STATISTICS IN MEDICINE
Statist. Med. **17**, 873–890 (1998)

INTERVAL ESTIMATION FOR THE DIFFERENCE BETWEEN INDEPENDENT PROPORTIONS: COMPARISON OF ELEVEN METHODS

ROBERT G. NEWCOMBE*

Senior Lecturer in Medical Statistics, University of Wales College of Medicine, Heath Park, Cardiff CF4 4XN, U.K.

SUMMARY

Several existing unconditional methods for setting confidence intervals for the difference between binomial proportions are evaluated. Computationally simpler methods are prone to a variety of aberrations and poor coverage properties. The closely interrelated methods of Mee and Miettinen and Nurminen perform well but require a computer program. Two new approaches which also avoid aberrations are developed and evaluated. A tail area profile likelihood based method produces the best coverage properties, but is difficult to calculate for large denominators. A method combining Wilson score intervals for the two proportions to be compared also performs well, and is readily implemented irrespective of sample size. © 1998 John Wiley & Sons, Ltd.

1. INTRODUCTION

Interval estimation for proportions and their differences encounters two problems that cannot arise in the continuous case: intervals that do not make sense, termed aberrations; and a coverage probability (achieved confidence level) that can be quite different to the intended nominal $1-\alpha$. Vollset[1] and Newcombe[2] are recent comparative evaluations of different available methods for the single proportion.

Similar issues apply to the difference between two proportions, a particularly important situation, arising naturally in prospective comparative studies such as the randomized controlled clinical trial. Unfortunately, standard statistical software including Minitab, SPSS and SAS, and even StatXact, has nothing to offer the user. Hence, by default, the computationally simplest asymptotic methods continue to be used, despite their known poor coverage characteristics and propensity to aberrations.

Table I shows the notation adopted for the comparison of two independent binomial proportions. It is assumed that the denominators m and n are fixed, leading to *unconditional* methods. Appendix I gives methods for the *ratio* of two proportions, which assume a different conditioning, namely $m+n$ fixed.

* Correspondence to: Robert G. Newcombe, Senior Lecturer in Medical Statistics, University of Wales College of Medicine, Heath Park, Cardiff CF4 4XN, U.K.

CCC 0277–6715/98/080873–18$17.50

Received May 1995
Revised July 1997

Table I. Notation for comparison of two independent proportions

Observed frequencies:

	Sample	
	1	2
+	a	b
−	c	d
Total	m	n

Theoretical proportions:	Observed proportions:
$\mathbf{E}A/m = \pi_1$	$p_1 = a/m$
$\mathbf{E}B/n = \pi_2$.	$p_2 = b/n$.

Here A and B denote random variables of which a and b are realizations.

Reparameterization:
Parameter of interest $\theta = \pi_1 - \pi_2$
Nuisance parameter $\psi = (\pi_1 + \pi_2)/2$.

Beal[3] reviewed and evaluated several asymptotic unconditional methods – see Section 2 for explicit formulae. All of these involve identifying the interval within which $(\theta - \hat{\theta})^2 \leqslant z^2 V(\tilde{\psi}, \tilde{\theta})$ where $\hat{\theta} = a/m - b/n$, and $V(\psi, \theta) = u\{4\psi(1-\psi) - \theta^2\} + 2v(1-2\psi)\theta = \pi_1(1-\pi_1)/m + \pi_2(1-\pi_2)/n$ is the variance of $\hat{\theta}$, $u = \frac{1}{4}(1/m + 1/n)$, $v = \frac{1}{4}(1/m - 1/n)$, and z is the standard Normal deviate associated with a two-tailed probability α. Here $\tilde{\psi}$ and $\tilde{\theta}$ denote hypothetical values of ψ and θ. The simple asymptotic method involves substitution of MLEs, $\tilde{\psi} = \hat{\psi}$ and $\tilde{\theta} = \hat{\theta}$. This may be improved upon in several ways using $\tilde{\theta} = \theta$ and solving for θ which is analogous to the method of Wilson[4] for the single proportion. A simple alternative class of estimators of this form involves setting $\tilde{\theta} = \theta$ and $\tilde{\psi}$ as a Bayes posterior estimate of ψ. Beal examined two resulting methods, termed the Haldane and Jeffreys–Perks methods; both performed generally better than the simple asymptotic method, and of the two, that of Jeffreys–Perks was preferable. These methods are, however, prone to certain novel anomalies, especially latent overshoot, as described later.

Beal also evaluated the closely interrelated methods of Mee[5] and Miettinen and Nurminen[6] which are based on, but superior to, Anbar.[7] Here $\tilde{\theta} = \theta$ and $\tilde{\psi} = \psi_\theta$, the profile estimate of ψ given θ, that is, the MLE of ψ conditional on the hypothesized value of θ. The form of ψ_θ is described in Appendix II, distinguishing the four cases NZ (no zero cells), OZ (one zero), RZ (two zeros in the same row) and DZ (two zeros on the same diagonal). The Miettinen–Nurminen method involves imputing to $\hat{\theta}$ a variance which is $(m+n)/(m+n-1)$ times as large as the expression for V above. Recently Wallenstein[8] published a closely related non-iterative method with similar coverage.

Miettinen and Nurminen also[6] considered a true profile likelihood method, involving $\{\theta: \ln \Lambda(\theta, \psi_\theta) - \ln \Lambda(\hat{\theta}, \hat{\psi}) \geqslant -z^2/2\}$, where Λ denotes the likelihood function. They concluded it was theoretically inferior, for reasons set out by Cox and Reid;[9] it is not amenable to continuity correction to mitigate its anti-conservatism.

An alternative approach involving precisely computed tail probabilities based on θ and ψ_θ was found greatly superior to existing methods for the paired difference case,[10] and is included in the present evaluation also. This approach leads naturally to consideration of a pair of methods.

A so-called 'exact' definition of tail probabilities aims to align the minimum coverage probability (CP) with $1-\alpha$. Alternatively, 'mid-p' tail areas[11–13] represent the attempt to achieve a mean coverage of $1-\alpha$.

The simple asymptotic method, without and with the continuity correction, and the methods described in the above four paragraphs constitute methods 1 to 9 of the 11 methods considered in the present evaluation; in general, there is a progressive improvement in performance from the simple asymptotic method 1 to method 9, at the cost of greatly increased computational complexity. Nevertheless, the tail area profile likelihood methods 8 and 9, which are the most complex of any evaluated here, but which as we will show have the best coverage and location properties, display a novel anomaly. Suppose a, m and the ratio $p_2=b/n$ are held constant, while $n\to\infty$ through values which keep b integer valued. We would expect that a good method for $a/m-b/n$ would produce a sequence of intervals, each nested within its predecessor, tending asymptotically towards some corresponding interval for the single proportion, shifted by the constant p_2. Yet these methods give a sequence of lower limits which increase up to a certain n, but subsequently decrease, violating the above consideration. Evaluation of just what the lower limit is converging towards is computationally prohibitive, but there is clearly an anomaly here.

Now, it is clear that the simple method's asymptotic behaviour is appropriate in this respect. The anomaly can only arise because the reparameterization (θ,ψ) leads to disregard of the fact that a/m and b/n are *independently* sufficient statistics for π_1 and π_2, respectively. A simple combination of single-sample intervals for a/m and b/n, which avoids the deficiencies of methods 1 and 2, is thus worth considering, indeed would seem to correspond most closely to the chosen conditioning on m and n only. We may combine Wilson[4] score intervals (without or with continuity correction) for each single proportion in much the same way as the simple method is constructed. The resulting methods avoid all aberrations.

2. METHODS COMPARED

Eleven unconditional methods were selected for comparison. Only methods 1 to 4 are capable of violating the $[-1,+1]$ boundaries, in which case the resulting interval is truncated. Methods 1 to 4 and 10 to 11 involve direct computation, methods 5 to 9 are iterative, of which methods 8 and 9 are the most complex. For notation see Table I, and Appendix II for the form of ψ_θ:

1. Simple asymptotic method, no continuity correction:

$$\hat{\theta}\pm z\surd(ac/m^3+bd/n^3).$$

2. Simple asymptotic method, with continuity correction (reference 14, p. 29):

$$\hat{\theta}\pm(z\surd\{ac/m^3+bd/n^3\}+(1/m+1/n)/2).$$

3. Beal's Haldane method:[3] limits are $\theta^*\pm w$ where

$$\theta^*=\frac{\hat{\theta}+z^2v(1-2\tilde{\psi})}{1+z^2u}$$

$$w=\frac{z}{1+z^2u}\surd[u\{4\tilde{\psi}(1-\tilde{\psi})-\hat{\theta}^2\}+2v(1-2\tilde{\psi})\hat{\theta}+4z^2u^2(1-\tilde{\psi})\tilde{\psi}+z^2v^2(1-2\tilde{\psi})^2]$$

$\tilde{\psi}=(a/m+b/n)/2$, $u=(1/m+1/n)/4$ and $v=(1/m-1/n)/4$.

4. Beal's Jeffreys–Perks method:[3] as above, but with

$$\tilde{\psi} = \frac{1}{2}\left(\frac{a+0{\cdot}5}{m+1} + \frac{b+0{\cdot}5}{n+1}\right).$$

5. Method of Mee:[5] the interval is

$$\left\{\theta\colon |\hat{\theta}-\theta| \leqslant z\sqrt{\left[\lambda\left\{\frac{(\psi_\theta+\theta/2)(1-\psi_\theta-\theta/2)}{m} + \frac{(\psi_\theta-\theta/2)(1-\psi_\theta+\theta/2)}{n}\right\}\right]}\right\}$$

where $\lambda = 1$.

6. Method of Miettinen and Nurminen (reference 6, equations 8 and 9 with Wilson-form variance): as method 5, but $\lambda = (m+n)/(m+n-1)$.
7. True profile likelihood method (reference 6, appendix III): the interval consists of all θ satisfying

$$a\ln\frac{\psi_\theta+\theta/2}{a/m} + b\ln\frac{\psi_\theta-\theta/2}{b/n} + c\ln\frac{1-\psi_\theta-\theta/2}{c/m} + d\ln\frac{1-\psi_\theta+\theta/2}{d/n} \geqslant -\frac{z^2}{2}$$

omitting any terms corresponding to empty cells.

8. Profile likelihood method based on 'exact' tail areas: interval is $L \leqslant \theta \leqslant U$ such that

(i) if $L \leqslant \theta \leqslant \hat{\theta}$, $kP_x + \sum_{1 \geqslant \xi > x} P_\xi \geqslant \frac{\alpha}{2}$

(ii) if $\hat{\theta} \leqslant \theta \leqslant U$, $kP_x + \sum_{-1 \leqslant \xi < x} P_\xi \geqslant \frac{\alpha}{2}$

where $P_\xi = \Pr[A/m - B/n = \xi \,|\, \theta, \psi_\theta]$, $x = a/m - b/n$ and $k = 1$.

9. Profile likelihood method based on 'mid-*p*' tail areas: as method 8, but with $k = 0{\cdot}5$.
10. Method based on the Wilson[4] score method for the single proportion, without continuity correction:

$L = \hat{\theta} - \delta$, $U = \hat{\theta} + \varepsilon$ where

$$\delta = \sqrt{\{(a/m - l_1)^2 + (u_2 - b/n)^2\}} = z\sqrt{\{l_1(1-l_1)/m + u_2(1-u_2)/n\}}$$
$$\varepsilon = \sqrt{\{(u_1 - a/m)^2 + (b/n - l_2)^2\}} = z\sqrt{\{u_1(1-u_1)/m + l_2(1-l_2)/n\}}$$

l_1 and u_1 are the roots of $|\pi_1 - a/m| = z\sqrt{\{\pi_1(1-\pi_1)/m\}}$, and l_2 and u_2 are the roots of $|\pi_2 - b/n| = z\sqrt{\{\pi_2(1-\pi_2)/n\}}$.

11. Method using continuity-corrected score intervals (reference 14, pp. 13–14): as method 10, but l_1 and u_1 delimit the interval

$$\{\pi_1\colon |\pi_1 - a/m| - 1/(2m) \leqslant z\sqrt{[\pi_1(1-\pi_1)/m]}\}.$$

Note that if $a = 0$, $l_1 = 0$; if $c = 0$, $u_1 = 1$. Similarly l_2 and u_2 delimit the interval

$$\{\pi_2\colon |\pi_2 - b/n| - 1/(2n) \leqslant z\sqrt{[\pi_2(1-\pi_2)/n]}\}.$$

Table II shows the eleven methods applied to eight selected combinations of a, m, b and n, representing all of cases NZ, OZ, RZ and DZ. Clearly when all four cell frequencies are large, as in example (a) (reference 14, p. 101), all methods produce rather similar intervals. Choice between

Table II. 95 per cent confidence intervals for selected contrasts, calculated using eleven methods. Asterisked values denote aberrations (limits beyond ± 1 or inappropriately equal to $\hat{\theta}$)

Method	Contrast							
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
	56/70–48/80	9/10–3/10	6/7–2/7	5/56–0/29	0/10–0/20	0/10–0/10	10/10–0/20	10/10–0/10
1. Asympt, no CC	+0·0575, +0·3425	+0·2605, +0·9395	+0·1481, +0·9947	+0·0146, +0·1640	0·0000*, 0·0000*	0·0000*, 0·0000*	+1·0000*, +1·0000	+1·0000*, +1·0000
2. Asympt, CC	+0·0441, +0·3559	+0·1605, > +1·0000*	+0·0053, > +1·0000*	−0·0116, +0·1901	−0·0750, +0·0750	−0·1000, +0·1000	+0·9250, > +1·0000*	+0·9000, > +1·0000*
3. Haldane	+0·0535, +0·3351	+0·1777, +0·8289	+0·0537, +0·8430	−0·0039, +0·1463	0·0000*, +0·0839	0·0000*, 0·0000*	+0·7482, +1·0000	+0·6777, +1·0000
4. Jeffreys–Perks	+0·0531, +0·3355	+0·1760, +0·8306	+0·0524, +0·8443	−0·0165, +0·1595	−0·0965, +0·1746	−0·1672, +0·1672	+0·7431, > +1·0000*	+0·6777, +1·0000
5. Mee	+0·0533, +0·3377	+0·1821, +0·8370	+0·0544, +0·8478	−0·0313, +0·1926	−0·1611, +0·2775	−0·2775, +0·2775	+0·7225, +1·0000	+0·6777, +1·0000
6. Miettinen–Nurminen	+0·0528, +0·3382	+0·1700, +0·8406	+0·0342, +0·8534	−0·0326, +0·1933	−0·1658, +0·2844	−0·2879, +0·2879	+0·7156, +1·0000	+0·6636, +1·0000
7. True profile	+0·0547, +0·3394	+0·2055, +0·8634	+0·0760, +0·8824	+0·0080, +0·1822	−0·0916, +0·1748	−0·1748, +0·1748	+0·8252, +1·0000	+0·8169, +1·0000
8. ‘Exact’	+0·0529, +0·3403	+0·1393, +0·8836	−0·0104, +0·9062	−0·0302, +0·1962	−0·1684, +0·3085	−0·3085, +0·3085	+0·6915, +1·0000	+0·6631, +1·0000
9. ‘Mid-*p*’	+0·0539, +0·3393	+0·1834, +0·8640	+0·0470, +0·8840	−0·0233, +0·1868	−0·1391, +0·2589	−0·2589, +0·2589	+0·7411, +1·0000	+0·7218, +1·0000
10. Score, no CC	+0·0524, +0·3339	+0·1705, +0·8090	+0·0582, +0·8062	−0·0381, +0·1926	−0·1611, +0·2775	−0·2775, +0·2775	+0·6791, +1·0000	+0·6075, +1·0000
11. Score, CC	+0·0428, +0·3422	+0·1013, +0·8387	−0·0290, +0·8423	−0·0667, +0·2037	−0·2005, +0·3445	−0·3445, +0·3445	+0·6014, +1·0000	+0·5128, +1·0000

CC: continuity correction

methods is more critical when the numbers are smaller, as in cases (b) to (h). The degree of concordance with hypothesis testing is limited, not surprisingly, as the conditioning is different: constrast (d) (Goodfield *et al.*,[15] cited by Altman and Stepniewska[16]) exemplifies the anti-conservatism of methods 1 and 7, whilst (c) suggests methods 8 and 11 are conservative. Overt overflow (see below) can occur with methods 2 and 4 and indeed also method 1. Limits inappropriately equal to $\hat{\theta}$ can occur with methods 1 and 3.

3. CRITERIA FOR EVALUATION

The present evaluation presupposes the principles set out in detail by Newcombe.[2] In brief, a good method will avoid all aberrations and produce an appropriate distribution of coverage probabilities. The coverage probability[2] CP is defined as $\Pr[L \leqslant \theta \leqslant U]$ where L and U are the calculated limits. The 'exact' criterion requires CP $\geqslant 1 - \alpha$ for all points in the parameter space, but by the smallest attainable margin. We interpret[2] the 'mid-*p*' criterion to imply that for any m and n the mean coverage probability $\overline{\mathrm{CP}}$ is to be close to but not below $1 - \alpha$, and $\min_{0<\theta<1}$ CP should not be too far below $1 - \alpha$, for chosen m and n or for any m and n. All methods evaluated aim to have $\alpha/2$ non-coverage at each end, except in boundary cases ($\hat{\theta} = \pm 1$). In recognition of the importance of interval location we examine symmetry as well as degree of coverage. When, as here, all methods are equivariant,[17] that is, show appropriate properties of symmetry about 0 when p_1 and p_2 are interchanged or replaced by $1 - p_1$ or $1 - p_2$, equality of left and right non-coverage as θ ranges from -1 to $+1$ is gratuitous. It is more pertinent to distinguish probabilities of non-coverage at the *mesial* (closer to 0) and *distal* (closer to ± 1) ends of the interval. Here, as with the single proportion, the relationship of coverage to the parameter value shows many discontinuities (see, for example, graphs in Vollset[1]). To minimize the potential for distortion as a result of this, θ and ψ are randomly sampled real numbers, not rationals, and similarly we avoid use of selected, round values of m and n.

The 'exactness' claimed for any method can only relate to its mode of derivation, and does not carry across to the achieved coverage probabilities for specific combinations of m, n, θ and ψ. It is necessary to evaluate these for a representative set of points in the parameter space. It is also important to examine the location of L and U in relation to each other, to boundaries that should not be violated (-1 and $+1$), and to boundaries that should be violable (that is, avoidance of inappropriate tethering to $\hat{\theta}$, as defined below).

For any method of setting a confidence interval $[L, U]$ for $\theta = \pi_1 - \pi_2$, with $L \leqslant \hat{\theta} \leqslant U$, two properties are considered desirable:

(i) Appropriate coverage and location: $L \leqslant \theta \leqslant U$ should occur with probability $1 - \alpha$ and $L > \theta$ and $U < \theta$ each with probability $\alpha/2$.
(ii) Avoidance of *aberrations*, defined as follows.

Several kinds of aberrations can arise, principally *point estimate tethering* and *overshoot*. Tethering occurs when one or both of the calculated limits L and U coincides with the point estimate $\hat{\theta}$. In the extreme case, where $\hat{\theta} = +1$, it is appropriate that $U = \hat{\theta}$, likewise that $L = \hat{\theta}$ when $\hat{\theta} = -1$. Otherwise this is an infringement of the principle that the CI should represent some 'margin of error' on both sides of $\hat{\theta}$, and is counted as adverse. Bilateral point estimate tethering, $L = \hat{\theta} = U$, constitutes a degenerate or *zero-width interval* (ZWI) and is always inappropriate.

Point estimate tethering can only occur in case RZ (two zeros in the same row) for methods 1 and 3, and in case DZ (two zeros on the same diagonal), exemplified in Table II, contrasts (e) to (h). The Haldane method produces appropriate, unilateral tethering in case DZ. In case RZ, it produces unilateral tethering if $m \neq n$, and a ZWI at $\hat{\theta} = 0$ if $m = n$, both of which are inappropriate. Method 1 produces a totally inappropriate ZWI at 0 in case RZ, and a ZWI at $+1$ (or -1) in case DZ, for which unilateral tethering would be appropriate. In case DZ, the Jeffreys–Perks method reduces to the Haldane method if $m = n$, otherwise produces overt overshoot.

Overt overshoot (OO) occurs when either calculated limit is outside $[-1, +1]$; $U = +1$ is not counted as aberrant when $\hat{\theta} = +1$, and correspondingly at -1. Methods 1 to 4 are liable to produce OO, in which case we truncate the resulting interval to be a subset of $[-1, +1]$; instances in which OO would otherwise occur are counted in the evaluation.

Furthermore, methods 3 and 4 substitute an estimate $\tilde{\psi}$ for ψ which is formed without reference to θ, and very often one or two of the implied parameters

$$\tilde{\pi}_{1L} = \tilde{\psi} + L/2, \quad \tilde{\pi}_{2L} = \tilde{\psi} - L/2, \quad \tilde{\pi}_{1U} = \tilde{\psi} + U/2, \quad \tilde{\pi}_{2U} = \tilde{\psi} - U/2$$

lie outside $[0, 1]$. This anomaly is termed *latent overshoot* (LO); overt overshoot always implies latent overshoot, but latent overshoot can occur in the absence of overt overshoot, when inherent bounds for θ are not violated but the bounding rhombus $\frac{1}{2}|\theta| \leqslant \psi \leqslant 1 - \frac{1}{2}|\theta|$ is. The formulae still work, and do not indicate anything peculiar has occurred. In this evaluation the frequency of occurrence of any LO (irrespective of whether involving one or two implied parameters, and of whether OO also occurs) is obtained for methods 3 and 4, using the chosen $\alpha = 0{\cdot}05$. Unlike overt overshoot, latent overshoot cannot effectively be eliminated by truncation; as well as affecting coverage, such truncation can produce inappropriate point estimate tethering, or even an interval that excludes $\hat{\theta}$. Latent overshoot and its sequelae, like the inappropriate asymptotic behaviour of methods 8 and 9, appears a consequence of losing the simplicity of using information concerning π_1 and π_2 separately.

4. EVALUATION OF THE ELEVEN METHODS

The main evaluation of coverage is based on a sample of 9200 parameter space points (PSPs) (m, n, ψ, θ), with m and n between 5 and 50 inclusive. This approach is adopted for reasons set out in a preceding article.[2] The computer-intensive part of the process is the setting up of 'tables' of intervals for each (m, n) pair for the iterative methods, especially 8 and 9. Therefore a subset of 230 out of the 2116 possible pairs was selected (Figure 1), comprising all 46 diagonal entries with $m = n$, together with 92 pairs (m, n) with $m \neq n$ and the corresponding reversed pairs (n, m) which require the same tables. For each of $m = 5, 6, \ldots, 50$, two values of n were chosen, avoiding diagonal elements, duplicates and mirror-image pairs. These were selected so that m and n should be uncorrelated and that distributions of $|m - n|$ and the highest common factor (HCF) of m and n should be very close to those for all 2070 off-diagonal points. The rationale for examining the HCF is that when tail areas are defined in terms of $a/m - b/n$, the difference between 'mid-p' and 'exact' limits could be great when $m = n$ but relatively small when m and n are coprime. The configuration shown, the result of an iterative search, has m and n uncorrelated (Pearson's $r = -0{\cdot}00006$), and Kolmogorov–Smirnov statistics for $|m - n|$ and HCF 0·011 and 0·012, respectively. Thus the chosen set of off-diagonal (m, n) combinations may be regarded as representative, in all important respects, of all 2070 possible ones; these are used together with a deliberate over-representation of diagonal pairs, in view of their commonness of occurrence, to give a set of (m, n) pairs which may be regarded as typical.

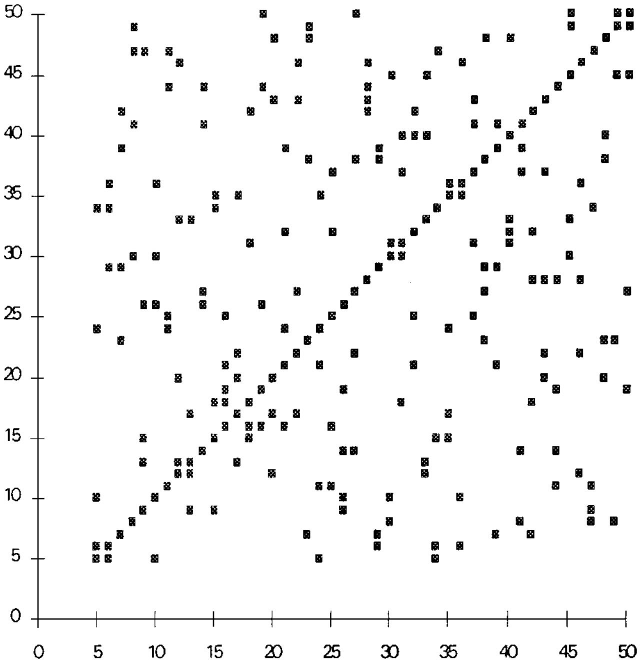

Figure 1. 230 (m, n) pairs chosen for the main evaluation

For each of the 230 (m, n) pairs, 40 (ψ, θ) pairs were chosen, with $\theta = \lambda\{1 - |2\psi - 1|\}$ and ψ and $\lambda \sim U(0, 1)$, all sampling being random and independent using algorithm AS183.[18] This resulted in a set of θ values with median 0·193, quartiles 0·070 and 0·393. For each chosen point (m, n, ψ, θ) of the parameter space, frequencies of all possible outcomes were determined as products of binomial probabilities. The achieved probability of coverage of θ by nominally 95 per cent confidence intervals calculated by each of the 11 methods was computed by summating all appropriate non-negligible terms. Mean coverage was also examined for subsets of the parameter space defined according to HCF, $\min(m, n)$, minimum expected frequency, ψ and θ in turn. Minimum coverage probabilities were also examined. Mesial and distal non-coverage rates were computed similarly, as were incidences of overt overshoot for methods 1 to 4, and latent overshoot at $1 - \alpha = 0{\cdot}95$ for methods 3 and 4. Probabilities of inappropriate tethering for methods 1 and 3 were computed by examining frequencies of occurrence of cases RZ and DZ, in conjunction with whether m and n were equal. Additionally, mean and minimum coverage probabilities for nominal 90 per cent and 99 per cent intervals for the same set of 9200 parameter space points were calculated.

To check further the effect of restricting to the chosen (m, n) pairs, coverage of 95 per cent intervals by methods 10 and 11 only was evaluated on a further set of 7544 parameter space points, 4 for each of the 1886 *unused* (m, n) pairs, with ψ and θ sampled as above.

A third evaluation examined the coverage of 95 per cent intervals by methods 10 and 11 only, when applied to the comparison of proportions with very large denominators but small to moderate numerators. 1000 parameter space points were chosen, with $\log_{10} m$ and

$\log_{10} n \sim U(2, 5)$, and $\log_{10}(4m\pi_1)$ and $\log_{10}(4n\pi_2) \sim U(0, 2)$, all sampling being independent and random.

In the first two evaluations, θ is positive, and left and right non-coverage are interpreted as mesial and distal non-coverage, respectively, with probabilities denoted here by MNCP and DNCP. Thus MNCP $= \sum_{\{a,b:\ l > \theta\}} p_{ab}$, and DNCP $= \sum_{\{a,b:\ u < \theta\}} p_{ab}$, where l and u are the calculated limits corresponding to observed numerators a and b, and $p_{ab} = \Pr[A = a, B = b \mid \theta, \psi]$. In the third evaluation, as in the intended application, θ may be of either sign, and mesial and distal non-coverage were imputed accordingly.

Expected interval width was calculated exactly for 95 per cent intervals by each method, truncated to lie within $[-1, +1]$ where necessary, for $\pi_1 = \pi_2 = 0{\cdot}5$ or $0{\cdot}01$ with m and n 10 or 100.

5. RESULTS

Table III shows that in the main evaluation the coverage probability of nominal 95 per cent intervals, averaged over the 9200 parameter space points, ranged from 0·881 (method 1) to 0·979 (method 11). In addition to method 1, method 3 was also anti-conservative on average, and method 7 slightly so; method 8 was slightly conservative, whereas other methods had appropriate mean coverage rates.

The *maximum* CP of method 1 in this evaluation was only 0·9656; for all other methods some parameter space points have CP = 1. The coverage probability of method 1 is arbitrarily close to 0 in extreme cases, either MNCP can approach 1 (when $\psi \sim 0$ or 1 and $\theta \sim 0$) or DNCP can (when $\psi \approx 0{\cdot}5$ and $\theta \sim 1$), due to ZWIs at 0 and 1, respectively. The continuity correction of method 2, though adequate to correct the mean CP, yields an unacceptably low min CP of 0·5137 in this evaluation. DNCP approaches $\exp(-0{\cdot}5) = 0{\cdot}6065$ with $n \ll m$ but $n \to \infty$, $\theta = \frac{1}{2}(1/m + 1/n) + \varepsilon$, $\pi_1 = 1 - \varepsilon$; a similar supremum applies to MNCP. The coverage of method 2 exhibited appropriate symmetry; for method 1, overall, distal non-coverage predominated in this evaluation.

Method 3, like method 1, can have CP arbitrarily close to 0, but only DNCP can approach 1, as $\psi \to 1$ (if $m \leqslant n$) or $\psi \to 0$ (if $m \geqslant n$), due to inappropriate tethering at 0. Method 4 eliminates this deficiency as well as the low mean CP, and reduces the preponderance of distal non-coverage.

Methods 5 and 6 have generally very similar coverage properties to each other, with overall coverage similar to method 4. The coverage probability was only 0·8516 when $m = 42$, $n = 7$, $\psi = 0{\cdot}9752$, $\theta = 0{\cdot}0253$, with MNCP = 0·1484, DCNP = 0; substantial distal non-coverage can also occur. These methods exhibited very good symmetry of coverage.

Coverage of method 7 was symmetrical but slightly anti-conservative on average, with min CP = 0·8299 ($m = 48$, $n = 23$, $\psi = 0{\cdot}9751$, $\theta = 0{\cdot}0806$, MNCP = 0·0344, DNCP = 0·1356). Values of either DNCP or MNCP around 0·14 can occur.

Even method 8 fails to be strictly conservative, the min CP in this evaluation being 0·9424, when $m = 32$, $n = 25$, $\psi = 0{\cdot}2640$ and $\theta = 0{\cdot}4016$. Here both MNCP (0·0279) and DNCP (0·0297) exceed the nominal $\alpha/2$; there are other parameter space points for which either exceeds 0·03. (This contrasts with the performance of the analogous method for the paired case[19] for which min CP was 0·9546, and DNCP (but not MNCP) was always less than 0·025.) The lowest coverage obtained for its 'mid-p' analogue, method 9, was 0·9131, at $m = n = 8$, $\psi = 0{\cdot}4890$, $\theta = 0{\cdot}4705$. Both these methods yielded symmetrical coverage.

For method 10, the lowest coverage obtained in the main evaluation was 0·8673, with $m = 35$, $n = 15$, $\psi = 0{\cdot}5087$, $\theta = 0{\cdot}9645$; this and other extremes arose entirely as distal non-coverage.

Table III. Estimated coverage probabilities for 95, 90 and 99 per cent confidence intervals calculated by 11 methods. Based on 9200 points in parameter space with $5 \leqslant m \leqslant 50$, $5 \leqslant n \leqslant 50$, $0 < \psi < 1$ and $0 < \theta < 1 - |2\psi - 1|$

Method	95% intervals						90% intervals		99% intervals	
	Coverage		Mesial non-coverage		Distal non-coverage		Coverage		Coverage	
	Mean	Minimum	Mean	Maximum	Mean	Maximum	Mean	Minimum	Mean	Minimum
1. Asympt, no CC	0·8807	0·0004	0·0417	0·7845	0·0775	0·9996	0·8322	0·0004	0·9253	0·0004
2. Asympt, CC	0·9623	0·5137	0·0183	0·4216	0·0194	0·4844	0·9401	0·5137	0·9811	0·5156
3. Haldane	0·9183	0·0035	0·0153	0·0656	0·0664	0·9965	0·8696	0·0035	0·9574	0·0035
4. Jeffreys–Perks	0·9561	0·8505	0·0140	0·0606	0·0299	0·1418	0·9123	0·7655	0·9896	0·9083
5. Mee	0·9562	0·8516	0·0207	0·1484	0·0231	0·1064	0·9076	0·8057	0·9919	0·9470
6. Miettinen–Nurminen	0·9584	0·8516	0·0196	0·1484	0·0220	0·1064	0·9114	0·8057	0·9925	0·9478
7. True profile	0·9454	0·8299	0·0268	0·1440	0·0278	0·1384	0·8912	0·6895	0·9893	0·9613
8. ‘Exact’	0·9680	0·9424	0·0149	0·0308	0·0170	0·0317	0·9305	0·8862	0·9948	0·9881
9. ‘Mid-*p*’	0·9591	0·9131	0·0197	0·04996	0·0212	0·0470	0·9116	0·8374	0·9933	0·9847
10. Score, no CC	0·9602	0·8673	0·0134	0·0660	0·0264	0·1327	0·9162	0·8226	0·9916	0·9173
11. Score, CC	0·9793	0·9339	0·0061	0·0271	0·0147	0·0661	0·9553	0·9012	0·9957	0·9399

CC: continuity correction

Table IV. Estimated coverage probabilities for 95 per cent confidence intervals calculated by 11 methods, for 9200 points in parameter space (PSPs). Determinants of coverage

Method	min (m, n)		min expected frequency		ψ in range		θ in range	
	5–9	30–50	0–1	5–25	<0·1 or >0·9	0·4–0·6	0–0·05	0·5–1
	Number of PSPs							
	1720	2360	3031	1875	1835	1907	1784	1422
1. Asympt, no CC	0·8063	0·9176	0·7981	0·9385	0·7416	0·9045	0·7604	0·8959
2. Asympt, CC	0·9470	0·9685	0·9616	0·9666	0·9665	0·9530	0·9791	0·9516
3. Haldane	0·9039	0·9276	0·8648	0·9480	0·8296	0·9474	0·8536	0·9437
4. Jeffreys–Perks	0·9617	0·9530	0·9649	0·9492	0·9793	0·9499	0·9782	0·9490
5. Mee	0·9602	0·9527	0·9679	0·9486	0·9732	0·9507	0·9675	0·9532
6. Miettinen–Nurminen	0·9632	0·9542	0·9697	0·9505	0·9744	0·9530	0·9692	0·9553
7. True profile	0·9476	0·9460	0·9561	0·9463	0·9566	0·9440	0·9553	0·9457
8. 'Exact'	0·9766	0·9628	0·9833	0·9551	0·9880	0·9612	0·9798	0·9670
9. 'Mid-p'	0·9674	0·9542	0·9745	0·9489	0·9803	0·9525	0·9734	0·9566
10. Score, no CC	0·9604	0·9584	0·9737	0·9486	0·9849	0·9468	0·9751	0·9497
11. Score, CC	0·9837	0·9750	0·9877	0·9682	0·9950	0·9698	0·9889	0·9715

CC: continuity correction

(The highest MNCP was 0·0660, but more extreme mesial non-coverage, up to 0·0998, occurred in the third evaluation with larger denominators (Table V)). MNCP can approach 0·1685, corresponding to the limiting non-coverage for the score method for the single proportion[2] when $\theta = L_1 - \varepsilon$, $\pi_2 = \varepsilon$, m small, $n \rightarrow \infty$, where L_1 is the lower score limit for $1/m$. Method 11 produced its lowest CP, 0·9339, when $m = n = 8$, $\psi = 0{\cdot}5160$, $\theta = 0{\cdot}9233$, entirely from distal non-coverage; this compares unfavourably with 0·949 for the corresponding method for the single proportion.[2] Both methods 10 and 11 yielded intervals erring towards mesial location.

Mean coverage was very similar whether m and n were equal, coprime or intermediate, except that for method 3, the mean CP was 0·9225 for $m \neq n$ but 0·9015 for $m = n$. Table IV illustrates the relation of mean coverage probability to other parameters: the lower of the two denominators m and n; the lowest of the four expected frequencies $m\pi_1$, $m(1 - \pi_1)$, $n\pi_2$ and $n(1 - \pi_2)$. Method 8, 11 and 6 (but not 5) were conservative on average in all zones of the parameter space examined. As expected, for most methods the mean coverage became closer to the nominal 0·95 as $\min(m, n)$ increased. Method 2 was anti-conservative for $\min(m, n)$ from 5 to 9, had mean CP 0·9628 for 10 to 19, which subsequently increased slightly. Method 7's mean coverage remained close to the overall mean of 0·9454. Method 11 remained very conservative for the larger values of m and n in the main evaluation.

The pattern according to minimum expected cell frequency (that is, $\min[m\pi_1, m(1 - \pi_1), n\pi_2, n(1 - \pi_2)]$ was generally similar, but more pronounced. Methods 2 and 7 were conservative on average even when the lowest expected frequency was under 1. For several methods whose overall mean coverage was over 0·95, the mean coverage dipped to slightly below 0·95 when all expected frequencies were over 5; this applies to methods 9, 10, 4 and 5, but not 6, apparently.

Table V. Estimated coverage probabilities for 95 per cent confidence intervals calculated by methods 10 and 11. Based on 1000 points in parameter space with $100 \leqslant m \leqslant 100{,}000$, $100 \leqslant n \leqslant 100{,}000$, $0 < \psi < 1$ and $0 < \theta < 1 - |2\psi - 1|$

	Coverage		Mesial non-coverage		Distal non-coverage	
	Mean	Minimum	Mean	Maximum	Mean	Maximum
10. Score, no CC	0·9618	0·9002	0·0265	0·0998	0·0116	0·0657
11. Score, CC	0·9785	0·9520	0·0153	0·0435	0·0062	0·0274

CC: continuity correction

Most methods had coverage closer to 0·95 for mesial ψ (0·4 to 0·6) than for distal ψ ($<0{\cdot}1$ or $>0{\cdot}9$). Method 10 was slightly anti-conservative for $0{\cdot}4 < \psi < 0{\cdot}6$, and also for $\theta > 0{\cdot}5$; method 4 was slightly anti-conservative for $0{\cdot}2 < \psi < 0{\cdot}8$, and for $\theta > 0{\cdot}25$. Conversely method 7 was conservative on average for ψ outwith (0·1, 0·9) and for $\theta < 0{\cdot}05$.

For method 1, the preponderance of distal non-coverage was gross for ψ outwith (0·1, 0·9) (MNCP = 0·0112, DNCP = 0·2472) or $\theta < 0{\cdot}05$. Conversely, for large θ, non-coverage was predominantly mesial ($0{\cdot}5 < \theta < 1$: MNCP = 0·0877, DNCP = 0·0164) – see examples (b) and (c) in Table II – with corresponding behaviour for mesial ψ. Thus when θ is large, method 1 does not err on the safe side, the interval fails to exclude values of θ which are considerably too large. For other methods, the mesial-distal location was less dependent upon θ and ψ.

Overt overshoot was common using methods 1 (mean probability 0·0270) and 2 (0·0594), less so for methods 3 (0·0012) and 4 (0·0052). It occurs with probability approaching 1 for method 2 as $\theta \rightarrow 1$, and also for method 4 provided $m \neq n$. For method 1, the maximum overshoot probability is lower, 0·7996, since as $\theta \rightarrow 1$, Pr[ZWI at 1] → 1. Latent overshoot occurred with probability around 0·5 for methods 3 and 4; some parameter space points produced latent overshoot with probability 1.

The Haldane method produced unilateral point estimate tethering with probability 0·0413 when $m \neq n$, and a ZWI at 0 with probability 0·0471 when $m = n$. Method 1 produced a ZWI at 0 in both these cases, and also a ZWI at 1 with probability 0·0035. For both methods the probability of inappropriate tethering approaches 1 as $\psi \rightarrow 0$ or 1.

Table III also shows coverage properties of 90 per cent and 99 per cent intervals, which where generally in line with the findings for 95 per cent intervals. Method 11 was strictly conservative for the chosen parameter space points at 90 per cent, but can be very anti-conservative at 99 per cent. The average coverage of method 4 became anti-conservative at 99 per cent.

In the second evaluation, using (m, n) pairs not included in the first, the coverage properties of methods 10 and 11 were virtually identical to those in the main evaluation.

In the third evaluation (Table V), with denominators ranging from 100 to 100,000, the mean coverage was very similar to the main evaluation, but the minimum CPs were more favourable, and contrastingly, the location of the interval was too *distal*. The lowest CP for method 10, 0·9002, resulted entirely from mesial non-coverage, with $m = 140$, $n = 67622$, $m\pi_1 = 0{\cdot}5349$, $n\pi_2 = 8{\cdot}9480$; the highest DNCP was only 0·0657. In the main evaluation, the location of these intervals was too mesial, but this asymmetry disappeared in the zone $\theta < 0{\cdot}05$.

Variation in expected interval width (Table VI) between different methods is most marked when any of the expected frequencies $m\pi_1$, $m(1 - \pi_1)$, $n\pi_2$ or $n(1 - \pi_2)$ is low. The width is then least for method 1 or 3, largely on account of the high frequency of degeneracy.

Table VI. Average interval width for 95 per cent confidence intervals calculated by 11 methods, for selected parameter space points

				m					
Method	10	10	10	100	100	100	100	100	100
				n					
	10	10	10	10	10	10	100	100	100
				π_1					
	0·01	0·5	0·95	0·01	0·5	0·95	0·01	0·5	0·95
				π_2					
	0·01	0·5	0·05	0·01	0·5	0·05	0·01	0·5	0·05
1. Asympt, no CC	0·0702	0·8302	0·2407	0·0635	0·6177	0·1996	0·0493	0·2758	0·1188
2. Asympt, CC	0·2702	1·0296	0·3420	0·1735	0·7277	0·2618	0·0693	0·2958	0·1385
3. Haldane	0·0646	0·7640	0·4316	0·1819	0·5904	0·2906	0·0487	0·2732	0·1227
4. Jeffreys–Perks	0·3580	0·7679	0·4327	0·2624	0·5930	0·3246	0·0644	0·2732	0·1227
5. Mee	0·5634	0·7737	0·4224	0·3286	0·5529	0·3507	0·0888	0·2732	0·1225
6. Miettinen–Nurminen	0·5840	0·7910	0·4371	0·3307	0·5549	0·3526	0·0891	0·2739	0·1228
7. True profile	0·3748	0·7990	0·3440	0·2233	0·5794	0·2871	0·0664	0·2745	0·1194
8. 'Exact'	0·6298	0·8801	0·4661	0·3372	0·5885	0·3541	0·0919	0·2840	0·1296
9. 'Mid-*p*'	0·5324	0·8128	0·4075	0·2978	0·5803	0·3384	0·0800	0·2749	0·1214
10. Score, no CC	0·5627	0·7231	0·4773	0·3289	0·5430	0·3522	0·0895	0·2707	0·1264
11. Score, CC	0·6945	0·8232	0·5744	0·4036	0·6121	0·4213	0·1061	0·2843	0·1398

CC: continuity correction

6. DISCUSSION

As for the single proportion, the only virtue of the simple asymptotic method is simplicity; the alignment of its coverage with $1-\alpha$ is merely nominal, an attribute of its method of construction, quite divorced from its attained coverage properties. Overt overshoot and inappropriate tethering are common, and coverage not symmetric.

Method 2 is a great improvement in terms of degree and symmetry of coverage, but the minimum coverage remains poor, and the improved mean coverage is attained at the cost of a still higher overshoot rate.

The Haldane and Jeffreys–Perks methods (3 and 4) attempt to overcome these deficiencies while maintaining closed-form tractability. The Haldane method suffers from poor mean and minimum coverage rates, severe asymmetry and proneness to aberrations. The Jeffreys–Perks method is a clear improvement, but still inadequate.

Methods 5 to 11 cannot produce any of the aberrations listed in Table V. The Mee and Miettinen–Nurminen methods (5 and 6) show slightly conservative mean coverage probability. Method 7, described[6] as theoretically and empirically somewhat inferior to method 6, is slightly anti-conservative; all three have a rather poor minimum coverage probability, however.

Methods 8 and 9 achieve much the performance they were designed to achieve, though method 8 as well as method 9 has a minimum coverage probability less than the nominal value. They require lengthy source code in any case, and also large amounts of computation time except when m and n are small.

Methods 10 and 11 are remarkable. They are computationally simpler than any of methods 5 to 9, and tractable for very large m and n. At least for 95 per cent and 90 per cent intervals, they attain a coverage distribution inferior only to their highly complex tail-area profile based counterparts, methods 9 and 8, respectively; the coverage distribution for method 10 is broadly similar to that of the fairly complex methods 5 and 6. (Method 10 is strictly less effective than method 9, in that its mean CP is higher but min CP lower; similarly for method 11 in comparison to method 8.) They avoid all aberrations, including the anomalous asymptotic behaviour of methods 8 and 9 as described in Section 1.

In Newcombe,[2] the Wilson method for the single proportion produced an interval that was located too mesially, with a distal non-coverage rate approximately 0·015 greater than the mesial. This is propagated into a similar difference, in both direction and size, between left and right non-coverage rates for method 10 in the main evaluation here. The same applies to the corresponding continuity-corrected methods, with a distal-mesial non-coverage preponderance approximately 0·01. The simple asymptotic method for the single proportion yielded an interval that was too distal, corresponding to the propensity to overshoot. Nevertheless, in the present evaluation method 1 intervals were located too mesially; the set of parameter space points was weighted away from high θ, for which location is too distal, ZWIs were common at $\hat{\theta} = 0$ but rare at $\hat{\theta} = 1$. Gart and Nam[20] found that a skewness correction, after Bartlett,[21] led to little improvement upon the Mee method.[5]

7. CONCLUSION

For the difference between independent proportions, a novel pair of methods (10 and 11) are presented, which are computationally very tractable irrespective of *m* and *n*, free from aberrations, and achieve better coverage properties than any except the most complex methods. In the absence of off-the-shelf software, these methods are strongly recommended over methods 1 and 2, the only ones commonly in use. Nevertheless, software producers are strongly urged to provide readily available routines for appropriate methods for the unpaired difference case, as also for the paired difference case and indeed the single proportion case.

APPENDIX I: MULTIPLICATIVE SCALE SYMMETRY OF CONDITIONAL INTERVALS DERIVED FROM THE WILSON SCORE INTERVAL

In a preceding paper[2] we showed that the Wilson score interval[4] is symmetrical on a logit scale. Interval estimates for certain types of *ratio* may be derived from intervals for *proportions*. These methods are *conditional* in nature in that the appropriate denominator *n* relates to only a subset of the *N* individuals studied, and has itself arisen as a result of sampling. From an appropriate interval estimate about $p = r/n$ we may derive a corresponding interval for any monotonic function of it, in particular the *odds* $r/(n - r)$. In the following examples, when the Wilson interval is used for r/n, the resulting derived interval is symmetrical on the multiplicative scale.

The simplest case[22–24] is the ratio of two Poisson counts, r_1/r_2. Armitage and Berry[23] considered the ratio of two bacterial colony counts of 13 and 31. A Wilson 95 per cent interval for the proportion 13/44 would be 0·1816 to 0·4422. This corresponds to an interval $0{\cdot}1816/(1 - 0{\cdot}1816) = 0{\cdot}2218$ to $0{\cdot}4422/(1 - 0{\cdot}4422) = 0{\cdot}7928$ for r_1/r_2, which is symmetrical about the point estimate 0·4194 on a multiplicative scale.

Furthermore, interest sometimes centres on the *ratio* of two proportions derived from independent samples, p_1/p_2 where $p_i = r_i/n_i$, $i = 1, 2$. Analogously, two rates expressed per person-year at risk may be contrasted by their ratio; here the odds ratio is inapplicable as no meaning attaches to $1-p_i$, $i=1,2$. Some existing methods are summarized by Miettinen and Nurminen[6] and Rothman.[25]

A simple method, appropriate for use when both proportions or rates are very small, may likewise be derived from an interval estimate for a single proportion. For example, 41 cases of breast carcinoma developed in a series[25,26] exposed to radiation, with 28,010 woman-years at risk, 1·856 times the rate in controls which was 15/19,017. We may base the calculation on any suitable interval for the proportion (41/56) of cases that arose in the exposed group. A Wilson interval for this proportion would be 0·604 to 0·830. The corresponding limits for the odds (41:15) are 1·526 and 4·897, which are then divided by the ratio of the denominators to yield limits 1·036 and 3·325 for the rate ratio. The method may be refined to standardize for age or other confounders, by adjusting the ratio of the denominators.[27]

Similarly, in an individually matched comparative retrospective study, the odds ratio ω is estimated by f/g, the ratio of the counts of discordant pairs. If (L, U) is a confidence interval for $f/(f+g)$, then $L/(1-L)$ and $U/(1-U)$ serve[28] as corresponding limits for ω. For example, Mills *et al.*[29] studied smoking habits in 106 psoriatics and individually matched controls, and obtained $f=38$, $g=14$, $\hat{\omega}=2{\cdot}71$. A score-based 95 per cent confidence interval for this odds ratio would be 1·48 to 4·96.

Coverage properties of such derived intervals follow from those of the interval method used for r/n. The Clopper–Pearson method[30] is often used,[28] and its conservatism carries across to conditional intervals derived from it. While choice of coverage intention is more important than symmetry, the latter is nonetheless a noteworthy property.

APPENDIX II: PROFILING OF ψ, THE NUISANCE PARAMETER, AS ψ_θ

When applied directly via the likelihood function $\Lambda(\theta,\psi)=\Pr[A=a \;\&\; B=b\,|\,\theta,\psi]$, by obtaining

$$\{\theta : \ln\Lambda(\theta,\psi_\theta)-\ln\Lambda(\hat{\theta},\hat{\psi})\geqslant -z^2/2\}$$

the profile likelihood approach has attracted censure as being in general anti-conservative.[9] Nevertheless the Mee and Miettinen–Nurminen methods for the unpaired difference, selected as optimal by Beal,[3] and the optimal methods for the paired difference[10,19] all involve substitution of the appropriate profile estimate for the nuisance parameter.

For the unpaired difference, with reparameterization as in Table I, the log-likelihood reduces (within an additive constant) to

$$\ln\Lambda = a\ln(\psi+\theta/2)+b\ln(\psi-\theta/2)+\mathrm{c}\ln(1-\psi-\theta/2)+d\ln(1-\psi+\theta/2)$$

with the understanding that terms corresponding to empty cells are omitted. The constraints $0\leqslant\pi_i\leqslant 1, i=1,2$ translate into restricting evaluation of $\ln\Lambda$ only within the bounding rhombus $\frac{1}{2}|\theta|\leqslant\psi\leqslant 1-\frac{1}{2}|\theta|$.

The four diagonal boundaries of this are: $\psi=-\theta/2$; $\theta/2$; $1-\theta/2$; and $1+\theta/2$. The likelihood is zero on each of these boundaries, unless the corresponding cell entry (a, b, c or d, respectively) is zero.

We distinguish four situations, according to the pattern of zero cells:

NZ: no zero cells; all four edges precipitous.

OZ: one cell zero, for example, $c = 0$, $abd > 0$; upper right edge $\psi = 1 - \theta/2$ is now available.

RZ: two cells in same row zero, for example $a = b = 0$, $cd > 0$; lower edges $\psi = \pm\,\theta/2$ are now available.

DZ: two cells on same diagonal zero, for example, $b = c = 0$, $ad > 0$; right-hand edges $\psi = \theta/2$ and $1 - \theta/2$ are now available.

These represent all situations allowed, as m and n must be greater than zero.

Case NZ: $abcd > 0$.

The log-likelihood function, regarded as a function of either θ or ψ, is a sum of terms of the form $\ln(\lambda\xi + \mu)$, where ξ can represent either θ or ψ, and $\lambda \neq 0$, whence

$$\frac{\partial^2}{\partial \xi^2} \ln(\lambda\xi + \mu) = \frac{-\lambda^2}{(\lambda\xi + \mu)^2} < 0.$$

So both $\partial^2/\partial\psi^2 \ln\Lambda$ and $\partial^2/\partial\theta^2 \ln\Lambda$ are negative throughout the bounding rhombus; the log-likelihood surface is smoothly convex, with chasms to $-\infty$ at all four edges in all cases. Thus there is a unique maximizing ψ_θ for each θ with $-1 < \theta < 1$, and $100(1 - \alpha)$ per cent profile likelihood limits for θ always exist, away from the boundary, for any $\alpha > 0$. $\partial/\partial\psi \ln\Lambda$ takes the form

$$\frac{C(\psi)}{\prod(\psi \pm \theta/2)(1 - \psi \pm \theta/2)}$$

where $C(\psi)$ is a cubic in ψ, but maximization is more readily performed iteratively. The MLE point is $\hat{\theta} = a/m - b/n$, $\hat{\psi} = (a/m + b/n)/2$; at $\theta = 0$, ψ_θ reduces to $\psi_0 = (a + b)/(m + n)$, the weighted proportion; as $\theta \to +1$ or -1, $\psi_\theta \to \frac{1}{2}$.

Case OZ: $c = 0$, $abd > 0$ say. $\hat{\pi}_1 = 1 > \hat{\pi}_2$

The attainable region here is the interior of the rhombus plus the edge $\psi = 1 - \theta/2, 0 < \theta < 1$. On this edge,

$$\frac{\partial}{\partial\psi} \ln\Lambda = a + \frac{b}{1 - \theta} - \frac{d}{\theta}$$

which tends to $-\infty$ as $\theta \to 0$ and to $+\infty$ as $\theta \to 1$. It has a unique zero at

$$\theta = \theta^* = \frac{a + n - \sqrt{\{(a + n)^2 - 4ad\}}}{2a}$$

where $0 < \theta^* < 1$.

For $\theta^* \leqslant \theta < 1$, ψ_θ is simply $1 - \theta/2$. For $-1 < \theta < \theta^*$, ψ_θ is within the rhombus and is $(B + \sqrt{\{B^2 - 4AC\}})/(2A)$ where $A = a + n$, $B = a(1 + \theta) + b$, $C = \{(a - b)(1 + \theta/2) - \mathrm{d}\theta/2\}\theta/2$. The MLE point is $\hat{\theta} = d/n$ where $\theta^* < \hat{\theta} < 1$, $\hat{\psi} = 1 - \hat{\theta}/2$, on the permitted edge.

Case RZ: $a = b = 0$, $cd > 0$ say. $\hat{\pi}_1 = \hat{\pi}_2 = \hat{\theta} = \hat{\psi} = 0$

This is the most tractable case for methods involving ψ_θ, though conversely the one that would invalidate the conditional approach for interval estimation of θ.

$$\frac{\partial}{\partial\psi} \ln\Lambda \leqslant 0 \text{ for all } \theta \text{ with } \tfrac{1}{2}|\psi| < \theta < 1 - \tfrac{1}{2}|\psi|$$

so the likelihood is maximized by $\psi_\theta = \frac{1}{2}|\theta|$, that is, on the two lower boundaries of the rhombus.

Thus for $\theta > 0$, $\psi = \psi_\theta$ implies $\pi_1 = \theta$, $\pi_2 = 0$; for $\theta < 0$, $\pi_1 = 0$, $\pi_2 = -\theta$.

Standard 95 per cent intervals for methods 7, 8 and 9 are then $(-1 + \gamma^{1/n}, 1 - \gamma^{1/m})$ where $\gamma = 0{\cdot}1465$, $0{\cdot}025$ and $0{\cdot}05$, respectively. These limits correspond to non-zero ones for the single proportions $0/n$ and $0/m$ (Miettinen and Nurminen,[6] equation 5, Clopper and Pearson[30] and Miettinen[31]). Methods 5 and 10 produce the limits $(-z^2/(n + z^2), z^2/(m + z^2))$, which correspond to Wilson[4] limits for $0/n$ and $0/m$; similarly for the analogous continuity-corrected methods.

Case DZ: $a = m$, $b = 0$, say. $\hat{\theta} = \hat{\pi}_1 - \hat{\pi}_2 = 1 - 0 = 1$

In this situation, a good method is expected to produce unilateral tethering. The right-hand boundaries of the rhombus, $\psi = \theta/2$ and $\psi = 1 - \theta/2$, are now attainable. The form of ψ_θ here depends on the relative sizes of m and n:

$$\text{If } m = n,\ \psi_\theta = \tfrac{1}{2} \text{ for all } \theta \text{ with } -1 < \theta < 1.$$

Otherwise, a pair of intersecting straight lines:

$$\text{If } m > n,\ \psi_\theta = 1 - \theta/2 \text{ when } n/m \leqslant \theta \leqslant 1,$$

$$\psi_\theta = \frac{m + (m - n)\theta/2}{m + n} \text{ when } -1 \leqslant \theta \leqslant n/m.$$

$$\text{If } m < n,\ \psi_\theta = \theta/2 \text{ when } m/n \leqslant \theta \leqslant 1,$$

$$\psi_\theta = \frac{m + (m - n)\theta/2}{m + n} \text{ when } -1 \leqslant \theta \leqslant m/n.$$

ACKNOWLEDGEMENTS

I am grateful to Dr. E. C. Coles for helpful discussion, to Mr. C. M. Berners-Lee for assistance with programming, and to two anonymous referees for many helpful suggestions.

REFERENCES

1. Vollset, S. E. 'Confidence intervals for a binomial proportion', *Statistics in Medicine*, **12**, 809–824 (1993).
2. Newcombe, R. G. 'Two-sided confidence intervals for the single proportion: comparison of seven methods', *Statistics in Medicine*, **17**, 857–872 (1998).
3. Beal, S. L. 'Asymptotic confidence intervals for the difference between two binomial parameters for use with small samples', *Biometrics*, **43**, 941–950 (1987).
4. Wilson, E. B. 'Probable inference, the law of succession, and statistical inference', *Journal of the American Statistical Association*, **22**, 209–212 (1927).
5. Mee, R. W. 'Confidence bounds for the difference between two probabilities', *Biometrics*, **40**, 1175–1176 (1984).

6. Miettinen, O. S. and Nurminen, M. 'Comparative analysis of two rates', *Statistics in Medicine*, **4**, 213–226 (1985).
7. Anbar, D. 'On estimating the difference between two probabilities with special reference to clinical trials', *Biometrics*, **39**, 257–262 (1983).
8. Wallenstein, S. 'A non-iterative accurate asymptotic confidence interval for the difference between two proportions', *Statistics in Medicine*, **16**, 1329–1336 (1997).
9. Cox, D. R. and Reid, N. 'Parameter orthogonality and approximate conditional inference', *Journal of the Royal Statistical Society, Series B*, **49**, 1–39 (1992).
10. Newcombe, R. G. 'Unconditional confidence interval methods for the difference between two binomial proportions based on paired data', 16th International Biometric Conference, Hamilton, New Zealand, 1992.
11. Lancaster, H. O. 'The combination of probabilities arising from data in discrete distributions', *Biometrika*, **36**, 370–382 (1949).
12. Stone, M. 'The role of significance testing. Some data with a message', *Biometrika*, **56**, 485–493 (1969).
13. Berry, G. and Armitage, P. 'Mid-P confidence intervals: a brief review', *Statistician*, **44**, 417–423 (1995).
14. Fleiss, J. L. *Statistical Methods for Rates and Proportions*, 2nd edn, Wiley, New York, 1981.
15. Goodfield, M. J. D., Andrew, L. and Evans, E. G. V. 'Short-term treatment of dermatophyte onychomycosis with terbinafine', *British Medical Journal*, **304**, 1151–1154 (1992).
16. Altman, D. G. and Stepniewska, K. A. 'Confidence intervals for small proportions, including zero', *British Medical Journal*, in submission.
17. Blyth, C. R. and Still, H. A. 'Binomial confidence intervals', *Journal of the American Statistical Association*, **78**, 108–116 (1983).
18. Wichmann, B. A. and Hill, I. D. 'An efficient and portable pseudo-random number generator', *in* Griffiths, P. and Hill, I. D. (eds), *Applied Statistics Algorithms*, Ellis Horwood, Chichester, 1985.
19. Newcombe, R. G. 'Improved confidence interval methods for the difference between binomial proportions based on paired data', Submitted for publication.
20. Gart, J. J. and Nam, J-M. 'Approximate interval estimation of the difference in binomial parameters: correction for skewness and extension to multiple tables', *Biometrics*, **46**, 637–643 (1990).
21. Bartlett, M. S. 'Approximate confidence intervals. III. A bias correction', *Biometrika*, **42**, 201–204 (1955).
22. Kahn, H. A. 'The relationship of reported coronary heart disease mortality to physical activity of work', *American Journal of Public Health*, **53**, 1058–1067 (1963).
23. Armitage, P. and Berry, G. *Statistical Methods in Medical Research*, 2nd edn, Wiley, New York, 1987.
24. Ederer, F. and Mantel, N. 'Confidence limits on the ratio of two Poisson variables', *American Journal of Epidemiology*, **100**, 165–167 (1974).
25. Rothman, K. *Modern Epidemiology*, Little Brown, Boston, 1986.
26. Boice, J. A. and Monson, R. R. 'Breast cancer in women after repeated fluoroscopic examination of the chest', *Journal of the National Cancer Institute*, **59**, 823–832 (1977).
27. Gardner, M. J. and Altman, D. G. (eds). *Statistics with Confidence. Confidence Intervals and Statistical Guidelines*, British Medical Journal, London, 1989.
28. Breslow, N. E. and Day, N. E. *Statistical Methods in Cancer Research*. 1. *The Analysis of Case-Control Studies*, IARC, Lyon, 1980.
29. Mills, C. M., Srivastava, E. D., Harvey, I. M., Swift, G. L., Newcombe, R. G., Holt, P. J. A. and Rhodes, J. 'Smoking habits in psoriasis: a case-control study', *British Journal of Dermatology*, **127**, 18–21 (1992).
30. Clopper, C. J. and Pearson, E. S. 'The use of confidence or fiducial limits illustrated in the case of the binomial', *Biometrika*, **26**, 404–413 (1934).
31. Miettinen, O. S. *Theoretical Epidemiology*, Wiley, New York, 1985, pp. 120–121.