*Biostatistics* (2002), **3**, 3, *pp.* 379–386
*Printed in Great Britain*

# Unconditional small-sample confidence intervals for the odds ratio

ALAN AGRESTI*, YONGYI MIN

*Department of Statistics, University of Florida, Gainesville, Florida 32611-8545, USA*
aa@stat.ufl.edu

## SUMMARY

The traditional approach to ‘exact’ small-sample interval estimation of the odds ratio for binomial, Poisson, or multinomial samples uses the conditional distribution to eliminate nuisance parameters. This approach can be very conservative. For two independent binomial samples, we study an unconditional approach with overall confidence level guaranteed to equal at least the nominal level. With small samples this interval tends to be shorter and have coverage probabilities nearer the nominal level.

*Keywords*: Binomial distribution; Conditional tests; Exact inference; Two-by-two contingency table.

## 1. INTRODUCTION

Table 1 shows results from a recent study of preterm infants. Of infants born between 27 and 32 weeks gestational age who required mechanical ventilation, the study compared 26 infants with periventricular leukomalacia (PVL) and 26 with normal development on various characteristics. Table 1 compares the groups on a neonatal adverse event (oliguria). The sample odds ratio equals 2.08. The researchers test of whether the true odds ratio $\theta = 1$ reported a $P$-value from Fisher’s exact test. This paper considers interval estimation of $\theta$ using small-sample confidence intervals that guarantee that the true coverage probability is at least the nominal level. The standard approach for doing this (Cornfield, 1956) uses a conditional distribution in order to eliminate a nuisance parameter. It provides the 95% confidence interval (0.1, 127.3). (The study samples were matched, but since information is not provided to utilize the matching, as in the source paper we treat samples as independent.)

With small samples a conditional approach has the disadvantage of increasing the discreteness and hence the conservatism of methods. This is well known for testing $H_0 : \theta = 1$, in which Fisher’s exact test can be very conservative (for example, Suissa and Shuster, 1985 showed comparisons with an unconditional test). This note discusses an unconditional approach to interval estimation for $\theta$ based on two independent binomial samples. This method also guarantees that the true coverage probability is at least the nominal level. For Table 1 it provides the 95% interval (0.2, 29.4).

Section 2 summarizes methods of forming confidence intervals for odds ratios. Section 3 introduces the unconditional approach. Section 4 compares coverage probabilities and lengths between conditional and unconditional intervals. With small samples, the unconditional interval tends to be shorter and the coverage probability tends to be closer to the nominal level than for conditional intervals. Section 5 discusses implementation.

*To whom correspondence should be addressed

Table 1. *Comparison of two infant groups on an adverse event* (*Okumura* et al., 2001)

| | Response | |
|---|---|---|
| Group | Yes | No |
| PVL | 2 | 24 |
| Control | 1 | 25 |

## 2. CONFIDENCE INTERVALS FOR THE ODDS RATIO

Consider a $2 \times 2$ contingency table, based on two independent binomial samples within rows or within columns or a multinomial sample over the four cells or four independent Poisson samples. Let $\{n_{ij}\}$ denote cell counts, with $n = \sum n_{ij}$ and $\hat{\theta} = (n_{11}n_{22})/(n_{12}n_{21})$. Let $z_{\alpha/2}$ denote the $1 - \alpha/2$ standard normal quantile. The large-sample approach for confidence intervals for $\theta$ relies on the asymptotic normality of $\log(\hat{\theta})$. The most popular large-sample $100(1-\alpha)\%$ interval (Woolf, 1955) exponentiates endpoints of

$$\log(\hat{\theta}) \pm z_{\alpha/2}\sqrt{n_{11}^{-1} + n_{12}^{-1} + n_{21}^{-1} + n_{22}^{-1}}.$$

Adjusted versions of this formula handle zero counts (for example, Gart, 1966 and Agresti, 1999). This interval inverts the Wald test for the log odds ratio using a standard error derived with the delta method. Perhaps surprisingly for a Wald method, it works quite well. Alternatively, one could invert the large-sample likelihood-ratio or score chi-squared tests about $\theta$.

With large-sample methods, at fixed parameter values the coverage probability converges to the nominal level as $n$ increases. However, for tables such as Table 1, one might feel safer using a small-sample approach. Construction of an interval guaranteed to achieve at least the nominal level for any sample size requires dealing with nuisance parameters. For instance, with a multinomial sample one can parametrize the mass function using $\theta$ and a marginal row and a marginal column probability. Historically, the most popular approach eliminates nuisance parameters by conditioning on their sufficient statistics. For inference about $\theta$ with standard sampling models, one conditions on row and column marginal totals. The result is the hypergeometric distribution

$$P(n_{11} = t | \{n_{i+}\}, \{n_{+j}\}; \theta) = \frac{\binom{n_{1+}}{t}\binom{n-n_{1+}}{n_{+1}-t}\theta^t}{\sum_s \binom{n_{1+}}{s}\binom{n-n_{1+}}{n_{+1}-s}\theta^s},$$

for $\max(0, n_{1+} + n_{+1} - n) \leqslant t \leqslant \min(n_{1+}, n_{+1})$.

With this distribution, Cornfield (1956) constructed a confidence interval by inverting two separate one-sided exact tests, each of size at most $\alpha/2$. For observed value $t_{\text{obs}}$ for $n_{11}$, the interval is defined by values $(\theta_{\text{L}}, \theta_{\text{U}})$ satisfying

$$P(n_{11} \leqslant t_{\text{obs}}; \theta_{\text{U}}) = \alpha/2, \qquad P(n_{11} \geqslant t_{\text{obs}}; \theta_{\text{L}}) = \alpha/2. \tag{1}$$

This confidence interval consists of the collection of $\theta_0$ for which the exact $P$-value exceeds $\alpha/2$ in testing $H_0 : \theta = \theta_0$ against each one-sided alternative. The exactness refers to the conditional distribution being free of nuisance parameters. The actual confidence coefficient, defined as the infimum of the coverage probabilities for all possible $\theta$, has the nominal level as a lower bound. Since the distribution of $n_{11}$ is discrete, for any value $\theta$ the coverage probability may actually be much greater than $1 - \alpha$ (Neyman, 1935).

An alternative 'exact' conditional confidence interval for $\theta$ inverts a single two-sided test instead of two separate one-sided tests. Sterne (1954) used this approach for interval estimation of a binomial parameter, and Baptista and Pike (1977) adapted it to the odds ratio. Their test formed the acceptance region using ordered null probabilities, with the highest added first. This leads to length optimality. Another approach inverts a two-sided test using a standard test criterion such as the likelihood-ratio or score statistic, but using its exact conditional distribution rather than its asymptotic distribution. For various parameters in 2 × 2 tables, Agresti and Min (2001a), discussed the advantage of basing an interval on a two-sided family of tests rather than two equal-tailed one-sided tests. For Table 1, for instance, method (1) gives (0.10, 127.2) whereas inverting the two-sided 'exact' conditional score test gives (0.15, 62.7).

One can also use the conditional distribution for large-sample inference. The likelihood-based interval is the set of odds ratio values for which twice the conditional log likelihood falls within $z^2_{\alpha/2}$ of its maximum (Aitkin *et al.*, 1989, p. 198). Or, one can invert the conditional score test. This interval consists of odds ratios resulting from expected frequencies having the same margins as the observed counts and for which the Pearson chi-squared statistic is no greater than $z^2_{\alpha/2}$ (Cornfield, 1956).

## 3. Unconditional confidence intervals for odds ratio

In a 2 × 2 table, assume now that $y_i = n_{i1}$ is a binomial variate with parameter $\pi_i$ and index $n_i = n_{i1} + n_{i2}$, and $y_1$ and $y_2$ are independent. (The approach below also applies for a single multinomial sample, after conditioning on the row totals.) Let $p_i = y_i/n_i$. The product binomial probability mass function of $(y_1, y_2)$ is

$$f(y_1, y_2; n_1, n_2, \pi_1, \pi_2) = \binom{n_1}{y_1}\pi_1^{y_1}(1-\pi_1)^{n_1-y_1}\binom{n_2}{y_2}\pi_2^{y_2}(1-\pi_2)^{n_2-y_2}.$$

For given $\theta$, $\pi_1$ is determined by $\pi_2$ and $\theta$. Thus, one can express this joint mass function as $f(y_1, y_2; n_1, n_2, \theta, \pi_2)$. For inference about $\theta$, $\pi_2$ is a nuisance parameter.

To guarantee an upper bound $\alpha$ for the size, an unconditional test eliminates $\pi_2$ by maximizing the $P$-value over its values. For testing $H_0 : \theta = \theta_0$ with a test statistic $T$ having observed value $t_{\text{obs}}$ such that larger values provide more evidence against $H_0$, the unconditional $P$-value uses $f(y_1, y_2; n_1, n_2, \theta_0, \pi_2)$ to calculate

$$P(\theta_0) = \sup_{\pi_2} P[T \geqslant t_{\text{obs}}; \theta_0, \pi_2]. \tag{2}$$

A small-sample unconditional confidence interval inverts a small-sample unconditional test. The interval is the set of $\theta_0$ for which $P(\theta_0) > \alpha$. The total number of outcomes of the two types (i.e. the column totals) is not fixed, so the relevant distribution is less discrete than the hypergeometric. This provides the potential to reduce conservatism. However, the potential also exists to increase conservatism by forming the $P$-value using the worst-case scenario (2) for the nuisance parameter. The method is invariant to expressing $f$ as $f(y_1, y_2; n_1, n_2, \theta, \pi_2)$ or $f(y_1, y_2; n_1, n_2, \theta, \pi_1)$.

For calculating the $P$-value (2), $T$ can be a standard criterion such as the score or likelihood-ratio statistic. Results shown in this paper use the score statistic. The score test of $H_0 : \theta = \theta_0$ has statistic of the form (Miettinen and Nurminen, 1985)

$$T = [n_1(p_1 - \hat{\pi}_1(\theta_0))]^2\left[\frac{1}{n_1\hat{\pi}_1(\theta_0)(1-\hat{\pi}_1(\theta_0))} + \frac{1}{n_2\hat{\pi}_2(\theta_0)(1-\hat{\pi}_2(\theta_0))}\right],$$

where $\hat{\pi}_i(\theta_0)$ denotes the ML estimate of $\pi_i$ under the constraint that

$$\frac{\hat{\pi}_1(\theta_0)/(1-\hat{\pi}_1(\theta_0))}{\hat{\pi}_2(\theta_0)/(1-\hat{\pi}_2(\theta_0))} = \theta_0.$$

For Table 1, inverting the unconditional test with score statistic gives 95% confidence interval (0.23, 29.4), much narrower than the conditional intervals of (0.10, 127.2) and (0.15, 62.7).

Let $L(\theta, \pi_2) = \log f(y_1, y_2; n_1, n_2, \theta, \pi_2)$. Alternatively, in (2) one could use the profile log likelihood $L(\theta_0, \hat{\pi}_2(\theta_0)) = \sup_{\pi_2} \log[f(y_1, y_2; n_1, n_2, \theta_0, \pi_2)]$, which occurs in the likelihood-ratio statistic,

$$T = -2[L(\theta_0, \hat{\pi}_2(\theta_0)) - L(\hat{\theta}, p_2)].$$

Or, one could adapt the Sterne (1954) approach of using ordered null probabilities to this unconditional setting, letting $T = -\log f(y_1, y_2; n_1, n_2, \theta, \pi_2)$, in which case (2) corresponds to

$$P(\theta_0) = \sup_{\pi_2} P[f(Y_1, Y_2; n_1, n_2, \theta_0, \pi_2) \leqslant f(y_1, y_2; n_1, n_2, \theta_0, \pi_2)].$$

Note this differs from the profile likelihood approach, which moves the sup inside $P$ and applies it to both $f$ terms in this expression in evaluating possible samples.

## 4. COMPARISONS OF COVERAGE PROBABILITIES AND LENGTHS

We studied coverage probabilities and expected lengths of small-sample conditional and unconditional confidence intervals for the odds ratio for various $n$ with independent binomial samples. This section briefly summarizes results. More detailed comparisons are available in a technical report (Agresti and Min, 2001b).

For $n_1 = n_2 = 10$, Figures 1 and 2 show coverage probabilities for three methods with nominal level 0.95: the conditional 'exact' intervals (1) based on inverting two one-sided tests (i.e. the Cornfield tail interval) and (2) based on a single two-sided 'exact' score test, and (3) the unconditional small-sample interval based on the score test. The figures plot the coverage probability as a function of $\pi_1$: Figure 1 when $\pi_2 = 0.1, 0.3$, and 0.5, and Figure 2 when $\theta = 1.0, 2.0$, and 4.0. (The two conditional approaches have identical coverage probability in the $\theta = 1.0$ graph.) The unconditional approach tends to give results closer to the nominal level. Similar results occurred for $(n_1, n_2) = (10, 20), (20, 20)$, and (30, 30), although the difference between unconditional and conditional methods becomes less pronounced as $\{n_i\}$ increases.

For the sample sizes considered, the unconditional interval is almost always contained in the Cornfield conditional interval and is usually contained in the two-sided conditional interval. The unconditional interval tends to be shorter than the latter except when $y_1$ and $y_2$ are near the middle of their range, in which case the lengths tend to be similar. In such cases the discreteness with the conditional approach is less severe. The unconditional interval tended to be consistently shorter when either $y_i$ was near the boundary, especially when $y_i \leqslant 1$ or $y_i \geqslant n_i - 1$.

We also conducted an evaluation using randomly sampled $(\pi_1, \pi_2)$ pairs over the unit square, comparing expected lengths (conditional on no count equaling 0) and coverage probabilities when $n_1 = n_2 = 10$, 20, or 30. Similar results occurred. The one situation in which the conditional two-sided method yielded smaller coverage probability and slightly narrower intervals was when the true log odds ratio is extremely large in absolute value. For instance, with $\pi_2 = 0.1$, this happens for many $\pi_1$ values above 0.6, especially when sample sizes increase and become more unbalanced. In practice, more important are relatively small true effects. For $(\pi_1, \pi_2)$ pairs sampled randomly over $0.5 < \theta < 2.0$, the unconditional approach fares very well. For instance, the proportion of parameter pairs for which the actual coverage probability of nominal 95% intervals exceeds 0.96 is 1.00 for the two conditional methods, but for the unconditional approach it is 0.34 when $n_1 = n_2 = 10$, 0.21 when $n_1 = n_2 = 20$, and 0.16 when $n_1 = n_2 = 30$.

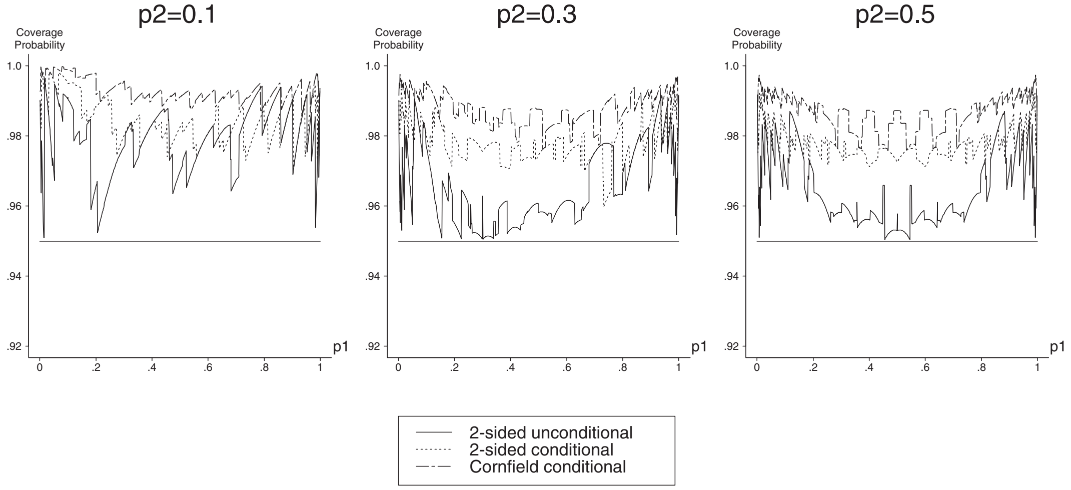

Fig. 1. Coverage probabilities for 95% confidence intervals for odds ratio, when $n_1 = n_2 = 10$.

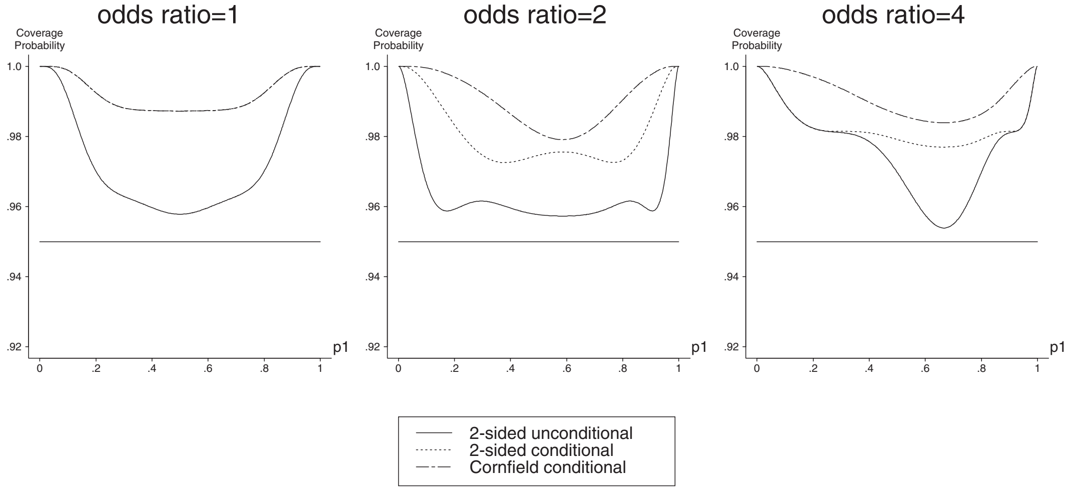

Fig. 2. Coverage probabilities for 95% confidence intervals for odds ratio, when $n_1 = n_2 = 10$.

## 5. Implementing the unconditional method

In implementing the unconditional method, we used the Cornfield conditional interval limits (1) as starting values. For the upper limit $\theta_\mathrm{U}$ of the Cornfield interval, suppose $P(\theta_\mathrm{U}) = \sup_{\pi_2} P[T \geqslant t_\mathrm{obs}; \theta_\mathrm{U}, \pi_2] < \alpha$. Then we iteratively decreased $\theta$ from $\theta_\mathrm{U}$ with a fine grid search until finding $\theta_0$ for which $P(\theta_0) = \alpha$. Likewise we iteratively searched from the lower limit of Cornfield's interval until finding $\theta_0$ for which $P(\theta_0) = \alpha$. For any particular $\theta$ we evaluated $P(\theta)$ by calculating $P[T \geqslant t_\mathrm{obs}; \theta, \pi_2]$ both for a fine grid search of $\pi_2$ values and for a separate large random selection of $\pi_2$ values. It is possible, though seems quite rare, that the set of $\theta_0$ for which $P(\theta_0) > \alpha$ is not a connected interval. Thus, we supplemented this initial interval determination by an intensive random search of parameter points corresponding to $\theta$ outside the working limits. The final interval is the set of $\theta_0$ from the smallest to the largest values satisfying $P(\theta_0) > \alpha$.

A disadvantage of the unconditional approach is that calculations are complex. There is no assurance that an algorithm such as just described produces the correct interval. After generating the intervals for all possible samples with the given $n_1$ and $n_2$, we tested results by randomly generating $100\,000(\pi_1, \pi_2)$ pairs in the unit square. At each $(\pi_1, \pi_2)$ we calculated the coverage probability of the unconditional intervals. For the examples we considered, the actual probability was never less than the nominal level.

A criticism of the unconditional approach, vocally made by Fisher about Barnard's small-sample test of $\pi_1 = \pi_2$, is that its sample space includes outcome totals very different from observed ones and potential probabilities very different from those suggested by the data. An alternative inversion of unconditional tests uses the Berger and Boos (1994) method of eliminating the nuisance parameter. It takes the supremum in (2) over a high confidence region (for example, 99.9%) for $\pi_2$ and adjusts the $P$-value (for example, by adding 0.001) so the overall nominal size is not exceeded. In a sense this addresses the criticism by restricting attention to nuisance parameter values supported by the data. We also implemented this approach for the unconditional score interval. However, it was similar and not consistently better or consistently poorer than the score interval based on taking the supremum over all possible $\pi_2$ values.

## 6. COMMENTS AND CHALLENGES FOR EXTENSIONS

For interval estimation of the odds ratio with independent binomial samples, several options now can ensure the true confidence level is at least the nominal one. These include Cornfield's 1956 conditional interval based on (1), conditional intervals inverting various two-sided tests such as score or likelihood-ratio tests using the hypergeometric distribution, and unconditional intervals using the same test statistics but with the product binomial distribution. Our research indicates that with small samples, inversion of a two-sided unconditional test tends to give narrower intervals than conditional methods, with actual coverage probability nearer the nominal level. The improvement increases as the discreteness does, that is, as sample sizes decrease and as the true probabilities both approach 0 or 1.

Fisher's criticism of the unconditional approach is part of a considerable debate over the years about ways of testing whether $\theta = 1$. See Sprott (2000, Section 6.4.4) for a recent cogent support of Fisher's arguments. The same arguments apply to interval estimation. Note, though, that a general approach to interval estimation is not possible with the conditional method, since it does not apply with non-canonical parameters. Thus, the unconditional approach is the basis of small-sample interval estimation for other parameters in $2 \times 2$ contingency tables, such as the difference of proportions and the relative risk (see, for example, Coe and Tamhane, 1993 and Agresti and Min (2001a). It seems surprising that the unconditional approach is not used with the odds ratio. Recently, however, Troendle and Frank (2001) proposed an unconditional interval by inverting a test using the sample odds ratio itself as test statistic. This also shows improvement over conditional intervals. We believe it more appropriate to use the likelihood-ratio or score test statistic; two samples with the same odds ratio could provide quite different evidence about whether a particular $\theta$ is plausible for the true value. See Chan and Zhang (1999) for related remarks about the difference of proportions.

The unconditional approach becomes computationally more difficult to implement as $\{n_i\}$ increases. As this happens, though, there is less need for it, since the conditional approach suffers less from discreteness. A challenging research problem is comparing asymptotic behavior of unconditional and conditional methods. In how broad a sense are these methods asymptotically equivalent to each other and to ordinary likelihood-based large-sample methods?

A more challenging computational problem yet is to extend the unconditional method to interval estimation of a common odds ratio for stratified $2 \times 2$ tables. With $K$ strata of independent binomials with $n$ observations per binomial, the number of possible samples is on the order of $n^{2K}$; with a grid search of $g$ values for each of the $K$ nuisance parameters, the probabilities of these samples must be evaluated for $g^K$ combinations of nuisance parameter values, which is itself problematic with a suitably large value for $g$. Simultaneously, to invert the unconditional test one must evaluate these probabilities with tests over a space of $\theta$ values. Currently, this extension requires special methods to be feasible, such as a way to determine iteratively a guess for the nuisance parameter combination that yields the supremum $P$-value. Rather than dealing with this supremum, a rougher approximation calculates $P(\theta_0)$ at the ML estimates of the binomial parameters within the $K$ strata under the constraint that the $K$ odds ratios equal $\theta_0$. This eliminates the search over the nuisance parameters but still requires enumerating all the possible samples and their product binomial probabilities. Although this uses binomial distributions rather than large-sample normality, the guarantee no longer holds that the true coverage probability is bounded below by the nominal level. The extension to stratified tables is an interesting research problem, but probably less important in practice since the discreteness issue with the conditional approach diminishes greatly as $K$ increases.

## Acknowledgements

This research was partially supported by grants from NIH and NSF.

## References

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

Agresti, A. and Min, Y. (2001a). On a small-sample confidence intervals for parameters in discrete distributions. *Biometrics* **57**, 963–971.

Agresti, A. and Min, Y. (2001b). Unconditional small-sample confidence intervals for the odds ratio, *Technical Report no. 2001-016*, Department of Statistics, University of Florida.

Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989). *Statistical Modelling in GLIM*. Oxford: Clarendon.

Baptista, J. and Pike, M. C. (1977). Exact two-sided confidence limits for the odds ratio in a 2 × 2 table. *Journal of the Royal Statistical Society,* Series C **26**, 214–220.

Berger, R. R. and Boos, D. D. (1994). *P* values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* **89**, 1012–1016.

Chan, I. S. F. and Zhang, Z. (1999). Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics* **55**, 1202–1209.

Coe, P. R. and Tamhane, A. C. (1993). Small sample confidence intervals for the difference, ratio and odds ratio of two success probabilities. *Communications in Statistics, Part B—Simulation and Computation* **22**, 925–938.

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* 4, pp. 135–148.

Gart, J. J. (1966). Alternative analyses of contingency tables. *Journal of the Royal Statistical Society,* Series B **28**, 164–179.

Miettinen, O. and Nurminen, M. (1985). Comparative analysis of two rates. *Statistics in Medicine* **4**, 213–226.

Neyman, J. (1935). On the problem of confidence limits. *Annals of Mathematical Statistics* **6**, 111–116.

Okumura, A. *et al.* (2001). *Pediatrics* **107**, 469–475.

Sprott, D. A. (2000). *Statistical Inference in Science*. New York: Springer.

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

SUISSA, S. AND SHUSTER, J. J. (1985). Exact unconditional sample sizes for the 2 by 2 binomial trial. *Journal of the Royal Statistical Society,* Series A **148**, 317–327.

TROENDLE, J. F. AND FRANK, J. (2001). Unbiased confidence intervals for the odds ratio of two independent binomial samples with application to case-control data. *Biometrics* **57**, 484–489.

WOOLF, B. (1955). On estimating the relation between blood group and disease. *Annals of Human Genetics* **19**, 251–253.

[*Received April 24, 2001*; *first revision June 25, 2001*; *second revision August 22, 2001*; *accepted for publication August 27, 2001*]