STATISTICS IN MEDICINE
Statist. Med. 2003; **22**:611–621 (DOI: 10.1002/sim.1320)

Confidence intervals for the binomial parameter: some new considerations

Jenő Reiczigel[*,†]

Szent István University, Faculty of Veterinary Science, Department of Biomathematics and Informatics, Budapest, István u. 2., H-1078, Hungary

SUMMARY

Several methods have been proposed to construct confidence intervals for the binomial parameter. Some recent papers introduced the 'mean coverage' criterion to evaluate the performance of confidence intervals and suggested that exact methods, because of their conservatism, are less useful than asymptotic ones. In these studies, however, exact intervals were always represented by the Clopper–Pearson interval (C–P). Now we focus on Sterne's interval, which is also exact and known to be better than the C–P in the two-sided case. Introducing a computer intensive level-adjustment procedure which allows constructing intervals that are exact in terms of mean coverage, we demonstrate that Sterne's interval performs better than the best asymptotic intervals, even in the mean coverage context. Level adjustment improves the C–P as well, which, with an appropriate level adjustment, becomes equivalent to the mid-*P* interval. Finally we show that the asymptotic behaviour of the mid-*P* method is far poorer than is generally expected. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: exact confidence interval; test inversion, minimum coverage; mean coverage; binomial proportion

1. INTRODUCTION

The construction of confidence intervals for the binomial parameter has an extensive literature. Several methods, asymptotic as well as exact, have been proposed. Vollset [1] compared 13 methods and recommended the use of the Clopper–Pearson interval (C–P), the mid-*P* interval, and the score interval with or without continuity correction. Recently, some authors adopted a new view based on the concept of 'mean coverage', in which strict conservatism is not a requirement. Newcombe [2] recommended the C–P interval in those cases when strict conservatism is needed, and otherwise the score and the mid-*P* intervals. Agresti and Coull [3] voted for using 'mean coverage' instead of the conventional minimum coverage and proposed

[*] Correspondence to: Jenő Reiczigel, Szent István University, Faculty of Veterinary Science, Department of Biomathematics and Informatics, Budapest, István u. 2., H-1078, Hungary.
[†] E-mail: jreiczig@univet.hu

Contract/grant sponsor: Hungarian Scientific Research Fund; contract/grant number: OTKA T035150.
Contract/grant sponsor: Hungarian Ministry of Education; contract/grant number: NKB-2001-KUT-5-018.

Received February 2001
Accepted April 2002

a new method, the so-called 'adjusted Wald interval', which they found superior to the score interval. These papers may suggest that exact methods, because of their conservatism, are less valuable than the asymptotic ones. Agresti and Coull [3] explicitly stated that exact methods might have an important role somewhere else in statistics but not in this area.

The aim of the present paper is to show that with an appropriate level adjustment exact methods can be applied in the mean coverage context and may be superior to asymptotic ones. To enable comparisons, computer programs were written for the Sterne interval and the above intervals that were found to be best in former comparisons. The programs determine also the coverage probabilities (see Section 3 for details).

2. STERNE'S EXACT INTERVAL

The interval proposed by Sterne [4] is defined by inverting the exact binomial test with acceptance regions including the most probable values of the binomial variable, taking the most probable, then the next most probable, until their total probability reaches the required level, for example, 95 per cent. (For the details of test inversion see Section 3.) Crow [5] noted that Sterne's procedure did not always result in a proper interval and corrected it. Additionally, he tried to improve it, but his modifications did not make the interval definitely better (for a summary, see Blyth and Still [6]). Blyth and Still, also starting from the Sterne interval, constructed their interval according to some further monotonically and smoothness conditions. Unfortunately, the Blyth and Still interval [6] is rather complicated to implement in a computer program. We propose using Sterne's original (but corrected) interval. It is easy to compute, strictly conservative but never too conservative. Its total length is near the possible minimum, and compared to the two-decimal tables in Blyth and Still's paper [6], there are only minor differences.

It should be emphasized that the Sterne interval is inherently two-sided, that is, it cannot be converted in the usual way into a one-sided interval. The reason for this is that a $(1-\alpha)$-level Sterne acceptance region may have the whole error probability α on one side, therefore one must not halve the error probability when looking at one side, for example, if taking the one-sided interval from 0 to the upper limit of a 90 per cent Sterne interval, it must be regarded again as a 90 per cent rather than a 95 per cent interval. This makes the use of the Sterne interval unreasonable if one-sided intervals are needed. In such cases the C–P interval offers a better solution.

Note that Sterne's method can be applied to a Poisson parameter too, and has the same advantages as in the binomial case (computer programs are available from the author).

3. CONFIDENCE INTERVAL CONSTRUCTION BY TEST INVERSION

For those not familiar with confidence interval construction methods, let us demonstrate through an example how confidence interval construction by test inversion works. This is the idea behind several confidence interval construction methods (for example, the C–P, the mid-P, the Sterne etc.). The methods differ just in the test to be inverted.

Suppose we want to invert a test of H_0: $p = p_0$ for the binomial parameter p to get a 90 per cent confidence interval for p based on, say, $n = 6$ observations. Denote by X the observed number of successes. The basic idea is that a 90 per cent confidence set should

Table I. The columns contain binomial probabilities ($n=6, p$ on the horizontal axis). In each column bold face probabilities indicate the 90 per cent acceptance regions of the Sterne test of H_0: $p=p_i$. The horizontal line indicates the 90 per cent Sterne confidence interval given $X=3$ is observed.

X	p_i										
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
6	0.0000	0.0000	0.0001	0.0007	0.0041	0.0156	0.0467	**0.1176**	**0.2621**	**0.5314**	**1**
5	0.0000	0.0001	0.0015	0.0102	0.0369	**0.0938**	**0.1866**	**0.3025**	**0.3932**	**0.3543**	0.0000
4	0.0000	0.0012	0.0154	0.0595	**0.1382**	**0.2344**	**0.3110**	**0.3241**	**0.2458**	**0.0984**	0.0000
3	0.0000	0.0146	0.0819	**0.1852**	**0.2765**	**0.3125**	**0.2765**	**0.1852**	0.0819	0.0146	0.0000
2	0.0000	**0.0984**	**0.2458**	**0.3241**	**0.3110**	**0.2344**	**0.1382**	0.0595	0.0154	0.0012	0.0000
1	0.0000	**0.3543**	**0.3932**	**0.3025**	**0.1866**	**0.0938**	0.0369	0.0102	0.0015	0.0001	0.0000
0	**1**	**0.5314**	**0.2621**	**0.1176**	0.0467	0.0156	0.0041	0.0007	0.0001	0.0000	0.0000

consists of all such values p_0 of the parameter for which H_0: $p=p_0$ is not rejected by the test at the 90 per cent level. Assume for simplicity that one-digit precision is sufficient for the interval endpoints, because in this case the procedure can be demonstrated using a small table of binomial probabilities (Table I). The most natural way to find the confidence interval (which is, however, computer intensive if higher precision is required) is to determine the acceptance regions of H_0: $p=p_i$ for each p_i ($p_i=0, 0.1, 0.2, \ldots, 0.9, 1$), and to take the first and the last p_i for which the acceptance region contains the observed X. Note that for some tests the endpoints of the acceptance regions do not increase monotonically with p. In such cases the set of those p_i values for which H_0: $p=p_i$ is not rejected at a given level does not form a proper interval, but it contains 'holes'. As Crow noticed, this may happen with the Sterne test. Of course, taking the smallest and largest p_i, as was suggested above, always results in a proper interval.

Bold face probabilities in Table I indicate the 90 per cent acceptance regions of the Sterne test, for example, for H_0: $p=0.2$, the acceptance region consists of the values $X=0,1,2$. Note that the type I error probability is always 1 minus the probability of the acceptance region (0.0989 in this particular case). Now suppose we have observed $X=3$. The 90 per cent confidence interval obtained from the acceptance regions of the Sterne test is (0.3 , 0.7), since for the p_i values in this interval, the hypothesis H_0: $p=p_i$ cannot be rejected at the 90 per cent level (Table I). The coverage probability given the true parameter is p_i is equal to the probability of the acceptance region belonging to p_i, which means that numerical estimates for minimum and mean coverage can be obtained immediately as side-products of the test inversion procedure. Let us calculate the probabilities of the acceptance regions for each p_i in Table I. For $p=0$ we get 1, for $p=0.1$ we get 0.9841 etc. As the smallest one is 0.9011 (for $p=0.2$), the estimated minimum coverage probability of the 90 per cent Sterne interval is 0.9011. Similarly, the estimated mean coverage is 0.9475. These estimates would be more precise if they were based on a finer division of the parameter space, that is, if using steps <0.1. In the present study, a step of 0.0001 was applied to provide sufficient accuracy for the coverage estimates.

Interval endpoints can be characterized as such parameter values, at which there is a change in the acceptance region. This may occur, for example, if the tail probability reaches the desired value. This happens in Table I between $p=0.2$ and $p=0.3$ (actually at about 0.2010), therefore the interval belonging to $X = 3$ starts here. Unlike methods handling the two tails independently (like, for example, the C–P), in the case of the Sterne interval there is a further reason for changes in the acceptance region, namely if the same probabilities appear at both ends. This is illustrated by $p = 0.5$ in Table I. Here, although the desired level is not reached, the acceptance region changes, as on the left of this point $P(X = 1) > P(X = 5)$, while on the right $P(X = 1) < P(X = 5)$ holds. Therefore the interval belonging to $X = 1$ ends, while that belonging to $X = 5$ starts here.

Some properties of the confidence intervals are easy to derive from the properties of the acceptance regions, for example, if the acceptance regions of one test are contained in those of another test for all p_i, then even the confidence intervals are contained in the intervals obtained from the other test. If all acceptance regions of one test are shorter than are those of another test, then the resulting confidence sets are also shorter, regarding total length. As the Sterne test uses the highest probabilities to construct the acceptance regions, it has the shortest possible acceptance regions, and so the Sterne confidence set has minimal total length, at least if it contains no 'holes'. If it does, and it is corrected by taking the smallest and largest p_i to form a proper interval, it loses this minimality property, but the increase is actually unobservably small. (Evaluating all 95 per cent intervals up to $n = 100$, the mean relative increase of total length was 0.04 per cent.)

4. OTHER INTERVALS TO BE COMPARED

For a short description of the intervals to be examined, let n, $\hat{p}$ and z denote the sample size, the observed probability and the $1 - \alpha/2$ critical value of the standard normal distribution. Let $b_{n,p}(.)$ and $B_{n,p}(.)$ denote the probability mass function and the CDF of the binomial distribution with parameter p.

1. Wilson's score interval [7] is defined by the formula

$$(2n\hat{p} + z^2 \pm z\sqrt{\{z^2 + 4n\hat{p}(1 - \hat{p})\}})/2(n + z^2)$$

The score interval is included in most textbooks. Its continuity corrected version [1] is not considered here because our experiences supported the findings of Agresti and Coull [3] that it performs similar to, if not a little worse than, the C–P interval.

2. The adjusted Wald interval has been proposed by Agresti and Coull [3]. It is calculated according to the Wald formula after adjusting the observed sample, that is

$$\tilde{p} \pm z\sqrt{\{\tilde{p}(1 - \tilde{p})/\tilde{n}\}}$$

where $\tilde{p} = (n\hat{p} + z^2/2)/(n + z^2)$ and $\tilde{n} = n + z^2$. Surprisingly, though it was derived as an approximation to the score interval, it has better coverage properties.

3. The Clopper–Pearson interval [8] is usually considered, in spite of some warnings [9], as the 'gold standard'. It is defined by inverting the exact binomial tests with equal-tailed acceptance regions (L_p, U_p), that is, where L_p is the largest value for which

$$B_{n,p}(L_p) \leqslant \alpha/2$$

and U_p is the smallest value for which

$$B_{n,p}(U_p+1) \geqslant 1-\alpha/2$$

The $(1-\alpha)$-level confidence interval consists of all p such as $L_p \leqslant \hat{p} \leqslant U_p$.

4. The mid-P interval [10] is defined similarly, but using slightly different (potentially narrower) acceptance regions (L'_p, U'_p). L'_p is the largest value for which

$$B_{n,p}(L'_p-1)+1/2b_{n,p}(L'_p-1) \leqslant \alpha/2$$

and U'_p is the smallest value for which

$$B_{n,p}(U'_p)+1/2b_{n,p}(U'_p) \geqslant 1-\alpha/2$$

Here too, the $(1-\alpha)$-level confidence interval consists of all p such as $L'_p \leqslant \hat{p} \leqslant U'_p$.

5. COMPUTER INTENSIVE LEVEL-ADJUSTMENT PROCEDURE

The C–P interval is often criticized for its conservatism. However, its conservatism is not because it is exact but because it aims to be symmetrical, that is, a 95 per cent interval aims to have 2.5 per cent error probability on each side, which, at worst, may actually result in a 97.5 per cent interval [11]. Other exact intervals releasing the condition of symmetry [4–6] are able to avoid conservatism even for small samples. Those insisting on symmetry and therefore choosing the C–P interval should keep in mind that even this method cannot guarantee symmetry, since the error probabilities are in fact not exactly 2.5 per cent, but less than that.

Conservatism is actually a discrepancy between what is expected and what is actually achieved (that is, the nominal and actual level). Thus the simplest way to avoid conservatism is determining and reporting the actual level rather than the nominal one. Even if doing so, the actual level may differ from the desired. To overcome this, a computer intensive level-adjustment procedure is proposed. The idea is very simple, namely varying the nominal level iteratively until the actual level gets close enough to the desired level. In the present implementation, the procedure finds the smallest nominal level with a precision of four decimal places for which the actual level is still above the desired one, for example, to have a C–P interval with an actual level of about 90 per cent for $n=5$, we should go down to the 83.59 per cent nominal level, so we obtain 90.88 per cent actual level. This is somewhat better than the 93.14 per cent that we get using the nominal 90 per cent interval. Another advantage of the procedure is that it can be applied in the same way for adjusting the mean coverage, which yields exact intervals in terms of mean coverage. Thus it is possible to construct a level-adjusted C–P interval for practically any prescribed mean coverage probability, for example, to get a C–P interval with 95 per cent mean coverage for $n=10$, one should go down to the 87.18 per cent nominal level. Note that this provides a minimum coverage of 88.95 per cent, still considerably higher than that of the 95 per cent score interval, which is 83.56 per cent.

The procedure can be applied in relation with other confidence interval construction methods as well, not exclusively with the C–P interval. Note that the Sterne interval needs level adjustment only in relation to the mean coverage.

6. RESULTS OF THE COMPARISONS

In the comparisons, both the minimum and the mean coverage criteria are considered. Any criterion is chosen, an interval is better than another if (i) its coverage is closer (possibly from above) to the desired value, and (ii) its fluctuation of coverage is less. The former can be measured simply as a distance; the latter however is more difficult to measure. In the present study, two measures are used for this purpose. One is the root mean square error of the coverage around the minimum or the mean coverage (depending on the criterion used), and the other is the proportion of the parameter space with coverage probability within given limits around the desired value [3], for example, for 95 per cent intervals, the range from 95 to 97 per cent is used in relation to minimum coverage and the range from 93 to 97 per cent in relation to mean coverage. If comparing various methods in terms of coverage fluctuation or interval width, it is desirable to compare intervals with the same actual rather than nominal coverage (either minimum or mean). To explain this, suppose that we are comparing a nominal 95 per cent C–P to a nominal 95 per cent score interval for $n=10$. It is not at all surprising that the score interval is shorter, because in terms of minimum coverage we are comparing a 96.1 per cent C–P to a 83.6 per cent score, or in terms of mean coverage, a 98.37 per cent C–P to a 95.41 per cent score. The level adjustment procedure is used here to provide that the intervals to be compared have the same actual level.

Choosing the minimum coverage criterion, only the C–P and the Sterne intervals were considered, of which obviously the Sterne interval was better in terms of minimum coverage as well as coverage fluctuation (Table II). Choosing the mean coverage criterion, the score, the mid-*P* and the adjusted Wald intervals were found to be the best in former comparisons [1–3]. Now they are compared to the mean coverage adjusted C–P and Sterne intervals. Owing to the mean coverage adjustment, the C–P and Sterne intervals are exact in terms of mean coverage, that is, in this respect they are superior to the asymptotic ones. For the comparisons, we adjusted the exact intervals to have the same mean coverage as the asymptotic interval, rather than to have the desired coverage. The reason for this is that fluctuation depends to some extent on the mean value, so it would not be fair to compare the coverage fluctuation of intervals with different mean coverage. Results show (Table III) that the Sterne interval has higher minimum coverage probability than the other intervals with the same mean coverage (also indicating that it shows less coverage fluctuation than those do). The mean coverage adjusted C–P interval has about the same minimum coverage as the mid-*P* and the adjusted

Table II. Comparison of the 95 per cent Clopper–Pearson (C–P) and the Sterne interval.

Sample size	Minimum coverage*	Root MSE of coverage (around the minimum coverage) C–P/Sterne	Proportion of the parameter space with $0.95 \leqslant \text{coverage} \leqslant 0.97$ C–P/Sterne
10	0.9610	0.0245/0.0201	0.078/0.201
15	0.9515	0.0300/0.0218	0.221/0.499
30	0.9505	0.0251/0.0162	0.396/0.803
50	0.9508	0.0209/0.0132	0.617/0.878
100	0.9503	0.0166/0.0105	0.832/0.946

*That of a nominal 95 per cent C–P interval. Sterne's interval was constructed to have the same minimum coverage.

Table III. Comparison of the nominal 95 per cent asymptotic intervals to the mean coverage adjusted Clopper–Pearson (C–P) and Sterne intervals.

(a) Score interval

Sample size	Mean coverage[*]	Minimum coverage Score/C–P/Sterne	Root MSE of coverage (around the mean coverage) Score/C–P/Sterne	Proportion of the parameter space with $0.93 \leqslant \text{coverage} \leqslant 0.97$ Score/C–P/Sterne	Total length Score/C–P/Sterne
10	0.9541	0.8356/0.8903/0.9243	0.0214/0.0245/0.0176	0.607/0.514/0.669	4.790/4.729/4.747
15	0.9534	0.8369/0.9078/0.9302	0.0183/0.0212/0.0161	0.777/0.671/0.839	5.899/5.825/5.839
30	0.9524	0.8373/0.9213/0.9344	0.0141/0.0173/0.0123	0.883/0.789/0.908	8.393/8.303/8.338
50	0.9518	0.8392/0.9246/0.9385	0.0114/0.0143/0.0097	0.940/0.870/0.941	10.860/10.760/10.813
100	0.9511	0.8435/0.9326/0.9425	0.0084/0.0111/0.0077	0.969/0.932/0.967	15.381/15.262/15.305

(b) Mid-*P* interval

Sample size	Mean coverage[†]	Minimum coverage Mid-*P*/C–P/Sterne	Root MSE of coverage (around the mean coverage) Mid-*P*/C–P/Sterne	Proportion of the parameter space with $0.93 \leqslant \text{coverage} \leqslant 0.97$ Mid-*P*/C–P/Sterne	Total length Mid-*P*/C–P/Sterne
10	0.9683	0.9266/0.9259/0.9455	0.0182/0.0177/0.0141	0.433/0.444/0.534	5.066/5.059/5.067
15	0.9637	0.9328/0.9321/0.9418	0.0164/0.0170/0.0122	0.638/0.634/0.724	6.128/6.113/6.147
30	0.9581	0.9245/0.9301/0.9411	0.0148/0.0154/0.0104	0.815/0.801/0.884	8.550/8.522/8.567
50	0.9553	0.9248/0.9307/0.9422	0.0124/0.0136/0.0090	0.878/0.855/0.930	10.975/10.931/10.979
100	0.9530	0.9207/0.9347/0.9439	0.0097/0.0107/0.0075	0.942/0.926/0.960	15.452/15.392/15.443

(c) Adjusted Wald interval

Sample size	Mean coverage[‡]	Minimum coverage Adjusted Wald/C–P/Sterne	Root MSE of coverage (around the mean coverage) Adjusted Wald/C–P/Sterne	Proportion of the parameter space with $0.93 \leqslant \text{coverage} \leqslant 0.97$ Adjusted Wald/C–P/Sterne	Total length Adjusted Wald/C–P/Sterne
10	0.9638	0.9168/0.9237/0.9398	0.0167/0.0190/0.0150	0.559/0.514/0.647	5.022/4.944/4.954
15	0.9624	0.9303/0.9166/0.9395	0.0148/0.0176/0.0125	0.699/0.650/0.751	6.145/6.073/6.121
30	0.9598	0.9339/0.9305/0.9439	0.0129/0.0148/0.0098	0.811/0.787/0.872	8.646/8.593/8.632
50	0.9578	0.9344/0.9336/0.9455	0.0117/0.0129/0.0085	0.861/0.836/0.920	11.100/11.061/11.105
100	0.9554	0.9389/0.9387/0.9462	0.0099/0.0104/0.0069	0.920/0.916/0.957	15.587/15.560/15.617

[*] That of a nominal 95 per cent score interval. The C–P and the Sterne intervals were constructed to have the same mean coverage.
[†] That of a nominal 95 per cent mid-*P* interval. The C–P and the Sterne intervals were constructed to have the same mean coverage.
[‡] That of a nominal 95 per cent adjusted Wald interval. The C–P and the Sterne intervals were constructed to have the same mean coverage.

Wald intervals. The score interval has considerably less minimum coverage than do all the others. Concerning coverage fluctuation, the asymptotic methods are better than the C–P while worse than the Sterne interval, according to both measures.

Total length of both the mean coverage adjusted C–P and Sterne intervals is a little less (≈ 0.4–1.3 per cent) less than that of the score interval for $n \leqslant 100$, given the mean coverage is the same. In the case of small samples ($n \leqslant 30$), they are also a little shorter (≈ 0.2–1.6 per cent) than the adjusted Wald interval. The total length of the mid-P interval is about the same as that of the mean coverage adjusted exact intervals (± 0.3 per cent). In fact these differences do not have much practical importance but show that much better coverage properties of the Sterne interval do not imply an increase in interval length. Using the score interval with the Blyth and Still limits for boundary outcomes [6], the minimum coverage increases to about 89 per cent but all other properties remain more or less the same.

For 90 per cent and 99 per cent intervals, results are similar except that the mean coverage of the 99 per cent score interval remains below the nominal level even for larger samples, that is, the 99 per cent score interval is anticonservative even in terms of mean coverage. (Its mean coverage was found to be 0.9875, 0.9888, 0.9895 for $n = 10, 30, 100$, respectively.)

7. DISCUSSION

The coverage probability of a confidence interval for a parameter is defined as the probability that the interval contains the parameter. Of course, this probability may depend on the true value of the parameter. Usually it is impossible to have the same coverage probability for all parameter values (like in the present case, due to the discreteness of the binomial distribution), which calls for a global measure. The traditional confidence coefficient (also called confidence level or minimum coverage) is the infimum, while the mean coverage [2, 3] is the average of the coverage probability over the whole parameter range (for the binomial parameter, over the $[0, 1]$ range). There is an important difference in the interpretation of these measures. The confidence coefficient can be interpreted as a lower bound of the coverage rate, guaranteed by the procedure. Mean coverage, however, being based on averaging over the whole parameter set, does not have any meaningful interpretation in relation to a single problem, just for a wide range of problems. It may have some special meaning for statisticians analysing lots of data sets with various true probabilities and aiming at a 'lifetime mean coverage' of 95 per cent, but could they really expect that all possible true probabilities occur uniformly in their practice? For example, epidemiologists studying the occurrence of rare diseases might protest against averaging over the whole [0,1] range. Mean coverage calls for the Bayesian framework where averaging is naturally made with respect to the prior distribution of the parameter, but in a frequentist context it is difficult to interpret. Nevertheless, it may be potentially useful in some situations.

Results show that the Sterne interval is superior in terms of minimum as well as mean coverage to all other intervals examined. This means that (i) its coverage probability is less fluctuating, (ii) in the case of equal mean coverage its minimum coverage is higher (the 'worst case risk' is less), and (iii) its total interval width is about the same or even a little less. For illustration, Figure 1 presents the coverage graphs of the score and the Sterne intervals. The Sterne interval exhibits a more regular, less fluctuating coverage function even for large samples. Note that it remains better than the score interval even if attention is restricted to

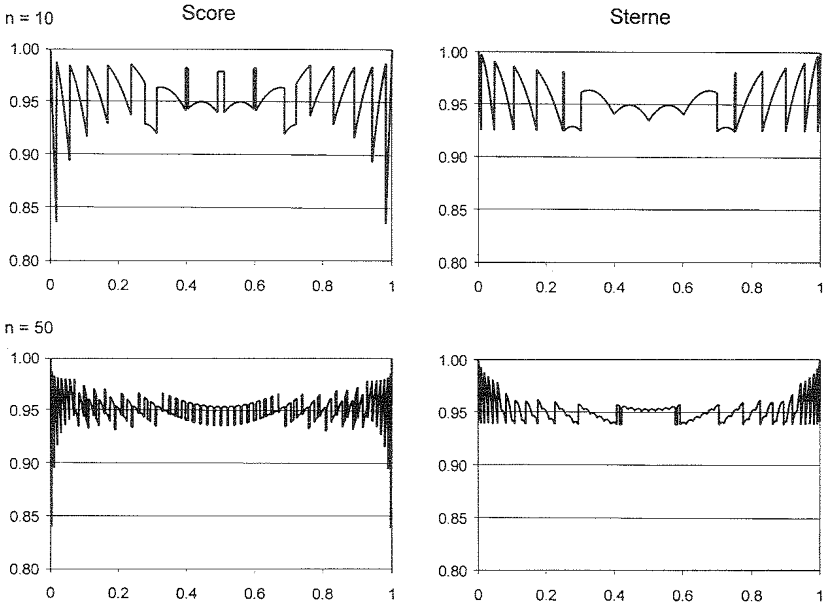

Figure 1. Coverage functions of the score and the mean coverage adjusted Sterne interval (true parameter value on the x-axis, coverage probability on the y-axis).

the parameter range of [0.1,0.9] where the score interval is regarded to perform acceptably (results not reported here).

Concerning the adjusted Wald interval, the present results support those reported by Agresti and Coull [3] about its superiority to the score interval. This is particularly important at the 99 per cent level where even the mean coverage of the score method is below the nominal 99 per cent.

We found that the level-adjusted C–P interval is practically equivalent to the mid-P interval (if adjusted to have the same confidence coefficient as the mid-P interval). This rather surprising thing means that the 'better performance' of the mid-P interval is purely due to its reduced confidence level, that is, the same performance can be reached by simply reducing the level of the C–P interval to the same extent. Thus, the only advantage of the mid-P interval is that the procedure does this level reduction automatically, so the users do not need to think about what they are willing to lose in terms of minimum coverage.

As our results suggested that the actual level of the mid-P interval does not converge to the nominal one, we determined its asymptotic minimum coverage, derived from the asymptotic probabilities of its acceptance regions (Section 3). Since for any λ the binomial probabilities with $p = \lambda/n$ converge to the Poisson probabilities with λ, binomial acceptance regions as well as their probabilities converge to those of the corresponding Poisson. Thus, these probabilities can be determined by applying the mid-P method to test the Poisson parameter (that is, H_0: $\lambda = \lambda_0$). By numerical evaluation of all acceptance regions of this test for $\lambda_0 = 0$ to 30 with steps of 0.001, minimum probabilities at the 90, 95 and 99 per cent level turned out to

Table IV. Ninety per cent acceptance regions of the Clopper–Pearson interval (solid line) and the Sterne interval (dotted line). For $X=0$, the Sterne interval is not contained in the C–P.

X	p			
	0.348	0.350	0.352	0.354
7	0.0006	0.0006	0.0007	0.0007
6	0.0081	0.0084	0.0086	0.0089
5	0.0456	0.0466	0.0477	0.0487
4	0.1423	0.1442	0.1462	0.1482
3	0.2666	0.2679	0.2692	0.2704
2	0.2997	0.2985	0.2973	0.2961
1	0.1871	0.1848	0.1824	0.1801
0	0.0501	0.0490	0.0480	0.0469

be 0.8574 (at $\lambda_0=4.275$), 0.9165 ($\lambda_0=3.061$) and 0.9861 ($\lambda_0=13.570$), respectively. This means that the mid-P interval fails to be asymptotically exact.

Blaker [12] has proposed a new exact interval recently, which is a very good alternative to the Sterne interval. The method has much common with the Sterne interval. It is based on inverting the exact test with acceptance regions including the most 'acceptable' values of the binomial variable, where acceptability of a value x is defined by the following function (called the 'acceptability function'):

$$\mathrm{ACC}(x)=\min\left\{\sum_{i=0}^{x} b_{n,p}(i), \sum_{j=x}^{n} b_{n,p}(j)\right\}$$

where $b_{n,p}(.)$ denotes the probability mass function of the binomial distribution with parameter p. To form the acceptance region, the most 'acceptable' values are taken until the desired level (for example, 95 per cent) is reached. This is clearly parallel to the definition of Sterne's acceptance region where the most probable values are taken in the same way. Blaker's interval is also inherently two-sided and its minimum coverage is always practically equal to the desired. Note that the definition of the acceptability function ensures that Blaker's interval is always contained in the C–P interval, which is not always true for the Sterne interval (see the example below). Otherwise the performance of the Blaker and the Sterne intervals is about the same; although the Sterne interval is typically (but not always) a little shorter and shows a little less coverage fluctuation, the difference is practically irrelevant.

The following example illustrates that the Sterne interval is not always contained in the C–P (Table IV). In this example, the upper endpoint of the Sterne interval belonging to $X=0$ goes beyond the corresponding C–P endpoint. It is clear that from $p=0.350$ up, since $P(X=0)\leqslant 0.05$, $X=0$ is not included in the C–P acceptance region. However, as the Sterne acceptance region consists of the highest probabilities, it includes $X=0$ rather than $X=5$ until $P(X=0)>P(X=5)$ holds, that is, up to $p=0.354$. That is, the Sterne test will not exclude an outcome with a higher probability in favour of another outcome with lower probability, whereas the C–P, handling the two tails independently of each other, may do this. In fact, this phenomenon is of theoretical importance only, as it occurs rarely and the difference is usually rather small.

Table V. Properties of the adjusted Wald and the Sterne interval for $n = 1000$.

Nominal level	Minimum coverage Sterne/adjusted Wald	Mean coverage (root MSE) Sterne/adjusted Wald	Total length Sterne/adjusted Wald
90%	0.9000/0.8832	0.9048 (0.0052)/0.9014 (0.0072)	41.225/40.907
95%	0.9500/0.9438	0.9527 (0.0030)/0.9511 (0.0041)	49.029/48.765
99%	0.9900/0.9890	0.9907 (0.0007)/0.9904 (0.0011)	64.337/64.153

8. CONCLUSION

Based on the results, we recommend using the Sterne interval (or the mean coverage adjusted Sterne interval if mean coverage is of interest) up to $n = 1000$ and the adjusted Wald interval beyond that. The Sterne interval can be replaced by Blaker's interval [12]. For one-sided intervals, the C–P method should be used up to $n = 1000$. Note that when switching from the exact to the asymptotic method, both the minimum and the mean coverage as well as the interval length becomes a little smaller (Table V).

ACKNOWLEDGEMENTS

This work was supported by the grants OTKA T035150 and NKB-2001-KUT-5-018. The author is indebted to three anonymous referees for their valuable suggestions.

REFERENCES

1. Vollset SE. Confidence intervals for a binomial proportion. *Statistics in Medicine* 1993; **12**:809–824.
2. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* 1998; **17**:857–872.
3. Agresti A, Coull BA. Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician* 1998; **52**:119–126.
4. Sterne TE. Some remarks on confidence or fiducial limits. *Biometrika* 1954; **41**:275–278.
5. Crow EL. Confidence intervals for a proportion. *Biometrika* 1956; **43**:423–435.
6. Blyth CR, Still HA. Binomial confidence intervals. *Journal of the American Statistical Association* 1983; **78**:108–116.
7. Wilson EB. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 1927; **22**:209–212.
8. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934; **26**:404–413.
9. Edwardes, M. Letter to the editor: 'Confidence intervals for a binomial proportion'. *Statistics in Medicine* 1994; **13**:1693–1698.
10. Lancaster HO. Significance tests in discrete distributions. *Journal of the American Statistical Association* 1961; **56**:223–234.
11. Angus JE, Schafer RE. Improved confidence statements for the binomial parameter. *The American Statistician* 1984; **38**:189–191.
12. Blaker, H. Confidence curves and improved exact confidence intervals for discrete distributions. *The Canadian Journal of Statistics* 2000; **28**:783–798.