STATISTICS IN MEDICINE
Statist. Med. **17**, 857–872 (1998)

TWO-SIDED CONFIDENCE INTERVALS FOR THE SINGLE PROPORTION: COMPARISON OF SEVEN METHODS

ROBERT G. NEWCOMBE*

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

SUMMARY

Simple interval estimate methods for proportions exhibit poor coverage and can produce evidently inappropriate intervals. Criteria appropriate to the evaluation of various proposed methods include: closeness of the achieved coverage probability to its nominal value; whether intervals are located too close to or too distant from the middle of the scale; expected interval width; avoidance of aberrations such as limits outside [0, 1] or zero width intervals; and ease of use, whether by tables, software or formulae. Seven methods for the single proportion are evaluated on 96,000 parameter space points. Intervals based on tail areas and the simpler score methods are recommended for use. In each case, methods are available that aim to align either the minimum or the mean coverage with the nominal $1 - \alpha$. © 1998 John Wiley & Sons, Ltd.

1. INTRODUCTION

Applied statisticians have long been aware of the serious limitations of hypothesis tests when used as the principal method of summarizing data. Following persuasion by medical statisticians, for several years the instructions and checklists issued by leading journals, including the *British Medical Journal* and the *American Journal of Public Health*, for the benefit of prospective authors have indicated that in general confidence intervals (CIs) are preferred to *p*-values in the presentation of results. The arguments on which this policy is based are set out by Gardner and Altman.[1] This shift of emphasis, albeit very welcome, presents considerable practical difficulties – these are perhaps greater than when hypothesis tests are used, because the optimization of the latter has received disproportionate attention. A major advantage of confidence intervals in the presentation of results is that interval estimates, in common with point estimates, are relatively close to the data, being on the same scale of measurement, whereas the *p*-value is a probabilistic abstraction. According to Rothman[2] (p. 121) 'confidence intervals convey information about magnitude and precision of effect simultaneously, keeping these two aspects of measurement closely linked'. The usual two-sided confidence interval is thus simply interpreted as a margin of error about a point estimate. Thus any proposed method for setting confidence intervals should be not only *a priori* reasonable, in terms of justifiable derivation and computed coverage probability, but also

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

CCC 0277–6715/98/080857–16$17.50

Received May 1995
Revised July 1997

a posteriori reasonable, preferably for every possible set of data. However, this latter consideration is not always achieved. There are several ways in which this can occur, which we term *aberrations*; the extent of the simplest intervals may be inappropriate due to the bounded nature of the parameter space. Moreover, on account of the discrete distributional form, it is not possible to attain exactly a preset nominal confidence level $1 - \alpha$.

The present investigation concerns the simplest case, the single proportion. Among the extensive literature on this issue, Vollset[3] has evaluated coverage and mean interval width for several methods including the seven evaluated in the present study. Complementary to Vollset's approach, we develop further criteria for examining performance: *exactly calculated coverage probability* based on a *sample* of *parameter space points* and then *summarized*, including calculation of *mean* coverage; the *balance* of *left and right non-coverage* as an indication of *location*; and the incidence of various aberrations. Moreover, this approach is designed to be particularly appropriate to related multiple parameter cases, in particular differences between proportions, for unpaired and paired data. We develop this approach largely to establish a basis for evaluating methods for these cases in subsequent papers.[4,5] The graphical approach,[3] which in the single parameter case produces coverage curves with many discontinuities, is of very limited applicability there.

In setting a confidence interval for a single proportion p, the familiar, asymptotic Gaussian approximation $p \pm z\sqrt{\{p(1-p)/n\}}$ is often used, where n denotes the sample size and z denotes the standard Normal deviate associated with a two-tailed probability α. As well as computational simplicity, this approach has the apparent advantage of producing intervals centred on the point estimate, thus resembling those for the mean of a continuous Normal variate. However, incorporating this kind of symmetry leads to two obvious defects or *aberrations*, namely *overshoot* and *degeneracy*. For low proportions such as prevalences, when the numerator is small the calculated lower limit can be below zero. Conversely, for proportions approaching one, such as the sensitivity and specificity of diagnostic or screening tests, the upper limit may exceed one. The glaring absurdity of overshoot is readily avoided by truncating the interval to lie within [0, 1], of course, but even this is not always done. And a degenerate, *zero width interval* (ZWI) occurs when $p = 0$ or 1, for any $1 - \alpha < 1$. Less obviously, coverage is also very poor. Use[6] of a continuity correction (CC) $1/(2n)$ improves coverage and avoids degeneracy but leads to more instances of overshoot.

These deficiencies, though well known to statisticians, are little heeded in leading journals of major areas of application, as evidenced by the examples[7] cited in Section 3. They may be avoided by a variety of alternative methods.

The 'exact' method of Clopper and Pearson[8] has often been regarded as definitive; it eliminates both aberrations and guarantees strict conservatism, in that the coverage probability is at least $1 - \alpha$ for all θ with $0 < \theta < 1$. It comprises all θ for which precisely computed, 'exact' aggregate tail areas are not less than $\alpha/2$. Numerical values may be obtained iteratively, or by use of published tables (Lentner,[9] pp. 89–102) or the F-distribution.[10] Statistical software that inverts the incomplete beta function may be used, for example, SAS BETAINV $(1 - \alpha/2, r + 1, n - r)$ or Minitab invcdf $1 - \alpha/2$; beta $r + 1, n - r$ produce a Clopper–Pearson upper limit. The Clopper–Pearson method is known to be unnecessarily conservative. A closely related method[11] uses a 'mid-p' enumeration of tail areas[12–14] to reduce conservatism.

A likelihood-based approach[15] has been suggested as theoretically most appealing,[16] by definition it already incorporates an important aspect of symmetry about the maximum

likelihood estimate (MLE) p. There is no question of either continuity 'correction' or mid-p modification to adjust this method's coverage properties systematically.

A computationally much simpler approach due to Wilson,[17] a refinement of the simple asymptotic method, is basically satisfactory; θ is imputed its *true* asymptotic variance $\theta(1-\theta)/n$ and the resulting quadratic is solved for θ. This is more plausible than use of the estimated variance $p(1-p)/n$, and we proceed to show that this results in a good degree and reasonable symmetry of coverage as well as avoidance of aberrations. It has the theoretical advantage amongst asymptotic methods of being derived from the 'efficient score' approach.[18] It has a logit scale symmetry property (Appendix), with consequent log scale symmetry for certain derived intervals.[4] Closed-form solutions for lower and upper limits are available, both without[17] and with[19] continuity correction.

2. METHODS COMPARED

Seven methods were selected for comparison. All are designed to produce two-sided intervals, whenever this is possible given the data; they are constructed so as to try to align lower and upper tail probabilities symmetrically with $\alpha/2$. Only methods 1 and 2 can produce limits outside [0, 1] which are then truncated:

1. Simple asymptotic method ('Wald method' in Vollset[3]) without continuity correction: $p \pm z\sqrt{(pq/n)}$, where z is the $1-\alpha/2$ point of the standard Normal distribution, and $q = 1-p$.
2. Asymptotic method with continuity correction:[6]

$$p \pm (z\sqrt{(pq/n) + 1/(2n)}).$$

3. Wilson[17] 'score' method using asymptotic variance $\theta(1-\theta)/n$ and solving for θ; no continuity correction:

$$(2np + z^2 \pm z\sqrt{(z^2 + 4npq)})/2(n + z^2).$$

4. Score method incorporating continuity correction.[6,19] The interval consists of all θ such that $|p-\theta| - 1/(2n) \leqslant z\sqrt{\{\theta(1-\theta)/n\}}$. Expressions for the lower and upper limits L and U in closed form are available:

$$L = \frac{2np + z^2 - 1 - z\sqrt{\{z^2 - 2 - 1/n + 4p(nq+1)\}}}{2(n+z^2)}$$

$$U = \frac{2np + z^2 + 1 + z\sqrt{\{z^2 + 2 - 1/n + 4p(nq-1)\}}}{2(n+z^2)}.$$

   However, if $p = 0$, L must be taken as 0; if $p = 1$, U is then 1.
5. Method using 'exact' binomial tail areas;[8] the interval is $[L, U]$, with $L \leqslant p \leqslant U$, such that for all θ in the interval:

   (i) if $L \leqslant \theta \leqslant p$

$$kp_r + \sum_{j:r<j\leqslant n} p_j \geqslant \alpha/2,$$

or equivalently

$$\sum_{j:0 \leqslant j<r} p_j + (1-k)p_r \leqslant 1 - \alpha/2;$$

(ii) if $p \leqslant \theta \leqslant U$

$$\sum_{j:0 \leqslant j<r} p_j + kp_r \geqslant \alpha/2$$

respectively, where

$$p_j = \Pr[R = j] = \binom{n}{j} \theta^j (1-\theta)^{n-j},$$

$j = 0, 1, \ldots, n$, R denoting the random variable of which r is the realization, and $k = 1$. As usual an empty summation is understood to be zero.

6. Method using 'mid-p' binomial tail areas:[11] as method 5, but with $k = 1/2$.
7. Likelihood-based method.[15] The interval comprises all θ satisfying

$$r \ln\theta + (n-r)\ln(1-\theta) \geqslant r \ln p + (n-r)\ln(1-p) - z^2/2.$$

The above are recognized not to constitute a complete summary of the literature, but include those methods in common use; many others have been proposed. There are several closed-form approximations to method 5 already mentioned – the Pratt method[20,21] being a very close approximation indeed.[3] Blyth and Still[6] reviewed 'shortened' intervals, defined to be the shortest intervals that ensure strict conservatism, and hence reduce the excess conservatism of method 5. Use[22] of method 6 when $r = 0$ or n slightly reduces the conservatism of method 5, in effect by expending the whole of α in a one-sided way, and conversely, reverses the anti-conservatism of method 7.

All the above methods are *equivariant*;[6] limits for $(n-r)/n$ are complements of those for r/n. Alternative methods[23] based on the Poisson distribution are sometimes used if $r \ll n$; suitable tables are available for constructing 'exact' (Lentner,[9] pp. 152 and 154) and mid-p[24] intervals. In this situation methods 5 and 6 are often computationally unfeasible, if n is very large, or if the proportion and its denominator are available only in rounded form. However, Poisson intervals are wider, by a factor of approximately $1/\sqrt{q}$, than those based on the binomial distribution, hence unnecessarily conservative; moreover, they are not equivariant. Methods 3 and 4 do not have these drawbacks, and are thus preferable, but use of Poisson intervals is unavoidable for rates per person-year of risk.

Bayesian limits consisting of $\alpha/2$ and $1 - \alpha/2$ quantiles of the posterior distribution for θ, for suitably uninformative priors ($\beta(-1, -1)$ or $\beta(0, 0)$) are also available.[25] They are not evaluated here, because the criteria employed, relating to coverage, are not germane to the Bayesian paradigm.

3. ILLUSTRATIVE EXAMPLES

Table I gives 95 per cent confidence intervals for four chosen, illustrative combinations of n and r calculated by the seven methods.

For $p = 81/263$[26] one would clearly expect crude methods to perform reasonably. All of the methods yield very similar intervals. The other three examples are taken from Turnbull *et al.*[7] For 15/148 the relatively low r and p mean that choice of method starts to be more critical. The last

Table I. 95 per cent confidence intervals for selected combinations of n and r, calculated using seven methods. Asterisked values demonstrate aberrations (directly calculated limits outside [0, 1], or zero-width interval)

Method	n / r	263 81	148 15	20 0	29 1
Simple asymptotic					
1 Without CC		0·2522, 0·3638	0·0527, 0·1500	0·0000, 0·0000*	< 0·0000*, 0·1009
2 With CC		0·2503, 0·3657	0·0494, 0·1534	< 0·0000*, 0·0250	< 0·0000*, 0·1181
Score method					
3 Without CC		0·2553, 0·3662	0·0624, 0·1605	0·0000, 0·1611	0·0061, 0·1718
4 With CC		0·2535, 0·3682	0·0598, 0·1644	0·0000, 0·2005	0·0018, 0·1963
Binomial-based					
5 'Exact'		0·2527, 0·3676	0·0578, 0·1617	0·0000, 0·1684	0·0009, 0·1776
6 Mid-p		0·2544, 0·3658	0·0601, 0·1581	0·0000, 0·1391	0·0017, 0·1585
Likelihood-based					
7		0·2542, 0·3655	0·0596, 0·1567	0·0000, 0·0916	0·0020, 0·1432

CC: continuity correction

two cases, 0/20 and 1/29, are clearly ones in which choice of method is very important. Method 2 violates the boundary at 0. So does method 1, when $r = 1$, but not when $r = 0$ – but the resulting degenerate interval is perhaps even less appropriate for interpretation as a margin of error.

It is noteworthy that, irrespective of whether such aberrations occur, the binomial-based methods 6 and 5 produce intervals shifted closer to 0·5, relative to their asymptotic counterparts, 1 and 2, which are constructed symmetrically about the point estimate p. This applies even when there is little obvious difference in interval width. Often in such a situation, though two-sided limits are given, it is the upper one that is paid the most attention, as a kind of upper limit on prevalence. It is of concern if this is too low and hence provides false reassurance. Accordingly, and in line with Rothman's description[2] of confidence intervals quoted in Section 1, we regard interval *location* as of great importance.

4. CRITERIA FOR EVALUATION

It must first be decided what are appropriate criteria by which to evaluate competing methods. The issue of coverage probability, conservatism and interval width is crucial. A confidence interval method is strictly conservative if for all θ, the coverage probability $\mathrm{CP} \geqslant 1 - \alpha$. The 'exact' method[8] seeks to align min CP with $1 - \alpha$, and does so, for reasons set out by Angus.[27] Alternatively, a method may be regarded as conservative on average if $\overline{\mathrm{CP}} = \int \mathrm{CP}(\theta)\,\mathrm{d}f(\theta) \geqslant 1 - \alpha$ for some density function $f(\theta)$ for θ: one may seek to align $\overline{\mathrm{CP}}$ with $1 - \alpha$. A criterion of strict conservatism certainly removes the need to make an essentially arbitrary choice for the *psuedo-prior f*. It also has connotations of 'playing safe', but this is arguably fallacious. Any probability is merely an average of ones and zeros in any case; given the true proportion θ, sometimes the interval computed from the data, $[L, U]$, includes θ, sometimes

it does not, and the coverage probability is the average of ones and zeros imputed in these cases, weighted by their probabilities given θ. While it is of some interest to examine $\min_{0<\theta<1}$ CP,[27,28] either for chosen n or for all n, in that a value $\ll 1-\alpha$ should be regarded as contraindicating the method, nevertheless to average a CP further, over a pseudo-prior f, or a series of representative points in the (n, θ) space sampled accordingly, does no harm, and is arguably more appropriate. If it is intended that the nominal $1-\alpha$ should represent a minimum, methods that are strictly conservative, but with CPs as little above $1-\alpha$ as possible, should be chosen. If $1-\alpha$ is construed as an average, $\overline{\text{CP}}$ should approximate to $1-\alpha$ – ideally a little over $1-\alpha$, with min CP a little under $1-\alpha$. For any n, and any interval (θ_1, θ_2) representing a plausible degree of prior uncertainty for θ, $\overline{\text{CP}}$ should be a little over $1-\alpha$. These two stances lead to different choices of interval, given that CP depends on θ in a discontinuous manner and accordingly shows wide variation. Vollset[3] regarded the very occasional dips in coverage below $1-\alpha$, which occur when the score method with continuity correction is used, as tolerable within the min CP $\geqslant 1-\alpha$ criterion.

The Clopper–Pearson method is frequently labelled 'exact'. This epithet conveys connotations of ideal, 'gold standard' status, so that other methods have been designed to approximate to it[20,21,29–31] or have been evaluated relative to it.[3,27] Any term used in antithesis to 'exact' risks being construed as pejorative. (In the same way, a continuity 'correction' may be more reasonably termed a continuity adjustment; the former term begs the issue of whether the adjustment will be beneficial.) Nevertheless 'exact' can be used to convey several different meanings:

(i) Strictly conservative.
(ii) Use of a multiplier 1 for the probability of the outcome observed, as well as those beyond it, by contrast to the 'mid-p' approach, in which the probability of the observed outcome appears with a coefficient 1/2 in the tail probability.
(iii) Being based on precise enumeration of an assumed underlying probability distribution, such as the binomial or Poisson, not on any asymptotic approximation involving use of a standard error.
(iv) Attaining a CP equal to the nominal $1-\alpha$ for all θ (and n) constituting the parameter space.

Both the Clopper–Pearson method and its 'mid-p' counterpart are exact in sense (iii), but only the former is exact in senses (ii) and (iii). No method can achieve exactness in sense (iv), on account of the discontinuous behaviour of the coverage probability as θ moves past the lower or upper limit corresponding to any possible $r = 0, 1, \ldots, n$. Yet this is the sense that consumers of information presented in CI form are likely to expect it will imply. Analogously, the Fisher test for the 2×2 contingency table is 'exact' in senses (i) to (iii), yet one direct consequence of its strict conservatism is an attained α that is too low, for almost all parameter values. Angus,[27] pointing out that 'the two-sided Clopper–Pearson interval is not 'exact'', is using the term in sense (iv), though this usage is rare in the literature. Angus and Schafer[28] 'make no claims for the optimality of the Clopper–Pearson two-sided CI'. Vollset[3] argues that the continuity-corrected score method may be preferable to the Clopper–Pearson method.

In some,[6] but not all evaluations, it is made explicit whether the nominal coverage probability is intended to be a minimum or an average over the parameter space. The incidences of such aberrations as degeneracy and overshoot have not attracted much attention – in the latter case, perhaps because truncation is an obvious (though unsatisfactory) remedy.

Evaluations are sometimes restricted to 'round' n and values of θ which may be rational with small 'round' denominators which are far from coprime to n. For example, Ghosh[32] considered $n = 15, 20, 30, 50, 100$ and 200; $\theta = 0{\cdot}01, 0{\cdot}05, 0{\cdot}1, 0{\cdot}2, \ldots, 0{\cdot}9, 0{\cdot}95$ and $0{\cdot}99$. The discrete nature of R causes the coverage probabilities to vary discontinuously as θ alters, therefore such choices for n and θ may be atypical and hence inappropriate.

Furthermore, in constructing a $100(1-\alpha)$ per cent confidence interval (L, U) for any parameter θ, the intention that coverage should be $1-\alpha$ imposes only one constraint on the choice of L and U. A family of intervals (L_λ, U_λ) may be constructed, indexed by a parameter λ, $0 \leqslant \lambda \leqslant 1$, where $\lambda_1 < \lambda_2$ implies $L_{\lambda_1} < L_{\lambda_2}$ and $U_{\lambda_1} < U_{\lambda_2}$, and $\lambda = 0$ and $\lambda = 1$ correspond to one-tailed $100(1-\alpha)$ per cent intervals. One criterion[33,34] for choice among such intervals is minimization of width, which has an appealing interpretation as an integral of the probability of including false values.[35] An alternative criterion[14,11] is equalization of tail probabilities. The two criteria lead to the same choice for interval estimation of the mean of a Normal distribution, but not necessarily in other contexts. The quotation from Rothman[2] (Section 1) suggests that as a prerequisite to meaningful interpretation interval location is as important as width, and arguably should not be left to follow as an indirect consequence of minimization of width. Nevertheless hitherto evidence on left and right non-coverage separately has been lacking.

Evaluation of equivariant methods may be restricted to θ between 0 and $0{\cdot}5$, without loss of generality, permitting a more pertinent assessment of symmetry of coverage. If θ has a symmetrical distribution on $[0, 1]$, the true left and right non-coverage probabilities are necessarily equal; comparing them cannot help to assess performance. When the distribution of θ is restricted to $[0, 0{\cdot}5]$, and the attained left and right non-coverage probabilities are enumerated separately, they are interpretable as *distal* and *mesial* non-coverage probabilities (DNCP and MNCP), respectively. It is desirable that these should be equal, otherwise the method is regarded as producing intervals that are either too close to $0{\cdot}5$ or (arguably more seriously) too far from $0{\cdot}5$. Likewise, violations of the nearby and remote boundaries at 0 and 1 are to be enumerated separately.

Vollset[3] presented graphs showing the relationship of CP to θ for $n = 10, 100$ and 1000 for several methods including those evaluated here, but did not attempt to assess mean coverage properties. Closed-form expressions for mean CP, DNCP and MNCP for a given conjugate prior distribution for θ may be obtained as weighted sums of incomplete beta integrals. An alternative approach is adopted here: parameter space points are sampled from a suitable pseudo-prior, permitting assessment of both average and extreme coverage properties.

Complementary to average coverage is average width. This may be computed for a given θ, averaging the widths of intervals for $r = 0, 1, \ldots, n$ according to their probabilities given θ, or averaged further with respect to a pseudo-prior $f(\theta)$. It is desirable to achieve the required coverage with the least width.

Meaningful evaluation of average width presupposes truncation of any overshoot. Interval width is not invariant under monotone transformation, and furthermore its direct application to inherently asymmetrical, right-unbounded measures such as the rate ratio, the odds ratio or the Poisson parameter would be problematic, but these points do not invalidate its use for the single proportion, nor for differences between proportions. According to the mid-p criterion, the degree to which an interval method's $\overline{\mathrm{CP}}$ exceeds $1-\alpha$ may be regarded as an expression of unnecessary width, but that does not make $\overline{\mathrm{CP}}$ a measure of width. For example, with $n = 5$ and $\theta \sim \mathrm{U}(0, 1)$, methods 2 and 3 have similar mean widths for 95 per cent intervals, $0{\cdot}562$ and $0{\cdot}558$, but very

different mean coverages, 0·815 and 0·955, respectively; the score method produces more appropriately located intervals and thus expends its width more effectively.

Overshoot, the violation of the boundaries inherent to the problem (0 and 1), incurs a serious risk of its transmission, unchecked, along the publication chain.[7] This can always be coped with by truncation, or equivalently by careful specification of algorithms in inequality form so as to avoid it, but that obscures the nature of the problem. Thus for $r = 1$, the method 1 lower limit is generally negative and would be truncated to 0, but $\theta = 0$ is ruled out by the data, $\Pr[R \geqslant 1|\theta = 0] = 0$. Closely adjoining parameter values are also highly implausible: for $\theta = \varepsilon/n$, $\Pr[R \geqslant 1\,|\,\theta] \cong \varepsilon$, which can be made arbitrarily small by choice of ε. To be plausible the lower limit needs to be positive, not zero.

5. EVALUATION OF THE SEVEN METHODS

In the main evaluation the performance of 95 per cent intervals calculated by the chosen methods was evaluated for 96,000 representative parameter space points (PSPs), with $5 \leqslant n \leqslant 100$ and $0 < \theta < 0{\cdot}5$. For each n independently, 1000 unround θ values were chosen randomly from uniform distributions on $[(j-1)/2000, j/2000]$, using algorithm AS183.[36] This simple sampling scheme was chosen after some experimentation to give a reasonable degree of weighting towards situations in which asymptotic behaviour would be a poor approximation. Though in many respects it resembles a joint prior for n and θ, this is not the intended interpretation – n would generally be predetermined, and what range of θ would be plausible varies according to context – use of this pseudo-prior is merely an expedient to smooth out discontinuities and approximate the performance that might be obtained in practice. The investigation was not oriented towards any prior partitioning of the parameter space, but a major objective was to determine in which parts of the parameter space computationally simple methods might be acceptable. Accordingly, the relationship of coverage properties of each method to n, θ and $n\theta$ was examined.

Programs were developed to generate 'tables' listing the seven types of CI for each $n = 5, 6, \ldots, 100$. For each n and each θ chosen to accompany it, the binomial probability $\Pr[R = r\,|\,n, \theta]$ was generated, for each r for which it was non-negligible (using a tolerance of 10^{-10}). Probabilities of each direction of non-coverage, boundary violation and degeneracy as described above were summated across all r with non-negligible probability, to give exactly computed measures for each chosen PSP. These were then summarized across the randomly chosen set of PSPs.

It is conceded that the minimum CP found by examining a large number of PSPs will not generally identify the absolute minimum over all possible parameter values. With a large number of PSPs, the empirical minimum approaches the true minimum, for example, for method 3, the true minimum is 0·831. The justification for the approach is (i) mean CP is estimated essentially by Monte Carlo integration using systematic sampling, and (ii) it facilitates estimation in cases involving several parameters[4,5] in which exact determination of min CP is possible but more difficult.

Additionally, mean and minimum coverage probabilities for nominal 90 per cent and 99 per cent intervals for the same set of 96,000 parameter space points were also calculated.

To examine coverage for proportions with large denominators but small to moderate numerators, as often encountered in epidemiology, 1000 parameter space points were chosen. $\log_{10} n$ was sampled from U(2, 5), and the resulting n rounded to the nearest integer. Independently,

$\log_{10}(4n\theta)$ was sampled from U(0, 2). Coverage of the resulting 95 per cent intervals was determined.

The above approach was chosen as appropriate for evaluation of coverage, for which the intended value is a constant but dependence on θ and consequently n also is locally highly discontinuous. A different approach is appropriate for evaluation of expected interval width, which is grossly dependent on n and θ, but in a smooth manner. Expected interval width is calculated exactly for 95 per cent intervals by each method, for selected combinations of n and θ: here, each of $n = 5, 20$ and 100 with $\theta = 0{\cdot}5, 0{\cdot}2$ and $0{\cdot}05$. Furthermore, for each of the above values of n, the expected width for θ sampled from U[0, 1] is obtained directly, as then $\Pr[R = r] = 1/(n + 1)$ for $r = 0, 1, \ldots, n$.

6. RESULTS

Table II shows the mean and minimum coverage probabilities, and mean and maximum distal and mesial non-coverage probabilities, based on all 96,000 chosen parameter space points, for 95 per cent intervals.

The overall average CP ranged from 0·881 (method 1) to 0·971 (methods 4 and 5). On average method 2 is anti-conservative despite the continuity correction. The likelihood-based method 7 is also slightly anti-conservative on average. Conversely methods 3 and 6 are slightly conservative on average, with average CP close above 0·95, despite being non-CC and mid-p, respectively. The average DNCP was 0·032 for method 3, and less than the nominal 0·025 for all other methods. The average MNCP was 0·101 for method 1, 0·063 for method 2, 0·029 for method 7, else $< 0{\cdot}025$.

Methods 1 and 2 produced many totally unacceptable CPs. Methods 3 and 7 are capable of yielding CP below 0·9. Method 5 is strictly conservative, by a minute margin: for $n = 45$, $\theta = 0{\cdot}4432$, we obtain DNCP = 0·02494, MNCP = 0·02493. Method 4 is capable of being slightly anti-conservative, for example, $n = 12$, $\theta = 0{\cdot}0043$, DNCP = 0·0509, MNCP = 0 (see Vollset[3]).

Methods 3 and 4, and to a lesser degree 5 and 6, produced intervals too close to 0·5. Conversely, method 7 produced intervals slightly too far away from 0·5.

As expected, the mean coverage of methods 1 and 2 was heavily dependent upon n, θ and $n\theta$ (Table III). For method 3, $\overline{\text{CP}}$ was remarkably close to being constant, with respect to each of n, θ and $n\theta$ separately, and indeed jointly. Methods 4 and 5 had a conservative $\overline{\text{CP}}$, especially for low values of n, θ and $n\theta$; cross-tabulations indicated $n\theta$ was the dominant determinant of $\overline{\text{CP}}$ (though this is not obvious from Table III). The pattern was similar for method 6, but the $\overline{\text{CP}}$ was close to 0·95 when n, θ or $n\theta$ was large. Method 7 was slightly anti-conservative, except for low θ; here θ is the dominant determinant of coverage.

Methods 3 to 7 are incapable of violating the inherent [0, 1] bounds for θ. The probability of obtaining an interval entirely within [0, 1] averaged 0·730 and 0·838 for methods 2 and 1 (Table IV). The boundary violated was almost always the nearby one at 0. Some combinations of n and θ yield very high probabilities of boundary violation with these methods, in particular the nearer boundary, though the probability of violating the distant boundary 1 can also approach 0·5.

Zero width intervals can occur only with method 1, and then with probability $\theta^n + (1 - \theta)^n$ if $0 < \theta < 1$. This can take values arbitrarily close to 1, as $\theta \to 0$ (or 1); correspondingly, MNCP is arbitrarily close to 1. For given n, and θ uniform on [0, 1], the ZWI probability is $2/(n + 1)$; averaged over $n = 5$ to 100, this is 0·0607.

Table II. Estimated coverage probabilities, for 95 per cent confidence intervals calculated by 7 methods. From 96,000 parameter space points (n, θ) with $5 \leqslant n \leqslant 100$, $0 < \theta < 0{\cdot}5$

Method	Coverage		Distal non-coverage		Mesial non-coverage	
	Mean	Minimum	Mean	Maximum	Mean	Maximum
Simple asymptotic						
1 Without CC	0·8814	0·0002	0·0172	0·1304	0·1014	0·9998
2 With CC	0·9257	0·3948	0·0113	0·0701	0·0630	0·6052
Score method						
3 Without CC	0·9521	0·8322	0·0317	0·1678	0·0162	0·0578
4 With CC	0·9707	0·9491	0·0196	0·0509	0·0097	0·0246
Binomial-based						
5 'Exact'	0·9710	0·9501	0·0163	0·0250	0·0127	0·0250
6 Mid-p	0·9572	0·9121	0·0233	0·0483	0·0196	0·0500
Likelihood based						
7	0·9477	0·8019	0·0238	0·0668	0·0285	0·1465

CC: continuity correction

Table III. Estimated coverage probabilities related to n, θ and $n\theta$

Method	Region of parameter space						
	All	n 5 to 10	n 91 to 100	θ 0 to 0·05	θ 0·45 to 0·5	$n\theta$ 0 to 5	$n\theta$ 45 to 50
	Number of points						
	96000	6000	10000	9600	9600	28584	569
Simple asymptotic							
1 Without CC	0·8814	0·7151	0·9211	0·5785	0·9358	0·7557	0·9455
2 With CC	0·9257	0·8482	0·9441	0·8225	0·9547	0·8623	0·9570
Score method							
3 Without CC	0·9521	0·9545	0·9512	0·9518	0·9504	0·9548	0·9502
4 With CC	0·9707	0·9844	0·9650	0·9795	0·9668	0·9799	0·9610
Binomial-based							
5 'Exact'	0·9710	0·9868	0·9648	0·9872	0·9656	0·9836	0·9605
6 Mid-p	0·9572	0·9726	0·9531	0·9767	0·9522	0·9689	0·9508
Likelihood based							
7	0·9477	0·9465	0·9486	0·9613	0·9481	0·9463	0·9494

CC: continuity correction

Generally, the coverage properties of 90 per cent and 99 per cent intervals (Table V) were in line with the findings for 95 per cent intervals, though method 3 was slightly anti-conservative on average at 99 per cent for the chosen set of parameter space points.

For larger values of n, coverage properties for 95 per cent intervals were generally maintained (Table VI), but those of methods 1 and 2 declined greatly whilst method 7 became conservative on average.

Table IV. Estimated probabilities of achieving an interval within [0, 1], and of directly calculated limits L and U violating bounds, for 95 per cent confidence intervals calculated by simple asymptotic methods. From 96,000 parameter space points (n, θ) with $5 \leqslant n \leqslant 100$, $0 < \theta < 0{\cdot}5$

Method	Within bounds probability Pr $[0 \leqslant L \,\&\, U \leqslant 1]$		Pr $[L < 0]$		Pr $[U > 1]$	
	Mean	Minimum	Mean	Maximum	Mean	Maximum
Simple asymptotic						
1 Without CC	0·8380	0·0625	0·1584	0·7598	0·0035	0·4683
2 With CC	0·7303	0·0	0·2637	1·0	0·0060	0·4995

CC: continuity correction

Table V. Estimated coverage probabilities, for 90 per cent and 99 per cent confidence intervals calculated by 7 methods. From 96,000 parameter space points (n, θ) with $5 \leqslant n \leqslant 100$, $0 < \theta < 0{\cdot}5$

Method	90% intervals		99% intervals	
	Mean	Minimum	Mean	Minimum
Simple asymptotic				
1 Without CC	0·8379	0·0002	0·9197	0·0002
2 With CC	0·8947	0·3947	0·9521	0·3948
Score method				
3 Without CC	0·9047	0·7909	0·9890	0·8874
4 With CC	0·9390	0·9009	0·9940	0·9676
Binomial-based				
5 'Exact'	0·9384	0·9001	0·9948	0·99001
6 Mid-p	0·9112	0·8254	0·9921	0·9824
Likelihood based				
7	0·8955	0·6514	0·9896	0·9369

CC: continuity correction

Variation in expected interval width for 95 per cent intervals (Table VII) between different methods is most marked when $n\theta$ (or $n(1 - \theta)$) is low. The width is then least for method 1, largely on account of the high ZWI probability.

7. DISCUSSION

Method 1, the simplest and most widely used, is very anti-conservative on average, with arbitrarily low CP for low θ. Indeed, the *maximum* coverage probability is only 0·959; min DNCP is 0 and min MNCP is 0·0205. In this evaluation with $\theta < 0{\cdot}5$, the deficient coverage probability stems from right non-coverage; the interval does not extend sufficiently far to the right, as evidenced by the high frequency of ZWIs and the fact that a large part of the calculated interval may lie beyond the nearer boundary, 0. For general θ, this means the interval is positioned too far from 0·5 to attain symmetry in the more pertinent sense of equalizing mesial and distal non-coverage.

Table VI. Estimated coverage probabilities, for 95 per cent confidence intervals calculated by 7 methods. From 1000 parameter space points (n, θ) with $100 \leqslant n \leqslant 100000$, $0{\cdot}25 < n\theta < 25$

Method	Coverage	
	Mean	Minimum
Simple asymptotic		
1 Without CC	0·7279	0·2229
2 With CC	0·8530	0·3938
Score method		
3 Without CC	0·9535	0·8949
4 With CC	0·9731	0·9520
Binomial-based		
5 'Exact'	0·9788	0·9507
6 Mid-p	0·9656	0·9165
Likelihood based		
7	0·9575	0·8411

Method 2, incorporating the continuity correction, is an improvement in some respects, but is still very inadequate, also being highly anti-conservative and asymmetrical in coverage, and incurs an even higher risk of violating the nearer boundary, largely but not entirely instead of the ZWIs.

Even though, for large n and mesial p (for example, for 81/263 in Table I), methods 1 and 2 approximate acceptably to the better methods, it is strongly recommended that intervals calculated by these methods should no longer be acceptable for the scientific literature; highly tractable alternatives are available which perform much better. Use of the simple asymptotic standard error of a proportion should be restricted to sample size planning (for which it is appropriate in any case) and introductory teaching purposes.

The average coverage probability of the score method 3 is very close to the nominal value. For some n and θ the CP can be considerably lower – for a 95 per cent interval, as low as 0·831, occurring at $\theta \cong 0{\cdot}18/n$, and 0·89 for a nominal 99 per cent interval. Average left and right non-coverage probabilities are 0·032 and 0·016, thus the interval tends to be located too close to 0·5 – an overcorrection of the asymmetry of method 1. However, these are its only drawbacks; it is nearly as easy to calculate as method 1, but greatly superior, and involves neither aberrations nor special cases when $r = 0$ or n.

The score method's continuity-corrected counterpart, method 4, is nearly strictly conservative, with minimum coverage 0·949. Consequently the average coverage, 0·971, is quite conservative, which may be interpreted to mean the interval is simply unnecessarily wide. With distal and mesial non-coverage probabilities 0·020 and 0·010, these intervals likewise are located too close to 0·5.

The classical Clopper–Pearson method 5, the 'gold standard' of the strictly conservative criterion, has average coverage characteristics similar to method 4; again the location is slightly too mesial, though less so than method 4. The empirical minimum coverage and maximum mesial and distal non-coverage are practically identical to the theoretical values of 0·95 and 0·025.

Table VII. Average width of 95 per cent confidence intervals calculated using seven methods. Selected values of n; selected values of θ, and θ uniform on [0, 1]

n	θ	Simple asymptotic		Score method		Binomial-based		Likelihood-based
		No CC	CC	No CC	CC	'Exact'	mid-p	
		1	2	3	4	5	6	7
5	0·5	0·6904	0·7904	0·6183	0·7225	0·7553	0·6981	0·6624
	0·2	0·4414	0·5414	0·5540	0·6573	0·6720	0·6111	0·5451
	0·05	0·1308	0·2308	0·4707	0·5733	0·5667	0·4991	0·3884
	Uniform	0·4600	0·5600	0·5581	0·6616	0·6779	0·6168	0·5516
20	0·5	0·4268	0·4768	0·3927	0·4342	0·4460	0·4129	0·4076
	0·2	0·3263	0·3659	0·3256	0·3667	0·3671	0·3362	0·3254
	0·05	0·1225	0·1479	0·2188	0·2586	0·2380	0·2095	0·1808
	Uniform	0·3160	0·3564	0·3254	0·3663	0·3661	0·3348	0·3218
100	0·5	0·1950	0·2050	0·1914	0·2010	0·2024	0·1936	0·1932
	0·2	0·1556	0·1656	0·1543	0·1639	0·1640	0·1555	0·1545
	0·05	0·0815	0·0896	0·0884	0·0979	0·0942	0·0867	0·0839
	Uniform	0·1518	0·1614	0·1523	0·1619	0·1614	0·1531	0·1517

CC: continuity correction

The 'mid-p' binomial-based method 6, with average coverage 0·957 and minimum 0·912, is highly acceptable according to the criterion that seeks to align $\overline{\mathrm{CP}}$ with $1-\alpha$. With average distal and mesial non-coverage probabilities 0·023 and 0·020, it is also located slightly too mesially.

The likelihood-based method 7, a 'worthy alternative' to method 6,[15] is in fact slightly anti-conservative, with average coverage probability 0·948. This is similar to the $\overline{\mathrm{CP}}$ of 0·949 obtained for the profile-likelihood-based unconditional CI for the paired difference of proportions.[5] On average, distal and mesial aspects of non-coverage are reasonably closely balanced, however, for some PSPs there is considerable mesial non-coverage, up to 0·1465 which is $\exp(-z^2/2)$. The minimum coverage is barely above 0·8.[37]

Methods 5 and 6, which were set up in terms of tail areas, thus have better total coverage properties than method 7, which is based on the likelihood function. The same occurs for the corresponding unconditional methods for unpaired[4] and paired[5] differences in proportions. This suggests that, generally, likelihood-based interval methods may not perform very well when evaluated in terms of coverage. An alternative interpretation is that they should rather be regarded as leading to a different kind of interval estimate, which should be called 'likelihood interval' to distinguish it from a confidence interval (or indeed a Bayes interval).

8. CONCLUSION

Choice of method must depend on an explicit decision whether to align minimum or mean coverage with $1-\alpha$. For the conservative criterion, the Clopper–Pearson method is readily available, from extensive tabulations, and also software. The Pratt closed form approximation is a very close one, but requires a programmable calculator or software. Vollset[3] argues for preferring method 4. The more complicated shortened intervals[6] are less conservative than Clopper–Pearson, and deserve to be made available in software. According to the $\overline{\mathrm{CP}} = 1-\alpha$ criterion, method 6 performs very well; method 3 also performs well, and has the advantage of a simple closed form, equally applicable whether n is 5 or 50 million.

The most widely-used general statistical software packages are largely oriented towards hypothesis testing and do not serve to encourage the user to present appropriate CIs for proportions or related quantities. Neither SAS nor Minitab draws attention to the availability of Clopper–Pearson intervals indirectly by using the inverse of the beta integral; SPSS provides nothing for what their authoring teams must have regarded as a trivial task. The package CIA,[38] designed specifically for calculating CIs, provides method 1 or 5 depending on n and r; the criteria determining the choice are not made clear. StatXact[22] uses a hybrid of methods 5 and 6, as described above. Of the methods that perform well, only the score method is calculator-friendly. Statistical package producers are strongly urged to direct users to appropriate procedures for the very basic, but computationally non-trivial, task of setting confidence intervals for proportions.

APPENDIX: LOGIT SCALE SYMMETRY OF THE WILSON SCORE INTERVAL

The anomalous behaviour of the simple asymptotic interval is a consequence of its symmetry on the additive scale. By imputing a variance based on θ instead of p, the Wilson[17] score interval replaces this property with a more appropriate logit scale symmetry.

The Wilson limits L and U are the roots of the quadratic

$$\mathbf{F}_p = \theta^2(1+a) - \theta(2p+a) + p^2 = 0$$

where $a = z^2/n$. Their product is thus $LU = p^2/(1+a)$. Similarly, $1-L$ and $1-U$ satisfy $\mathbf{F}_q = 0$ where $q = 1-p$, so $(1-L)(1-U) = q^2/(1+a)$. Consequently, assuming $q \neq 0$, $(L/(1-L))(U/(1-U)) = p^2/q^2$, thus $\text{logit}(p) - \text{logit}(L) = \text{logit}(U) - \text{logit}(p)$, and the interval for p/q is symmetrical on a multiplicative scale. The same property applies in a nugatory way if q or p is zero.

ACKNOWLEDGEMENTS

I thank Professor G. A. Barnard for helpful discussion of the likelihood-based method, Professor O. S. Miettinen for correspondence relating to Bayesian methods, and two anonymous referees for many helpful suggestions.

REFERENCES

1. Gardner, M. J. and Altman, D. G. (eds). *Statistics with Confidence. Confidence Intervals and Statistical Guidelines*, British Medical Journal, London, 1989.
2. Rothman, K. *Modern Epidemiology*. Little Brown, Boston, 1986.
3. Vollset, S. E. 'Confidence intervals for a binomial proportion', *Statistics in Medicine*, **12**, 809–824 (1993).
4. Newcombe, R. G. 'Interval estimation for the difference between independent proportions: comparison of eleven methods', *Statistics in Medicine*, **17**, 873–890 (1998).
5. Newcombe, R. G. 'Improved confidence interval methods for the difference between binomial proportions based on paired data'. Submitted for publication.
6. Blyth, C. R. and Still, H. A. 'Binomial confidence intervals', *Journal of the American Statistical Association*, **78**, 108–116 (1983).
7. Turnbull, P. J., Stimson, G. V. and Dolan, K. A. 'Prevalence of HIV infection among ex-prisoners in England', *British Medical Journal*, **304**, 90–91 (1992).
8. 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).
9. Lentner, C. (ed). *Geigy Scientific Tables*, 8th edition, volume 2, Ciba-Geigy, Basle, 1982.
10. Miettinen, O. S. 'Estimation of relative risk from individually matched series', *Biometrics*, **26**, 75–86 (1970).
11. Miettinen, O. S. *Theoretical Epidemiology*, Wiley, New York, 1985, pp. 120–121.
12. Lancaster, H. O. 'The combination of probabilities arising from data in discrete distributions', *Biometrika*, **36**, 370–382 (1949).
13. Stone, M. 'The role of significance testing. Some data with a message', *Biometrika*, **56**, 485–493 (1969).
14. Berry, G. and Armitage, P. 'Mid-P confidence intervals: a brief review', *Statistician*, **44**, 417–423 (1995).
15. Miettinen, O. S. and Nurminen, M. 'Comparative analysis of two rates', *Statistics in Medicine*, **4**, 213–226 (1985).
16. Barnard, G. A. Personal communication, 1992.
17. Wilson, E. B. 'Probable inference, the law of succession, and statistical inference', *Journal of the American Statistical Association*, **22**, 209–212 (1927).
18. Cox, D. R. and Hinkley, D. V. *Theoretical Statistics*, Chapman and Hall, London, 1974.
19. Fleiss, J. L. *Statistical Methods for Rates and Proportions*, 2nd edn, Wiley, New York, 1981.
20. Pratt, J. W. 'A normal approximation for binomial, *F*, beta, and other common, related tail probabilities. II', *Journal of the American Statistical Association*, **63**, 1457–1483 (1968).
21. Blyth, C. R. 'Approximate binomial confidence limits', *Journal of the American Statistical Association*, **81**, 843–855 (1986).
22. Mehta, C. and Patel, N. *StatXact. Statistical Software for Exact Nonparametric Inference*, Version 2. Cytel, Cambridge, MA, 1991.
23. Garwood, F. 'Fiducial limits for the Poisson distribution', *Biometrika*, **28**, 437–442 (1936).
24. Cohen, G. R. and Yang, S. Y. 'Mid-*p* confidence intervals for the Poisson expectation', *Statistics in Medicine*, **13**, 2189–2203 (1994).

25. Lindley, D. V. *Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2. Inference*, Cambridge University Press, Cambridge, 1965, pp. 141–148.
26. Altman, D. G. *Practical Statistics for Medical Research*, Chapman and Hall, London, 1991, pp.161–165.
27. Angus, J. E. 'Confidence coefficient of approximate two-sided confidence intervals for the binomial probability', *Naval Research Logistics*, **34**, 845–851 (1987).
28. Angus, J. E. and Schafer, R. E. 'Improved confidence statements for the binomial parameter', *American Statistician*, **38**, 189–191 (1984).
29. Anderson, T. W. and Burstein, H. 'Approximating the upper binomial confidence limit', *Journal of the American Statistical Association*, **62**, 857–861 (1967).
30. Anderson, T. W. and Burstein, H. 'Approximating the lower binomial confidence limit', *Journal of the American Statistical Association*, **63**, 1413–1415 (1968).
31. Fujino, Y. 'Approximate binomial confidence limits', *Biometrika*, **67**, 677–681 (1980).
32. Ghosh, B. K. 'A comparison of some approximate confidence intervals for the binomial parameter', *Journal of the American Statistical Association*, **74**, 894–900 (1979).
33. Kendall, M. G. and Stuart, A. *The Advanced Theory of Statistics. Volume 2. Inference and Relationship*, 2nd edn, Griffin, London, 1967, pp. 101–102.
34. Bickel, P. J. and Doksum, K. A. *Mathematical Statistics*: *Basic Ideas and Selected Topics*, Holden-Day, San Francisco, 1977, p. 155.
35. Pratt, J. W. and Gibbons, J. D. *Concepts of Nonparametric Theory*, Springer Verlag, New York, 1981.
36. Wichmann, B. A. and Hill, I. D. 'An efficient and portable pseudorandom number generator', *in* Griffiths, P. and Hill, I. D. (eds), *Applied Statistics Algorithms*, Ellis Horwood, Chichester, 1985.
37. Newcombe, R. G. 'Confidence intervals for a binomial proportion', *Statistics in Medicine*, **13**, 1283–1285 (1994).
38. Gardner, M. J. *Confidence Interval Analysis*, British Medical Journal, London, 1989.

