STATISTICS IN MEDICINE
Statist. Med. **17**, 2635–2650 (1998)

IMPROVED CONFIDENCE INTERVALS FOR THE DIFFERENCE BETWEEN BINOMIAL PROPORTIONS BASED ON PAIRED DATA

ROBERT G. NEWCOMBE*

Department of Medical Computing and Statistics, University of Wales College of Medicine, Heath Park, Cardiff, CF4 4XN, U.K.

SUMMARY

Existing methods for setting confidence intervals for the difference θ between binomial proportions based on paired data perform inadequately. The asymptotic method can produce limits outside the range of validity. The 'exact' conditional method can yield an interval which is effectively only one-sided. Both these methods also have poor coverage properties. Better methods are described, based on the profile likelihood obtained by conditionally maximizing the proportion of discordant pairs. A refinement (methods 5 and 6) which aligns $1 - \alpha$ with an aggregate of tail areas produces appropriate coverage properties. A computationally simpler method based on the score interval for the single proportion also performs well (method 10). © 1998 John Wiley & Sons, Ltd.

1. INTRODUCTION

Interval estimation for proportions encounters two characteristic problems. First, the intended coverage probability cannot be achieved exactly, because of discreteness. Secondly, in many instances the familiar simple formulae fail to yield sensible intervals in the contexts described in Sections 2 and 6. Vollset[1] and Newcombe[2] provide recent comparative evaluations of different methods of interval evaluation for the simplest case, the single proportion. For the contrast between two proportions based on individually paired data, previous methods are highly unsatisfactory; acceptable alternatives are developed in this paper.

Table I summarizes notation, including several parameters pertinent to the different methods considered. The null hypothesis in the McNemar[3] test may be expressed as a zero difference, $\theta = \pi_2 - \pi_3 = 0$. In a matched comparative retrospective study, association between outcome and exposure is expressed by the odds ratio $\omega \equiv \pi_2/\pi_3$, the null hypothesis then being $\omega = 1$. In both situations, $\psi = \pi_2 + \pi_3$ plays an important role as a nuisance parameter. Contexts in which the difference θ is useful include the following.

(i) Method comparison studies for a binary attribute with assessment of bias of one method relative to the other. Thus Hope *et al.*[4] evaluated a new faecal occult blood test together with two existing ones on the same subjects, and gave confidence intervals for differences in

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

CCC 0277–6715/98/222635–16$17.50

Received May 1995
Accepted February 1998

Table I. Notation for comparison of proportions from paired binary data

First classification	Second classification	Observed frequencies	Observed proportions	Theoretical proportions
+	+	e	$p_1 = e/n$	π_1
+	−	f	$p_2 = f/n$	π_2
−	+	g	$p_3 = g/n$	π_3
−	−	h	$p_4 = h/n$	π_4
Total		n	1	1

Unconditional model:		
Parameter of interest	$\theta = (\pi_1 + \pi_2) - (\pi_1 + \pi_3) = \pi_2 - \pi_3$	(1)
Nuisance parameter	$\psi = \pi_2 + \pi_3$	(2)
Conditional model:		
Conditional probability	$\mu = \pi_2/(\pi_2 + \pi_3)$	(3)
Odds ratio for paired data	$\omega = \pi_2/\pi_3$	(4)
Additional parameters:		
Phi coefficient	$\phi = \dfrac{\pi_1\pi_4 - \pi_2\pi_3}{\sqrt{\{(\pi_1 + \pi_2)(\pi_3 + \pi_4)(\pi_1 + \pi_3)(\pi_2 + \pi_4)\}}}$	(5)
Inverse tetrachoric odds-ratio	$\eta = \pi_2\pi_3/\pi_1\pi_4$	(6)
Diagonal split parameters	$\nu = \pi_1/(\pi_1 + \pi_4)$ μ as above	(7)

sensitivity and specificity, using method 10 below. Simultaneous comparison of sensitivity and specificity is clearly desirable.[5]

(ii) Prospective studies in which subjects are assessed on two occasions. These may involve natural progression over time or some therapeutic intervention. In the latter case, the study may be poorly controlled (as in the examples cited in references 6 and 7), or may be a two-period cross-over trial. Here, if numbers of subjects in the two treatment order groups are equal, the analysis may be simplified by disregarding the order in which the treatments were administered. The methods evaluated here are then applicable.

In these situations θ is estimated by $\hat{\theta} = (f - g)/n$, the maximum likelihood estimate, as well as the obvious empirical one.

In Section 2 we introduce existing interval estimation methods for θ and show how they incur problems. In Section 3 we develop methods involving substitution of a profile estimate for the nuisance parameter ψ which obviate these problems. In Section 4 we introduce computationally simpler, effective methods. Section 5 presents formulae for the ten methods evaluated, and some numerical examples. The principles, plan and results of the evaluation are set out in Sections 6, 7 and 8, respectively.

2. THE PROBLEM

Testing the null hypothesis $\pi_1 + \pi_2 = \pi_1 + \pi_3$ is performed *conditional* on the observed split into $f + g$ discordant (informative) pairs and $e + h$ concordant (uninformative) ones; either an

asymptotic or an 'exact' test of H_0: $\mu = 1/2$ is performed. Accordingly, a method for the single proportion may be used to derive confidence limits L and U for the conditional probability μ, which are then transformed into corresponding limits for the parameter of interest. Thus for the odds ratio $\omega = \pi_2/\pi_3 = \mu/(1-\mu)$, $L/(1-L)$ and $U/(1-U)$ are appropriate limits;[8] coverage properties for ω will correspond to those for the method chosen for the single proportion.

However, suppose that the parameter of interest is the difference between proportions, θ. Corresponding limits for this parameter may be obtained, conditional of $f+g$, by substituting $\hat{\psi} = (f+g)/n$. We show that even when optimal methods are used to obtain L and U, limits for θ of the form $(2L-1)\hat{\psi}$ and $(2U-1)\hat{\psi}$ tend to perform unsatisfactorily.

Thus, the 'exact' conditional method[9,10] incurs an anomaly which we call *point estimate tethering* if $f = 0$ or $g = 0$. When $g = 0$, the Clopper–Pearson[11] interval for μ is one-sided, $[L, 1]$, the upper limit coinciding with $\hat{\mu}$, the point estimate of μ, which is appropriate in a boundary case. However, the corresponding upper limit for θ is now $\hat{\psi} = f/n$, identical to $\hat{\theta}$. Thus the method fails to produce a two-sided interval, even though values of $\pi_2 - \pi_3$ between f/n and one are not ruled out by the data – an undesirable consequence of the conditioning on $\psi = \hat{\psi}$. Point estimate tethering can occur with the unpaired difference also,[12] but only at $\hat{\theta} = 0$ or ± 1; here it can occur at any valid θ.

Conversely, the asymptotic variance interval $\hat{\theta} \pm z\sqrt{\{((e+h)(f+g)+4fg)/n^3\}}$ (where z denotes the $1-\alpha/2$ point of the standard Normal distribution) does not do this, but can violate the $[-1, +1]$ bounds on θ; for some combinations of n, ψ and θ, it usually does so.

Consider an extreme example in which both methods are unsatisfactory: $e+h = 2$, $f = 98$, $g = 0$, and $\hat{\theta} = 0.98$. The asymptotic 95 per cent interval without continuity correction, calculated directly, is 0·953 to 1·007. The 'exact' conditional 95 per cent interval, based on a Clopper–Pearson interval for μ, is 0·908 to 0·98; the upper limit is 0·98 regardless of $1-\alpha$. The objective of the confidence interval approach is to present results of a study in such a way that whenever possible a direct, clear interpretation is facilitated. This can be achieved here: unconditional methods for a confidence interval for θ are described, which yield upper limits between 0·99 and 1 in this example, and have good prior coverage probability characteristics.

3. METHODS BASED ON PROFILE LIKELIHOODS

The likelihood function Λ, which is proportional to $\pi_1^e \pi_2^f \pi_3^g \pi_4^h$, may be reparameterized in terms of the three parameters θ, ψ and v of Table I, to become $v^e(1-v)^h(1-\psi)^{e+h}((\psi+\theta)/2)^f((\psi-\theta)/2)^g$. By the sufficiency principle, the terms in v, the parameter determining the $e:h$ split may be disregarded as not contributing to inferences concerning θ. Thus here, as in existing methods, the distinction between the two cells representing concordant pairs is regarded as uninformative, and the likelihood is essentially trinomial. The true profile-likelihood based $1-\alpha$ confidence region (method 7 below) consists of all values of θ for which $\ln\Lambda \geqslant \ln\Lambda_{\max} - z^2/2$, where $\ln\Lambda_{\max}$ denotes the natural logarithm of the maximum likelihood. In the likelihood $\Lambda(\theta, \psi)$, the *profile estimate* ψ_θ, the MLE for ψ conditional on θ as derived in the Appendix, is substituted for ψ. The likelihood then reduces to a function of the single parameter θ, which has a unique turning point at $\theta = \hat{\theta}$, so the confidence region reduces to an interval. Lower and upper limits at which $\ln\Lambda$ takes the value $\ln\Lambda_{\max} - z^2/2$ are obtained iteratively. When $\hat{\theta} = \pm 1$, that is, $f = n$ or $g = n$, only one limit can be obtained in this way, the other limit being $\hat{\theta}$.

The anti-conservatism that is a drawback of the profile likelihood method in general is already known.[13] Use of a modified profile likelihood[14] would obviate this, though this approach is not

pursued here. Instead, we describe below a method (5) which generalizes the tail-area-based Clopper–Pearson[11] method for the single proportion by choosing limits L and U for θ such that

$$\Pr[\text{tables with more extreme } f-g \mid \theta, \psi_\theta] = \alpha/2.$$

A mid-p interval[15–18] (method 6) may be constructed similarly. The above three methods avoid all aberrations.

4. SIMPLER METHODS BASED ON SCORE INTERVALS

The methods described in Section 3 can only be implemented using programs which include nested iterative processes. Closed-form methods with appropriate properties would be more readily implemented. Wilson[19] proposed an asymptotic score interval for the single proportion p, which is a great improvement on $p \pm z\sqrt{\{p(1-p)/n\}}$. Newcombe[12] derived methods for a difference between independent proportions, in which score intervals for the two proportions are combined; these methods have favourable properties. Corresponding methods for the paired difference may be obtained by incorporating a correction for non-independence based on ϕ, the Pearson product-moment correlation applied to the 2×2 table (Reference 20, p. 59). The resulting intervals remain within $[-1, +1]$ whatever value in this range is substituted for ϕ. A conventional continuity correction may be incorporated in the score intervals.

Careful choice of the estimate $\hat{\phi}$ to substitute is necessary whenever any of the four marginals $e+f$, $g+h$, $e+g$ or $f+h$ is zero. With the usual notation, define $\hat{\phi} = \sqrt{(\chi^2/n)}$ (with the same sign as $eh-fg$) where χ^2 is obtained by summation over all cells with expected frequency > 0. Then $\chi^2 = 0$ in these cases. Hence it is reasonable to substitute 0 for $\hat{\phi}$ in the event of one or more zero cells.

In the absence of a continuity correction the estimated mean coverage is very close to $1-\alpha$, at the expense of serious dips of coverage, especially when v is close to 0·5 and n is small. This suggests an alternative approach, incorporating a continuity correction in $\hat{\phi}$ – as it turns out, preferably only when $\hat{\phi} > 0$.

The three resulting methods have an unusual property: the $e:h$ split is not disregarded. $\hat{\phi}$ is maximal when e or h is $[(e+h)/2]$ and reduces monotonically as $e/(e+h) \to 0$ or 1. Often, though not always, the lower and upper limits for θ display a similar monotonicity property, as shown in Table III.

5. METHODS COMPARED

Ten methods were selected for comparison of which only the first two can violate the $[-1, +1]$ boundaries, in which case the resulting interval is truncated.

1. Asymptotic method without continuity correction ('Unmodified Wald method')[21,7]: $\hat{\theta} \pm z$ se where $\hat{\theta} = (f-g)/n$. The standard error (se) is calculated *without* assuming H_0 (reference 20, p. 117):

$$\text{se} = \{\sqrt{(f+g-(f-g)^2/n)}\}/n = \sqrt{\{((e+h)(f+g)+4fg)/n^3\}}.$$

2. Asymptotic method with continuity correction (reference 20, pp. 116–119): $\hat{\theta} \pm (z\,\text{se} + 1/n)$.

3. 'Exact' method conditional on ψ:[9,10] $(2L_\mu - 1)\hat{\psi}$ to $(2U_\mu - 1)\hat{\psi}$ where $\hat{\psi} = (f+g)/n$ and (L_μ, U_μ) is an 'exact' Clopper–Pearson interval for $\mu = \pi_2/(\pi_2 + \pi_3)$, defined as in reference 2, method 5.
4. 'Mid-p' method conditional on ψ: as method 3, but using a 'mid-p' interval for μ (reference 2, method 6).
5. Unconditional profile likelihood method based on 'exact' tail areas: interval for θ is $[L, U]$ such that

$$\text{(i) if } L \leqslant \theta \leqslant \hat{\theta},\ kP_x + \sum_{x < \xi \leqslant n} P_\xi \geqslant \alpha/2$$

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

   where $P_\xi = \Pr[F - G = \xi \,|\, \theta, \psi_\theta]$, $x = f - g$, and $k = 1$. ψ_θ denotes the MLE of ψ given θ. F and G denote the random variables of which f and g are realizations.
6. Unconditional profile likelihood method based on 'mid-p' tail areas: as method 5, but with $k = 1/2$.
7. True profile likelihood method: all $\theta \in [-1, +1]$ satisfying

$$(e+h)\{\ln(1-\psi_\theta) - \ln(1-\hat{\psi})\} + f\{\ln(\psi_\theta + \theta) - \ln(\hat{\psi} + \hat{\theta})\}$$

$$+ g\{\ln(\psi_\theta - \theta) - \ln(\hat{\psi} - \hat{\theta})\} \geqslant -z^2/2.$$

   Terms corresponding to $e + h, f$ or g equal to zero are omitted.
8. Method based on Wilson[19] score interval for the single proportion without continuity correction. Interval is $\hat{\theta} - \delta$ to $\hat{\theta} + \varepsilon$ where δ and ε are the positive values

$$\delta = \sqrt{(\mathrm{d}l_2^2 - 2\hat{\phi}\mathrm{d}l_2\mathrm{d}u_3 + \mathrm{d}u_3^2)}, \quad \varepsilon = \sqrt{(\mathrm{d}u_2^2 - 2\hat{\phi}\mathrm{d}u_2\mathrm{d}l_3 + \mathrm{d}l_3^2)}.$$

   Here $\mathrm{d}l_2 = (e+f)/n - l_2$, $\mathrm{d}u_2 = u_2 - (e+f)/n$ where l_2 and u_2 are roots of $|\xi - (e+f)/n| = z\sqrt{\{\xi(1-\xi)/n\}}$. Likewise $\mathrm{d}l_3 = (e+g)/n - l_3$, $\mathrm{d}u_3 = u_3 - (e+g)/n$ where l_3 and u_3 are roots of $|\xi - (e+g)/n| = z\sqrt{\{\xi(1-\xi)/n\}}$. Also $\hat{\phi} = (eh - fg)/\sqrt{\{(e+f)(g+h)(e+g)(f+h)\}}$, but $\hat{\phi} = 0$ if this denominator is 0.
9. Method using continuity-corrected score intervals (reference 20, pp. 13–14): as above, but l_2 and u_2 delimit the interval

$$\{\xi : |\xi - (e+f)/n| - 1/(2n) \leqslant z\sqrt{(\xi(1-\xi)/n)}\}$$

   However, if $e + f = 0$, $l_2 = 0$; if $e + f = n$, $u_2 = 1$. Similarly for l_3 and u_3.
10. Method using score intervals but continuity corrected $\hat{\phi}$: as method 8 above, but with the numerator of $\hat{\phi}$ replaced by $\max(eh - fg - n/2, 0)$ if $eh > fg$.

Table II shows 95 per cent confidence intervals calculated by methods 1 to 7 for chosen combinations of $e + h, f$ and g. Table III presents corresponding intervals for methods 8 to 10, showing the effect of varying the $e : h$ split. Asterisks indicate overshoot and tethering aberrations. Though not intended to be a representative selection, these examples show how methods 5 to 7 and 10 obviate the aberrations of the traditional ones.

Table II. 95 per cent confidence intervals for selected combinations of $e + h$, f and g, calculated using seven methods, numbered as in text

Method	Cell frequencies						
	$e + h = 36$ $f = 12$ $g = 2$	$e + h = 36$ $f = 14$ $g = 0$	$e + h = 2$ $f = 97$ $g = 1$	$e + h = 0$ $f = 29$ $g = 1$	$e + h = 2$ $f = 98$ $g = 0$	$e + h = 0$ $f = 30$ $g = 0$	$e + h = 54$ $f = 0$ $g = 0$
1	0·0642, 0·3358	0·1555, 0·4045	0·9126, > 1*	0·8049, > 1*	0·9526, > 1*	1·0*, 1·0	0·0*, 0·0*
2	0·0442, 0·3558	0·1355, 0·4245	0·9026, > 1*	0·7715, > 1*	0·9426, > 1*	0·9667, > 1*	− 0·0185, 0·0185
3	0·0402, 0·2700	0·1503, 0·2800*	0·8711, 0·9795	0·6557, 0·9983	0·9076, 0·9800*	0·7686, 1·0	0·0*, 0·0*
4	0·0575, 0·2662	0·1721, 0·2800*	0·8834, 0·9790	0·6928, 0·9967	0·9210, 0·9800*	0·8099, 1·0	0·0*, 0·0*
5	0·0497, 0·3539	0·1619, 0·4249	0·8752, 0·9916	0·6557, 0·9983	0·9132, 0·9976	0·7686, 1·0	− 0·0660, 0·0660
6	0·0594, 0·3447	0·1691, 0·4158	0·8823, 0·9900	0·6928, 0·9967	0·9216, 0·9966	0·8099, 1·0	− 0·0540, 0·0540
7	0·0645, 0·3418	0·1686, 0·4134	0·8891, 0·9904	0·7226, 0·9961	0·9349, 0·9966	0·8760, 1·0	− 0·0349, 0·0349

*tethering or overshoot aberrations

Table III. 95 per cent confidence intervals for selected combinations of e, f, g and h, calculated using three methods based on score intervals for the single proportion, numbered as in test

e	f	g	h	Method	Interval
36	12	2	0	8	0·0569, 0·3404
				9	0·0407, 0·3522
				10	0·0569, 0·3404
20	12	2	16†	8	0·0618, 0·3242
				9	0·0520, 0·3329
				10	0·0562, 0·3292
18	12	2	18	8	0·0618, 0·3239
				9	0·0520, 0·3327
				10	0·0562, 0·3290
36	14	0	0	8	0·1528[a], 0·4167
				9	0·1360[b], 0·4271
				10	0·1528, 0·4167[c]
35	14	0	1	8	0·1573[a], 0·4149
				9	0·1435[b], 0·4249
				10	0·1461, 0·4175[c]
18	14	0	18	8	0·1504[a], 0·3910
				9	0·1410[b], 0·3989
				10	0·1441, 0·3963[c]
2	97	1	0	8	0·8721, 0·9854
				9	0·8589, 0·9887
				10	0·8721, 0·9854
1	97	1	1	8	0·8737, 0·9850
				9	0·8610, 0·9885
				10	0·8736, 0·9850
0	29	1	0	8	0·6666, 0·9882
				9	0·6189, 0·9965
				10	0·6666, 0·9882
2	98	0	0	8	0·9178, 0·9945
				9	0·9064, 0·9965
				10	0·9178, 0·9945
1	98	0	1	8	0·9174, 0·9916
				9	0·9063, 0·9933
				10	0·9171, 0·9916
0	30	0	0	8	0·8395, 1·0
				9	0·8001, 1·0
				10	0·8395, 1·0
54	0	0	0	8	−0·0664, 0·0664
				9	−0·0827, 0·0827
				10	−0·0664[d], 0·0664[e]

Table III (continued).

e	f	g	h	Method	Interval
53	0	0	1	8	− 0·0640, 0·0640
				9	− 0·0758, 0·0758
				10	− 0·0729[d],0·0729[e]
30	0	0	24	8	− 0·0074, 0·0074
				9	− 0·0079, 0·0079
				10	− 0·0358[d],0·0358[e]
29	0	0	25	8	− 0·0049, 0·0049
				9	− 0·0053, 0·0053
				10	− 0·0354[d],0·0354[e]
28	0	0	26	8	− 0·0025, 0·0025
				9	− 0·0026, 0·0026
				10	− 0·0352[d],0·0352[e]
27	0	0	27	8	0·0*, 0·0*
				9	0·0*, 0·0*
				10	− 0·0351[d],0·0351[e]

* tethering aberrations
Superscripts a to e denote series of limits calculated by the same method showing non-monotonic behaviour as the e:h split varies
† From reference 10, p. 122

6. PRINCIPLES FOR EVALUATION

The present evaluation presupposes the principles summarized by Newcombe.[2,12] In this instance, the following evaluation criteria are considered relevant:

(i) Degree and symmetry of coverage: $L \leqslant \theta \leqslant U$ should occur with probability $1 - \alpha$; $L > \theta$ and $U < \theta$ each with probability $\alpha/2$.
(ii) Expected interval width: as narrow possible, to achieve the desired coverage.
(iii) Avoidance of overshoot: directly calculated limits should satisfy $-1 \leqslant L$ and $U \leqslant 1$.
(iv) Avoidance of point estimate tethering: $\hat{\theta} = L$ or $\hat{\theta} = U$ is regarded as disadvantageous except if $\hat{\theta} = \pm 1$, when it is inevitable. A zero width interval (ZWI) with $L = \hat{\theta} = U$ is a special, bilateral case of tethering, which is always inappropriate. Three (mutually exclusive) circumstances leading to tethering may be identified – corresponding closely (not exactly) to cases 2 to 4 in the Appendix.
   (a) $f = 0$ or $g = 0$ (but not both); $e + h > 0$. Here methods 3 and 4 produce tethering, unnecessarily.
   (b) $f = 0$ or $g = 0$ (but not both); $e + h = 0$. Here $\hat{\theta} = \pm 1$: methods 3 to 10 produce appropriate intervals showing unilateral tethering. Method 1, however, produces a ZWI at ± 1. Method 2's continuity correction avoids this but causes overshoot.
   (c) $f = g = 0$; $e + h > 0$. Here methods 1, 3 and 4 produce ZWIs at 0. Methods 8 and 9, but not 10, produce a ZWI at 0 if also $e = h = n/2$ for n even.

Table IV. Estimated coverage probabilities for 95 per cent confidence intervals calculated by 10 methods. 9100 parameter space points with $10 \leqslant n \leqslant 100$, $0 < \psi < 1$, $0 < \theta < \psi$

	Coverage		Mesial non-coverage		Distal non-coverage	
	Mean	Minimum	Mean	Maximum	Mean	Maximum
Asymptotic						
1 Without CC	0·8543	0·0006	0·1094	0·3170	0·1262	0·9994
2 With CC	0·9690	0·6542	0·0091	0·3170	0·0219	0·3458
Conditional						
3 'Exact'	0·7816	0·0006	0·0106	0·0829	0·2079	0·9994
4 Mid-p	0·7637	0·0006	0·0196	0·1188	0·2166	0·9994
Unconditional						
5 'Exact'	0·9766	0·9546	0·0117	0·0263	0·0117	0·0239
6 Mid-p	0·9657	0·9332	0·0170	0·0372	0·0173	0·0465
7 Profile likelihood	0·9488	0·8539	0·0242	0·0590	0·0270	0·1387
Score						
8 Without CC	0·9505	0·6388	0·0150	0·0474	0·0345	0·3610
9 CC to score limits	0·9643	0·6388	0·0094	0·0277	0·0263	0·3610
10 CC to $\hat{\phi}$	0·9672	0·9031	0·0114	0·0285	0·0214	0·0960

CC: continuity correction

7. EVALUATION OF THE TEN METHODS

The main evaluation (Table IV) is of exactly calculated coverage of nominal 95 per cent intervals based on a large number of parameter space points (PSPs), sampled by random selection of parameter values. In choosing an appropriate pseudo-prior distribution, it is important to note the strong inverse relation between ψ and ϕ; when $\psi > 0{\cdot}5$, $\phi < 0$ except for extreme values of v. To be plausible for the paired case, a pseudo-prior distribution must ensure $\phi \geqslant 0$, though occasional instances of $\hat{\phi} < 0$ are permissible. Pseudo-priors which constrain either ϕ or the inverse tetrachoric odds-ratio $\eta = \pi_2\pi_3/\pi_1\pi_4$ to range on $(0, 1)$ achieve this. After some experimentation the following were selected.

For each of $n = 10, 11, \ldots, 100$, one hundred triples (ϕ, v, μ) were obtained, with ϕ, v and μ sampled independently from $\mathscr{U}(0, 1)$, $\mathscr{U}(1/2, 1)$, $\mathscr{U}(1/2, 1)$, respectively. Pseudo-random numbers were generated using algorithm AS183.[22] Given these parameter values, the corresponding ψ and θ (with $0 < \theta < \psi < 1$) are derived by solving the equations (3), (5) and (7) in Table I by a straightforward iterative process. Their distributions are highly skewed, very rarely approaching the theoretical limit of $+1$, and having means 0·220 and 0·117, medians 0·191 and 0·068 for ψ and θ in the chosen sample of 9100 parameter space points. It is conceded that the lower values of n chosen would yield inadequate power to detect a θ of this order; nevertheless $10 \leqslant n \leqslant 100$ was selected, as a range of values commonly encountered, and generally performance improves as n increases.

To evaluate methods 8 to 10, probabilities of all combinations of e, f, g and h were generated for each sampled PSP, disregarding those below a tolerance of 10^{-12}. Mesial (left) and distal (right) non-coverage probabilities are defined as

$$\mathrm{MNCP} = \sum_{\{e,f,g:\, l > \theta\}} p_{efg}, \quad \mathrm{DNCP} = \sum_{\{e,f,g:\, u < \theta\}} p_{efg}$$

Table V. Estimated coverage probabilities for 90 per cent and 99 per cent confidence intervals calculated by 10 methods. 9100 parameter space points with $10 \leqslant n \leqslant 100$, $0 < \psi < 1$, $0 < \theta < \psi$

	90% intervals		99% intervals	
	Mean	Minimum	Mean	Maximum
Asymptotic				
1 Without CC	0·8089	0·0006	0·8918	0·0006
2 With CC	0·9464	0·6537	0·9860	0·6542
Conditional				
3 'Exact'	0·7569	0·0006	0·8020	0·0006
4 Mid-p	0·7232	0·0006	0·7977	0·0006
Unconditional				
5 'Exact'	0·9453	0·9048	0·9969	0·9909
6 Mid-p	0·9211	0·8695	0·9953	0·9870
7 Profile likelihood	0·8959	0·7100	0·9905	0·9648
Score				
8 Without CC	0·9045	0·5175	0·9859	0·7257
9 CC to score limits	0·9312	0·5681	0·9892	0·7257
10 CC to $\hat{\phi}$	0·9301	0·8572	0·9934	0·9390

CC: continuity correction

where l and u are the calculated limits corresponding to observed cell frequencies e, f, g and $h = n - e - f - g$, and $p_{efg} = \Pr[E = e, F = f, G = g | \psi, \theta, \nu]$. For methods 1 to 7, attention was restricted to $e + h$, f and g, and probabilities below 10^{-10} disregarded. For each method, the probabilities of mesial and of distal non-coverage (Table IV), boundary violation and inappropriate tethering (Table V) were calculated for each PSP by summation, then means and extreme values over the 9100 PSPs obtained.

Additionally, mean and minimum coverage probabilities for nominal 90 per cent and 99 per cent intervals for the same set of 9100 PSPs were calculated (Table V).

To examine coverage of the computationally simpler methods for large denominators but small to moderate discordant cell frequencies, 1000 parameter space points were chosen. $\log_{10}(n)$ was sampled from $\mathcal{U}(3, 5)$, and the resulting n rounded to the nearest integer. Independently, $\log_{10}(2n\psi)$ was sampled from $\mathcal{U}(0, 2)$ and ν and μ from $\mathcal{U}(1/2, 1)$. Coverage of the resulting 95 per cent intervals by methods 1, 2, 8, 9 and 10 was determined.

Expected interval width was calculated exactly for 95 per cent intervals by each method, truncated where necessary, for $n = 10$ and 100 with $\theta = 0$, 0·5 and 0·9 (Table VI).

8. RESULTS

Table IV shows that the coverage probability (CP) for 95 per cent intervals, averaged over the 9100 PSPs, ranges from 0·764 (method 4) to 0·977 (method 5). Methods 1, 3 (despite being based on an 'exact' Clopper–Pearson interval for μ) and 4 are grossly anti-conservative on average, and right non-coverage, due to a ZWI at 0, occurs with probability which tends to 1 as $\psi \to 0$, holding $\mu > 0$ fixed. Even for $50 \leqslant n \leqslant 100$, these methods have mean coverage below 0·9. The *maximum*

CP for method 1 was only 0·9574. Method 2 is very much better, somewhat conservative on average, but occasionally incurs very high non-coverage either to the right or to the left. For these four methods the location of the interval is too mesial.

Method 5 has rather similar coverage characteristics to the Clopper–Pearson method for the single proportion, with a minimum CP of 0·955, appropriate to the 'exact' paradigm, and consequently a rather high mean CP. However the MNCP exceeded 0·025 slightly for 18 of the 9100 PSPs (the most extreme being $n = 48$, $\psi = 0{\cdot}2463$, $\theta = 0{\cdot}1865$, MNCP = 0·0263, DNCP = 0·0179), suggesting that the existence of PSPs for which CP $< 1 - \alpha$ cannot be ruled out. Method 6 yields a somewhat conservative mean CP (as does the mid-p method for a single proportion), and a respectable minimum of 0·9332 ($n = 100$, $\psi = 0{\cdot}0667$, $\theta = 0{\cdot}0660$, MNCP = 0·0321, DNCP = 0·0347). Method 7 is anti-conservative, to a slight degree on average, but with CP only 0·8539 for $n = 64$, $\psi = 0{\cdot}0318$, $\theta = 0{\cdot}0305$, whence MNCP = 0·0141, DNCP = 0·1320. These three methods achieve highly symmetrical coverage, on average. Thus in terms of total coverage probability, methods 5 and 6 are highly appropriate to the 'exact' and mid-p interpretations of $1 - \alpha$, respectively, but are computationally intensive.

The much simpler score-based method 8 achieves a mean CP of almost exactly 0·95, at the cost of many PSPs, not only isolated examples, yielding a CP lower than this. Mean coverage was examined for several zones of values of each of the parameters n, θ, $n\theta$, ψ, $n\psi$, μ, ν and ϕ in turn – for example, for each of $0{\cdot}5 < \nu < 0{\cdot}6$, $0{\cdot}6 < \nu < 0{\cdot}7$, $0{\cdot}7 < \nu < 0{\cdot}8$, $0{\cdot}8 < \nu < 0{\cdot}9$ and $0{\cdot}9 < \nu < 1{\cdot}0$. Method 8's mean CP was anti-conservative for many of these zones, aggregating to the following: $10 \leqslant n < 25$; $0{\cdot}05 < \theta < 1$; $1 < n\theta < 100$; $0{\cdot}1 < \psi < 1$; $5 < n\psi < 100$; $0{\cdot}8 < \mu < 1$; $0{\cdot}5 < \nu < 0{\cdot}8$; and $0 < \phi < 0{\cdot}6$. The dominant determinant of CP was ν, with mean CP only 0·9279 for $0{\cdot}5 < \nu < 0{\cdot}6$. The lowest CP attained was 0·6388, with $\nu = 0{\cdot}5198$ ($n = 54$, $\psi = 0{\cdot}0105$, $\theta = 0{\cdot}0094$, MNCP = 0·0002, DNCP = 0·3610, the result of a probability 0·0585 of a ZWI at 0, together with 6 adjacent near-ZWI configurations, as in the last block of Table III).

The effect of a conventional continuity correction (method 9) is to increase the overall mean CP, and produce a mean CP over 0·95 in all zones examined, except for $0{\cdot}5 < \nu < 0{\cdot}6$ (mean CP 0·9436). The minimum at 0·6388 is unaltered. Applying the continuity correction to ϕ instead (method 10) results in coverage only slightly inferior to method 6, by eliminating the most severe dips; as for methods 5 and 6, the mean CP exceeded 0·95 for each of the zones of the parameter space examined. The three score-based methods err towards mesial location.

Only methods 1 and 2 can yield calculated limits outside $[-1, +1]$. With positive values for θ, the incidence of truncation at -1 was naturally low (2×10^{-7}, 3×10^{-6}, respectively). Truncation at $+1$ occurred with probability 0·0007 for method 1, 0·0020 for method 2. These are much lower than for the single proportion and unpaired difference cases, because high values of θ are barely compatible with the plausibility constraint $\phi > 0$ here. Nevertheless some parameter combinations gave a high overshoot probability for method 1 ($n = 14$, $\psi = 0{\cdot}8680$, $\theta = 0{\cdot}8623$, $\phi = 0{\cdot}0162$, overshoot probability = 0·7596). For method 2, the overshoot probability becomes arbitrarily close to $+1$ as ψ and $\mu \to 1$.

The extreme configurations (2) to (4) in the Appendix which lead to inappropriate tethering or degeneracy, can each occur with arbitrarily high probability for suitable points in the parameter space. Tables with one of the off-diagonal cell frequencies zero, leading to unilateral tethering for methods 3 and 4, occur often, with probability around 0·26 in this evaluation. ZWIs at zero occur with frequency 9·5 per cent for methods 1, 3 and 4, but only 0·1 per cent for methods 8 and 9. Method 1 produces ZWIs at ± 1 much more rarely. These frequencies apply irrespective of the $1 - \alpha$ chosen.

Generally, the coverage properties for 90 per cent and 99 per cent intervals were in line with the findings for 95 per cent intervals (Table V). For methods 1 to 4 the minimum coverage probabilities bore no relation to the intended $1 - \alpha$. At 99 per cent method 7 became conservative on average whilst methods 2, 8 and 9 became anti-conservative, for the chosen set of parameter space points.

For larger values of n, the coverage properties were generally broadly similar to those for $n \leqslant 100$, except that the mean coverage for method 10 was higher at 0·9776; the minimum coverage of method 1 was 0·3960, and the minimum coverages of methods 8 and 9 were poorer at 0·5125 and 0·5164, respectively.

Variation in expected interval width (Table VI) between different methods is most marked when $n\psi$ is low. The width is then least for methods 1, 3 and 4, largely on account of the high ZWI probability.

9. DISCUSSION

The profile-based confidence interval methods for θ effectively overcome the deficiencies of existing methods. They cannot perform anomalously with respect to appropriate (-1 or $+1$) or inappropriate ($\hat{\theta}$) boundaries, nor produce zero-width intervals. In most respects method 7 is an improvement over existing methods, but tends to anti-conservatism. As expected, method 5, with an 'exact' criterion for enumerating tail probabilities, was conservative for all combinations of n, ψ and θ evaluated. The 'mid-p' method 6 has slightly conservative coverage when averaged over these PSPs, and errs only slightly on the anti-conservative side for a few of them.

Thus when interest centres on $\theta = \pi_2 - \pi_3$ the properties of methods 5 and 6, the unconditional method with 'exact' and 'mid-p' criteria for enumerating tail probabilities, are clearly superior to those of existing methods. Methods 3 and 4, which condition on ψ, perform particularly poorly, despite being derived from excellent methods for the single proportion π. The score-based methods for the paired difference do not perform quite as well here as for the single proportion or the unpaired difference, nevertheless method 10, with continuity correction to ϕ, has coverage properties generally similar to those of method 6, yet can be considerably narrower.

Paired designs are often used when the objective is to demonstrate equivalence,[5] in which situation the methods developed here yield appropriate interval estimates. Often equivalence is assessed by testing $H_0: \theta = \Delta$ against $H_1: \theta < \Delta$, for some prespecified $\Delta > 0$, or by two one-sided tests involving Δ and $-\Delta$. The methods presented here may be adapted for use as hypothesis tests in this situation. Thus a p-value may be obtained as that α for which a two-sided $1–2\alpha$ confidence interval just reaches Δ. Alignment of *distal* non-coverage with α is appropriate here.

Methods 5 and 6 are not exclusive to paired binary data; they could be applied to any situation in which there is an underlying trinomial distribution, characterized by probabilities $\pi_{14} + \pi_2 + \pi_3 = 1$, in which π_{14} (replacing $\pi_1 + \pi_4$) is a nuisance parameter and interest centres on $\pi_2 - \pi_3$. For example, subjects may be asked to state their preference between two alternatives A and B. Then p_2 and p_3 may be the proportions of respondents preferring A and B, the remainder expressing no preference. The more tractable score-based methods cannot be applied here without some assumption about the $e:h$ split.

Among alternative approaches not presented in detail in this paper, Lloyd[23] developed a general approach for deriving a confidence interval for a parameter in the presence of a nuisance parameter. The resulting method, which is free of aberrations, was evaluated alongside those included in this paper; even with a 'mid-p' modification, it was much more conservative than

Table VI. Average width of 95 per cent confidence intervals calculated by 10 methods, for selected parameter space points

n	10	10	10	10	10	10	100	100	100	100	100	100
π_1	0·49	0·30	0·20	0·05	0·04	0·01	0·49	0·30	0·20	0·05	0·04	0·01
π_2	0·01	0·20	0·55	0·70	0·91	0·94	0·01	0·20	0·55	0·70	0·91	0·94
π_3	0·01	0·20	0·05	0·20	0·01	0·04	0·01	0·20	0·05	0·20	0·01	0·04
π_4	0·49	0·30	0·20	0·05	0·04	0·01	0·49	0·30	0·20	0·05	0·04	0·01
ψ	0·02	0·40	0·60	0·90	0·92	0·98	0·02	0·40	0·60	0·90	0·92	0·98
θ	0·00	0·00	0·50	0·50	0·90	0·90	0·00	0·00	0·50	0·50	0·90	0·90
Asymptotic												
1 Without CC	0·0706	0·7249	0·6735	0·8740	0·2440	0·2573	0·0496	0·2462	0·2303	0·3138	0·1269	0·1553
2 With CC	0·2706	0·9247	0·8471	1·0114	0·3453	0·3577	0·0696	0·2662	0·2503	0·3338	0·1462	0·1725
Conditional												
3 'Exact'	0·0385	0·6228	0·6388	0·9752	0·6325	0·7147	0·0343	0·2557	0·1830	0·3229	0·1044	0·1717
4 Mid-p	0·0373	0·5788	0·5670	0·8864	0·5374	0·6197	0·0324	0·2400	0·1683	0·3060	0·0912	0·1571
Unconditional												
5 'Exact'	0·6334	0·8794	0·8298	0·9882	0·6647	0·7192	0·0925	0·2589	0·2414	0·3220	0·1421	0·1730
6 Mid-p	0·5357	0·8020	0·7547	0·9157	0·5853	0·6299	0·0804	0·2495	0·2321	0·3128	0·1331	0·1636
7 Profile likelihood	0·3785	0·7448	0·6811	0·8772	0·4579	0·5061	0·0667	0·2473	0·2304	0·3119	0·1291	0·1595
Score												
8 Without CC	0·1784	0·6369	0·6302	0·8418	0·4897	0·5411	0·0499	0·2417	0·2275	0·3100	0·1350	0·1670
9 CC to score limits	0·2046	0·7252	0·7256	0·9685	0·5889	0·6493	0·0524	0·2538	0·2406	0·3279	0·1494	0·1849
10 CC to $\hat{\phi}$	0·3957	0·6957	0·6736	0·8428	0·4934	0·5413	0·0650	0·2447	0·2299	0·3100	0·1359	0·1670

CC: continuity correction

other methods, with mean coverage over 0·99 for a 95 per cent interval. In a very recent publication May and Johnson[7] describe and evaluate two closed-form methods as alternatives to method 1. Of these, the modified Wald method is prone to overshoot much as method 2, whilst the Quesenberry–Hurst method yields a ZWI when $f = g = 0$.

Unfortunately the currently most widely used statistical software does not provide adequate procedures for computing intervals for $\pi_2 - \pi_3$, and until recently no adequate method either in closed form or based on existing tabulations has been available. Provision of software is desirable for score-based methods, which are of closed form but tedious; for methods 5 and 6, essential. Arcus for Windows[24] now incorporates a routine that performs method 5 if $n \leqslant 200$, else method 10. While further refinements or new approaches are worth developing, statistical package producers are strongly urged to make appropriate procedures available for this computationally non-trivial problem.

APPENDIX: DERIVATION OF ψ_θ AS A FUNCTION OF θ

Suppose, without loss of generality, that $f \geqslant g$. There are then four possible situations.

1. *f, g, $e + h$ all > 0.*

Given θ, the likelihood function is a polynomial in ψ, vanishing at $\psi = \pm\theta$ and $\psi = +1$, and positive for any permissible combination of θ and ψ (that is, with $0 \leqslant |\theta| \leqslant \psi \leqslant 1$).

If $\theta = 0$, the likelihood reduces to $(1-\psi)^{e+h}(\psi/2)^{f+g}$ which vanishes only at $\psi = 0$ and 1, is positive between these values, and has a single maximum at the ordinary MLE $\hat{\psi} = p_2 + p_3$.

Otherwise, when $\theta \neq 0$

$$\frac{\partial}{\partial\psi}\ln\Lambda = -\frac{e+h}{1-\psi} + \frac{f}{\psi+\theta} + \frac{g}{\psi-\theta} = \frac{Q(\psi)}{(1-\psi)(\psi^2-\theta^2)}$$

where Q is quadratic in ψ. In terms of its behaviour on $(-\infty, +\infty)$ this derivative has two zeros, those of its numerator, and three discontinuities, at $\psi = \pm\theta$ and $+1$. The second derivative is $\leqslant 0$ for all ψ except these discontinuity points. The maximizing ψ is the larger root of $Q(\psi)$, which always satisfies the constraint $|\theta| < \psi < 1$: $\psi_\theta = B + \sqrt{(B^2 - C)}$, where

$$B = \tfrac{1}{2}(p_2 + p_3) + \tfrac{1}{2}\theta(p_2 - p_3), \text{ and } C = (p_2 - p_3)\theta - (p_1 + p_4)\theta^2.$$

At the usual MLE point $\theta = p_2 - p_3$, $\psi = p_2 + p_3$, the derivative is zero; as this ψ is within the interval $[|\theta|, +1]$ and there is only one zero of $Q(\psi)$ within this interval, it follows that when $\theta = \hat{\theta}$, ψ_θ is simply $\hat{\psi}$. Since from above ψ_θ also takes the value $\hat{\psi}$ when $\theta = 0$, the relationship of ψ_θ to θ is not monotone. Because of the constraint $\psi \geqslant |\theta|$ which is built into the likelihood function, $\psi_\theta = 1$ at $\theta = \pm 1$. For intermediate values of θ, $\psi_\theta < 1$, and ψ_θ regarded as a function of θ has a unique minimum between $\theta = 0$ and $\theta = \hat{\theta}$.

2. *$g = 0$; f and $e + h > 0$.*

In this case the likelihood function reduces to $\Lambda \propto (1-\psi)^{e+h}((\psi+\theta)/2)^f$:

$$\frac{\partial}{\partial\psi}\ln\Lambda = -\frac{e+h}{1-\psi} + \frac{f}{\psi+\theta} = \frac{n(\psi^* - \psi)}{(1-\psi)(\psi+\theta)}$$

since $e + f + h = n$, and where ψ^* (a function of θ) is $p_2 - (1 - p_2)\theta$. This vanishes when $\psi = \psi^*$, which may or may not be in the acceptable domain bounded by $\psi \geqslant |\theta|$, $0 < \psi < 1$. Within the contraints, the likelihood is maximized by $\psi_\theta = \max(\theta, \psi^*)$. ψ^* is a decreasing linear function of θ, and the dependence of ψ_θ on θ is represented by a pair of intersecting straight lines, the decreasing one having $\psi_\theta = 1$ at $\theta = -1$, the increasing one being $\psi = \theta$.

3. $e = h = 0$.

The likelihood reduces to $((\psi + \theta)/2)^f((\psi - \theta)/2)^g$. For any $\theta \in [-1, +1]$, and irrespective of whether both or just one of f and g are non-zero, this is an increasing function of ψ, maximized by $\psi_\theta = 1$. In this situation, under either the 'exact' or the mid-p option, the method reduces to the corresponding conditional one. Method 5 limits for θ are of the form $2\mu_i - 1$, where μ_1 and μ_2 are lower and upper Clopper–Pearson[11] limits for μ.

4. $f = g = 0$; $e + h > 0$.

Here the likelihood $(1 - \psi)^{e+h}$ is maximized by minimizing ψ within the constraint $\psi \geqslant |\theta|$, that is, by taking $\psi_\theta = |\theta|$. When cases 2 and 4 yield a maximizing ψ_θ equal to θ, the corresponding method 5 confidence limit for θ is simply the Clopper–Pearson limit for ψ based on $f + g$ out of n 'successes' (for example, 0·9976, 0·0660 in Table II).

ACKNOWLEDGEMENTS

I thank Professor George Barnard and my colleagues Tim Peters and Ted Coles for their helpful comments, and 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. McNemar, Q. 'Note on the sampling error of the difference between correlated proportions or percentages', *Psychometrika*, **17**, 153–157 (1947).
4. Hope, R. L., Chu, G., Hope, A. H., Newcombe, R. G., Gillespie, P. E. and Williams, S. J. 'Comparison of three faecal occult blood tests in the detection of colorectal neoplasia', *Gut*, **39**, 722–725 (1996).
5. Lu, Y. and Bean, J. A. 'On the sample size for one-sided equivalence of sensitivities based upon McNemar's test', *Statistics in Medicine*, **14**, 1831–1839 (1995).
6. May, W. L. and Johnson, W. D. 'The validity and power of tests for equality of two correlated proportions', *Statistics in Medicine*, **16**, 1081–1096 (1997).
7. May, W. L. and Johnson, W. D. 'Confidence intervals for differences in correlated binary proportions', *Statistics in Medicine*, **16**, 2127–2136 (1997).
8. Breslow, N. E. and Day, N. E. *Statistical Methods in Cancer Research*. 1. *The Analysis of Case-Control Studies*, IARC, Lyon, 1980, pp. 165–166.
9. Liddell, F. D. K. 'Simplified exact analysis of case-referent studies: matched pairs; dichotomous outcome', *Journal of Epidemiology and Community Health*, **37**, 82–84 (1983).
10. Armitage, P. and Berry, G. *Statistical Methods in Medical Research*, 2nd edn, Blackwell, Oxford, 1987, p. 123.
11. 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).
12. Newcombe, R. G. 'Interval estimation for the difference between independent proportions: comparison of eleven methods', *Statistics in Medicine*, **17**, 873–890 (1998).

13. Cox, D. R. and Reid, N. 'Parameter orthogonality and approximate conditional inference', *Journal of the Royal Statistical Society*, *Series B*, **49**, 1–39 (1992).
14. Barndorff-Nielsen, O. E. 'On a formula for the distribution of the maximum likelihood estimate', *Biometrika*, **70**, 343–365 (1983).
15. Lancaster, H. O. 'The combination of probabilities arising from data in discrete distributions', *Biometrika*, **36**, 370–382 (1949).
16. Stone, M. 'The role of significance testing. Some data with a message', *Biometrika*, **56**, 485–493 (1969).
17. Berry, G. and Armitage, P. 'Mid-P confidence intervals: a brief review', *Statistician*, **44**, 412–423 (1995).
18. Miettinen, O. S. *Theoretical Epidemiology*, Wiley, New York, 1985, pp. 120–121.
19. Wilson, E. B. 'Probable inference, the law of succession, and statistical inference', *Journal of the American Statistical Association*, **22**, 209–212 (1927).
20. Fleiss, J. L. *Statistical Methods for Rates and Proportions*, 2nd edn, Wiley, New York, 1981.
21. Altman, D. G. *Practical Statistics for Medical Research*, Chapman and Hall, London, 1991, pp. 236–241.
22. 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.
23. Lloyd, C. J. 'Confidence intervals from the difference between two correlated proportions', *Journal of the American Statistical Association*, **85**, 1154–1158 (1990).
24. Buchan, I. *Arcus QuickStat* (*Biomedical*), Addison Wesley Longman, Cambridge, 1997.

