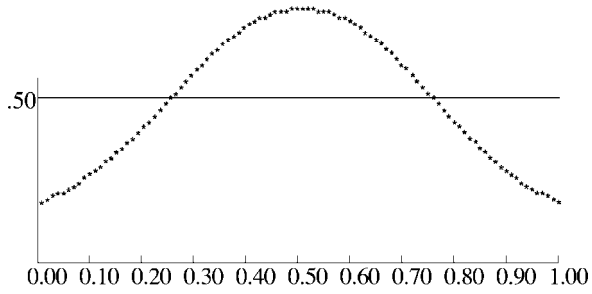

Figure 6. Probability that the Wilson mid-P method yields shorter confidence intervals than the uncorrected Cornfield method, $\psi = 1.00$, $p_1 = 0.01(0.01)0.99$, $m = n = 10$.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

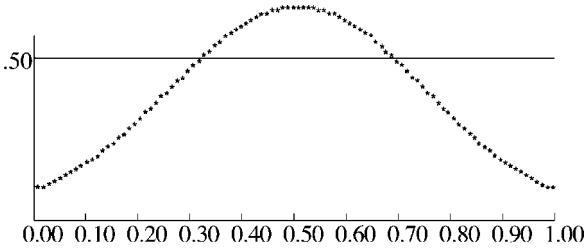

Figure 7. Probability that the Wilson mid-*P* method yields shorter confidence intervals than the adjusted Woolf method, $\psi = 1.0$, $p_1 = 0.01(0.01)0.99$, $m = n = 10$.

P-value) method produced shorter confidence intervals than the Clopper–Pearson mid-*P* method was consistently less than 0.50, indicating that the definition of *p*-value had a larger effect on the length of the confidence interval than the selection rule. All of these results are consistent with our previous conclusions based on coverage probabilities.

Figures 6 and 7 show examples of the length-comparison plots for some of the methods that performed well based on the previous criteria in the case where $\psi = 1$ and $m = n = 10$. Length comparison plots for the Wilson (ordinary *P*-value) and uncorrected Cornfield methods (see Fig. 6) indicate that the two methods are approximately equally desirable.

The unadjusted Woolf method was consistently outperformed by the Wilson mid-*P* method according to this length of interval criterion. This was due in part to the relatively large number of tables where the limits were zero to infinity, resulting in wide confidence intervals.

The adjusted Woolf and Wilson mid-*P* methods are compared in Fig. 7. Surprisingly, the adjusted Woolf method produced confidence intervals that were relatively short. (In terms of length of confidence interval, the only method that sometimes outperformed it was the Crow mid-*P* method, which requires prohibitive calculations.)

RECOMMENDATIONS

Based on our three criteria, Cornfield's method, without the continuity correction, performed the best of the methods examined here. The reader should decide whether the noncoverage above target, as illustrated by Fig. 4a, is acceptable. Crow's method, with use of a mid-*P*, while performing well, is too cumbersome to calculate to be readily used in practice and is only a slight improvement over the Wilson mid-*P* method.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

The Woolf method (with the Haldane–Anscombe correction), while slightly inferior in terms of coverage, performed equally well in terms of length of confidence interval. The intervals for this method are trivial to calculate, and this makes us recommend use of that method for small sample confidence intervals for the odds ratio.

REFERENCES

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statist. Sci.* 7(1):131–177.

Agresti, A. (1999). On logit confidence intervals for the odds ratio with small samples. *Biometrics* 55:597–602.

Agresti, A., Min, Y. (2002). Unconditional small-sample confidence intervals for the odds ratio. *Biostatistics* 3:379–386.

Angus, J. E., Schafer, R. E. (1984). Improved confidence statements for the binomial parameter. *Am. Statist.* 38(3):189–191.

Anscombe, F. J. (1956). On estimating binomial response relations. *Biometrika* 43:461–464.

Barnard, G. A. (1945). A new test for 2×2 tables. *Nature* 156:388.

Barnard, G. A. (1947). Significance tests for 2×2 tables. *Biometrika* 34:123–128.

Barnard, G. A. (1989). On alleged gains in power from lower *P*-values. *Statist. Med.* 8:1469–1477.

Barnard, G. A. (1990). Must clinical trials be large? The interpretation of P-values and the combination of test results. *Statist. Med.* 9: 601–614.

Berry, G., Armitage, P. (1995). Mid-*P* confidence intervals: A brief review. *Statist.* 44(4):417–423.

Clopper, C., Pearson, E. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26:404–413.

Conover, W. J. (1974). Some reasons for not using the Yates continuity correction on 2×2 contingency tables. *Am. Statist.* 69(346): 374–376.

Cornfield, J. (1956). A statistical problem arising from retrospective studies. In: Neyman, J., ed. *Proceedings of the Third Berkeley Symposium On Mathematical Statistics and Probability,* IV. Berkeley: University of California Press, pp. 135–148.

Cox, D. R. (1970). The continuity correction. *Biometrika* 57(1):217–219.

Crow, E. (1956). Confidence intervals for a proportion. *Biometrika* 43: 423–435.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Fisher, R. (1935). The logic of inductive inference. *J. R. Statist. Soc.* 98(1):39–82.
Fisher, R. (1962). Confidence limits for a cross-product ratio. *Aust. J. Statist.* 4(1):41.
Fleiss, J. (1979). Confidence intervals for the odds ratio in case-control studies: The state of the art. *J. Chronic Dis.* 32:69–77.
Gart, J. (1962). Approximate confidence limits for the relative risk. *J. R. Statist. Soc. Ser. B* 24:454–463.
Gart, J. (1971). The comparison of proportions: A review of significance tests, confidence intervals and adjustments for stratification. *Rev. Int. Statist. Inst.* 39(2):148–169.
Gart, J. J., Zweifel, J. R. (1967). On the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika* 54(1 and 2):181–187.
Gart, J., Thomas, D. (1972). Numerical results on approximate confidence limits for the odds ratio. *J. R. Statist. Soc. Ser. B* 34:441–447.
Grizzle, J. E. (1967). Continuity correction in the χ^2-test for the 2×2 table. *Am. Statist.* 21(4):28–32.
Haber, M. (1986). An exact unconditional test for the 2×2 comparative trial. *Psychol. Bull.* 99(1):129–132.
Haldane, J. B. S. (1940). The mean and variance of the moments of χ^2, when used as a test of homogeneity, when expectations are small. *Biometrika* 29:133–143.
Hannan, J., Harkness, W. (1963). Normal approximation to the distribution of two independent binomials, conditional on a fixed sum. *Ann. Math. Statist.* 34:1593–1595.
Haviland, M. (1990). Yates's correction for continuity and the analysis of 2×2 contingency tables. *Statist. Med.* 9:363–367.
Hirji, K. F. (1991). A comparison of exact, mid-*P*, and score tests for matched case-control studies. *Biometrics* 47:487–496.
Hirji, K. F., Tan, S., Elashoff, R. (1991). A quasi-exact test for comparing two binomial proportions. *Statist. Med.* 10:1137–1153.
Lancaster, H. O. (1949). The combination of probabilities arising from data in discrete distributions. *Biometrika* 36:370–382.
Lancaster, H. O. (1952). Statistical control of counting experiments. *Biometrika* 39:419–422.
Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. New York: Wiley.
Mantel, N., Greenhouse, S. (1968). What is the continuity correction? *Am. Statist.* 22(5):27–30.
Miettinen, O. (1974). Comment. *J. Am. Statist. Assoc.* 69:380–382.
Pearson, E. S. (1947). The choice of statistical tests illustrated on the interpretation of data classed in a 2×2 table. *Biometrika* 34:139–167.

Plackett, R. L. (1964). The continuity correction in 2×2 tables. *Biometrika* 51(3 and 4):327–337.

Rice, W. R. (1988). A new probability model for determining exact *P*-values for contingency tables when comparing binomial proportions. *Biometrics* 44:1–22.

Sterne, T. (1954). Some remarks on confidence or fiducial limits. *Biometrika* 41:275–278.

Stevens, W. L. (1951). Mean and variance of an entry in a contingency table. *Biometrika* 38:468–470.

Upton, G. J. G. (1982). A comparison of alternative tests for the 2×2 comparative trial. *J. R. Statist. Soc. Ser. A* 145(Part 1):86–105.

Upton, G. J. G. (1992). Fisher's exact test. *J. R. Statist. Soc. Ser. A* 155(Part 3):395–402.

Wilson, E. B. (1942). On confidence intervals. *Proc. Nat. Acad. Sci.* 28:88–93.

Wilson, P., Langenberg, P. (1999). Usual and shortest confidence intervals on odds ratios from logistic regression. *Am. Statist.* 53(4):332–335.

Williams, D. A. (1988). Tests for differences between several small proportions. *Appl. Statist.* 37(3):421–434.

Woolf, B. (1955). On estimating the relation between blood group and disease. *Ann. Human Genet.* 19:251–253.

Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *J. R. Statist. Soc. Ser. B* 1(2):217–235.

Yates, F. (1984). Tests of significance for 2×2 contingency tables. *J. R. Statist. Soc. Ser. A* 147(Part 3):426–463.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016