2011

Were Clopper & Pearson (1934) too careful?

Frank Tuyl
University of Newcastle, frank.tuyl@newcastle.edu.au

Were Clopper & Pearson (1934) too careful?

Abstract

The ‘exact’ interval due to Clopper & Pearson (1934) is often considered to be the gold standard for estimating the binomial parameter. However, for practical purposes it is also often considered to be too conservative, when mean rather than minimum coverage close to nominal could be more appropriate. It is argued that (1) Clopper & Pearson themselves changed between these two criteria, (2) ‘approximate’ intervals are preferable to ‘exact’ intervals, and (3) approximate intervals are well represented by Bayesian intervals based on a uniform prior.

Keywords

Exact inference, confidence interval, binomial distribution, Jereys prior, Bayes-Laplace prior

Publication Details

Tuyl, Frank, Were Clopper & Pearson (1934) too careful?, Proceedings of the Fourth Annual ASEARC Conference, 17-18 February 2011, University of Western Sydney, Paramatta, Australia.

Were Clopper & Pearson (1934) too careful?

Frank Tuyl

The University of Newcastle, Australia
frank.tuyl@newcastle.edu.au

Abstract

The 'exact' interval due to Clopper & Pearson (1934) is often considered to be the gold standard for estimating the binomial parameter. However, for practical purposes it is also often considered to be too conservative, when mean rather than minimum coverage close to nominal could be more appropriate. It is argued that (1) Clopper & Pearson themselves changed between these two criteria, (2) 'approximate' intervals are preferable to 'exact' intervals, and (3) approximate intervals are well represented by Bayesian intervals based on a uniform prior.

Key words: Exact inference, confidence interval, binomial distribution, Jeffreys prior, Bayes-Laplace prior

1. Introduction

The 'gold standard' for estimating the binomial parameter is due to Clopper & Pearson (C&P) [1]. The $(1-\alpha)100\%$ C&P interval is based on the inversion of two separate hypothesis tests, the resulting lower limit θ_l being calculated from

$$\sum_{r=x}^{n} \binom{n}{r} \theta^r (1-\theta)^{n-r} = \alpha/2 \tag{1}$$

and the upper limit θ_u from

$$\sum_{r=0}^{x} \binom{n}{r} \theta^r (1-\theta)^{n-r} = \alpha/2. \tag{2}$$

Interestingly, using the relationship between binomial summations and beta integrals, the central Bayesian interval corresponding to a beta(a, b) prior follows from equating

$$I_\theta(x+a, n-x+b) = \sum_{r=x+a}^{n+a+b-1} \binom{n+a+b-1}{r} \theta^r (1-\theta)^{n+a+b-1-r} \tag{3}$$

to $\alpha/2$ to obtain the lower limit and doing the same with

$$1-I_\theta(x+a, n-x+b) = \sum_{r=0}^{x+a-1} \binom{n+a+b-1}{r} \theta^r (1-\theta)^{n+a+b-1-r} \tag{4}$$

to obtain the upper limit, where I_θ is the incomplete beta function. It follows that the C&P lower limit can be seen to correspond to a beta$(0, 1)$, and the C&P upper limit to a beta$(1, 0)$ prior. (Calculation of C&P intervals by using the inverse beta distribution, in Excel for example, is much more straightforward than using the inverse F distribution suggested by, among many others, Blaker [2].) This means that, compared with the Bayesian interval based on the (uniform) beta$(1, 1)$ or Bayes-Laplace (B-L) prior, the C&P lower limit is based on subtracting a success from the sample and the C&P upper limit on subtracting a failure. As a result, strictly speaking the C&P lower (upper) limit calculation breaks down when $x = 0\,(n)$ and is set to $0\,(1)$.

The effect of including x in the binomial summations (1) and (2) is that the C&P interval is 'exact': frequentist coverage, defined as $C(\theta) = \sum_{r=0}^{n} p(r|\theta) I(r, \theta)$, is at least equal to nominal for *any* value of θ, as illustrated in Figure 1. (Here $p(r|\theta)$ is the binomial pdf and $I(r, \theta)$ an indicator function: it is 1 when the interval corresponding to outcome r covers θ and 0 otherwise.) In fact, the mid-P interval [3, 4] is based on including *half* of $p(x|\theta)$ in (1) and (2), leading to an 'approximate' interval: this family of intervals aims for mean coverage to be close to nominal without compromising minimum coverage too much. The common Wald interval, based on the standard Normal approximation, is a poor example due to its serious below-nominal coverage, as also shown in Figure 1.

Similar to the Wald interval, the central B-L interval has zero minimum coverage (near the extremes). This is avoided by hybrid intervals that are one-sided for $x = 0\,(n)$, central otherwise [5], but a better interval is the one based on highest posterior density (HPD) and shown in Figure 1. In fact, all B-L intervals have *nominal* mean coverage, and the HPD interval performs well with respect to minimum coverage also.

Our discussion of this Bayesian interval is relevant as in Section 2 we point to an apparent contradiction in

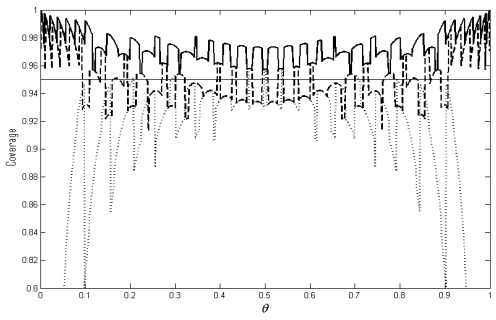

Figure 1: Coverage of the binomial parameter for $n = 30$ and $\alpha = 0.05$: Clopper & Pearson (solid), Bayes-Laplace HPD (dashed) and Wald (dotted) methods.

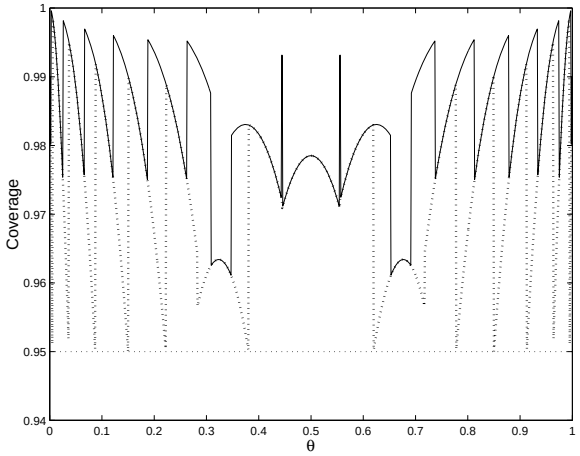

Figure 2: Coverage of the binomial parameter for $n = 10$ and $\alpha = 0.05$: Clopper & Pearson (solid) and Blaker (dotted) methods.

Clopper & Pearson's article and in Section 3 we argue that exact intervals may be seen to be more conservative than is usually shown in coverage graphs. In Section 4 we suggest the B-L HPD interval as an excellent representative of the 'approximate' family.

2. "Statistical experience"

It appears that C&P contradicted themselves with respect to which criterion, minimum or mean coverage, is more reasonable. On their first page (p.404), after a rather Bayesian reference to the *probability* of a parameter lying between two limits, they stated, "In our statistical experience it is likely that we shall meet many values of n and x; a rule must be laid down for determining θ_l and θ_u given n and x. Our confidence that θ lies within the interval (θ_l, θ_u) will depend upon the proportion of times that this prediction is correct in the long run of statistical experience, and this may be termed the confidence coefficient."

However, after showing graphically intervals for $n = 10$, C&P [1, p.406] changed the meaning of 'statistical experience': "It follows that in the long run of our statistical experience from whatever populations random samples of 10 are drawn, we may expect at least 95% of the points (x, θ) will lie inside the lozenge shaped belt, not more than 2.5% on or above the upper boundary and not more than 2.5% on or below the lower boundary." The addition of "at least" is understandable due to the discreteness of the Binomial distribution, but the "random samples of 10" phrase is crucial. We argue that in effect C&P and other exact intervals are based on the assumption, in addition to the hypothetical repeated sampling concept, that Nature is not only malicious, but omniscient as well: in effect, the exact method prepares for Nature choosing a true 'bad' value of θ based on knowledge of the sample size *and* level of confidence involved!

In fact, for the choice $n = 10$, C&P coverage is strictly above-nominal, which is improved elegantly (see Figure 2) by Blaker's method [2], based on considering confidence curves; due to its nesting property, the Blaker interval appears superior to other short exact intervals [6, 7, 8]. However, in the next section we argue that, from a practical point of view, *all* exact intervals are overly conservative.

3. "Approximate is better than exact"

The title of this section is a reference to Agresti & Coull (A&C) [9] who argued in favour of approximate intervals for most applications. We agree with their statement (p.125) that even though such intervals are technically not confidence intervals, "the operational performance of those methods is better than the exact interval in terms of how most practitioners interpret that term.". The implication is that, from a practical point of view, "narrower intervals for which the actual coverage probability could be less than .95 but is usually quite *close* to .95" are preferable.

Similarly, Brown, Cai & DasGupta (BCD) [5, p.113] considered the C&P approach "wastefully conservative and not a good choice for practical use, unless strict adherence to the prescription [coverage $\geq 1 - \alpha$] is demanded." It seems clear that A&C and BCD criticised the C&P interval because they consider mean coverage a better criterion than minimum coverage. We now show that even with respect to minimum coverage, C&P and other exact intervals are conservative.

A purist frequentist could claim that eventually a true physical constant, in the form of a proportion, could become known with substantial accuracy and that if this value were to be an "unlucky" one, exactness of previously calculated intervals would have been preferable. However, this argument is weakened if sample sizes are allowed to vary over time, as we now illustrate.

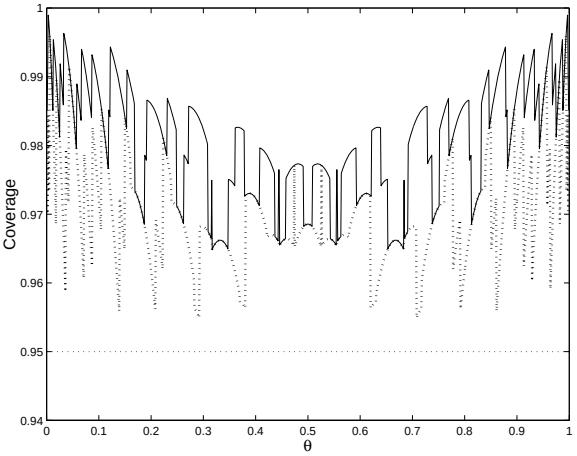

Figure 3: Coverage of the binomial parameter based on $\alpha = 0.05$ and averaging $n = 10$ and $n = 20$: Clopper & Pearson (solid) and Blaker (dotted) methods.

Supposing that the practitioner did always want to apply $\alpha = 0.05$, if in fact they considered a varying sample size (under repeated sampling), immediately Nature's scope for choosing 'bad' values of θ would be greatly reduced. Even short exact methods like Blaker's [2] turn strictly conservative as soon as repeated sampling takes place for two sample sizes (instead of one). Figure 3 is an example of this, based on adding $n = 20$ to the fixed $n = 10$ from Figure 2.

Thus, we do not even need "many values of n" to question the need for exact intervals! Finally, Figure 3 seems to lend even greater support to Stevens [10], who also argued against exact limits (p.121):

> It is the very basis of any theory of estimation, that the statistician shall be permitted to be wrong a certain proportion of times. Working within that permitted proportion, it is his job to find a pair of limits as narrow as he can possibly make them. If, however, when he presents us with his calculated limits, he says that his probability of being wrong is less than his permitted probability, we can only reply that his limits are unnecessarily wide and that he should narrow them until he is running the stipulated risk. Thus we reach the important, if at first sight paradoxical conclusion, that *it is the statistician's duty to be wrong* the stated proportion of times, and failure to reach this proportion is equivalent to using an inefficient in place of an efficient method of estimation.

In fact, Stevens showed that by taking a randomised weighted average of the two beta distributions implied by the C&P approach, the coverage probability, for *any* value of θ, equals the nominal confidence level. However, this approach leads to serious practical problems; for example, the resulting interval is typically asymmetric around $x = n/2$ if that is in fact the sample outcome, which is surely unacceptable from a practical point of view. About this interval, Lancaster [4] stated, "In certain experimental situations, this procedure would be time-consuming and even embarrassing to the statistician." It would thus appear that a 'regular' approximate interval with moderate below-nominal minimum coverage is preferable.

4. Which 'approximate' interval is best?

As stated in the introduction, intervals based on the B-L prior have nominal mean coverage, which, as far as we know, is not achieved by any 'approximate' intervals. The B-L HPD interval adds reasonable minimum coverage to this property. A more practical requirement dictates that no *individual* intervals are unreasonably short or wide, arguably satisfied by this likelihood-based interval also, but not necessarily by the approximate intervals recommended by BCD, for example.

It is unfortunate that review articles of approximate methods, such as the one by BCD, tend to consider, as representatives of the Bayesian approach, intervals based on the Jeffreys beta($\frac{1}{2}, \frac{1}{2}$) prior only. For example, minimum coverage of the Jeffreys HPD interval converges (as $n \to \infty$) to 84.0%, as opposed to 92.7% for the B-L HPD interval. Due to the Jeffreys prior's weight near the extremes, corresponding intervals appear particularly short when $x = 0\,(n)$, simultaneously causing this low minimum coverage. Even the hybrid Jeffreys interval recommended by BCD, despite their earlier criticism of methods that are data-based [5, p.106], has limiting minimum coverage of only 89.8%. These results follow quite simply from considering Poisson intervals.

A&C derived simple approximate intervals from the Score method due to Wilson [11]. These have excellent minimum coverage, at the expense of somewhat conservative mean coverage. (Note that, in contrast, the Score interval has mean coverage closer to nominal, but limiting minimum coverage of 83.8%.) However, individual Score and A&C intervals are undesirable when they are wider than corresponding C&P intervals: when this occurs, it would seem difficult to justify such an approximate interval to a client, based on the notion that the C&P interval is 'conservative'!

It is no surprise that Score-based intervals have this undesirable property, as the Normal approximation they are based on is inadequate when x is close to 0 or n. The mid-P interval appears to be a better choice, and, in fact, is quite similar to the B-L HPD interval for such x. However, as pointed out by A&C also, the mid-P interval is conservative (on average) for small n; for $n \leq 4$, for example, coverage is strictly above-nominal.

In short, we propose the B-L HPD interval as the preferred candidate for estimation of the binomial parameter, from both Bayesian and frequentist points

of view; see also [12] and [13]. About HPD intervals, BCD stated, "The psychological resistance among some to using this interval is because of the inability to compute the endpoints at ease without software.", but implementation is straightforward, even in Excel (using its Solver function). To emphasise, in effect, this probability-based interval is derived from the normalised likelihood function and has the desirable property that no values outside it have higher likelihood than the values inside it.

5. Conclusion

We conclude that based on their original intention, which was to allow for "many values of x and n", C&P's interval is too conservative: when allowing n to vary, coverage is strictly above-nominal for all values of θ. Short exact methods only address C&P conservativeness resulting from the dual one-sided testing aspect, and lead to similar behaviour. We consider this to be an additional argument against exact methods, which appear inconsistent with C&P's "long run of statistical experience".

We suggest that the B-L HPD interval, with its mean coverage equal to nominal and minimum coverage that would seem acceptable for most practical purposes, is preferable to the approximate intervals recommended by BCD, and should be adopted by Bayesians and frequentists alike.

References

[1] C. J. Clopper, E. S. Pearson, The use of confidence or fiducial limits illustrated in the case of the binomial, Biometrika 26 (1934) 404–416.

[2] H. Blaker, Confidence curves and improved exact confidence intervals for discrete distributions, The Canadian Journal of Statistics 28 (4) (2000) 783–798.

[3] H. O. Lancaster, The combination of probabilities arising from data in discrete distributions, Biometrika 36 (3/4) (1949) 370–382.

[4] H. O. Lancaster, Significance tests in discrete distributions, Journal of the American Statistical Association 56 (294) (1961) 223–234.

[5] L. D. Brown, T. Cai, A. DasGupta, Interval estimation for a binomial proportion, Statistical Science 16 (2) (2001) 101–133 (with discussion).

[6] E. L. Crow, Confidence intervals for a proportion, Biometrika 43 (1956) 423–435.

[7] C. R. Blyth, H. A. Still, Binomial confidence intervals, Journal of the American Statistical Association 78 (381) (1983) 108–116.

[8] G. Casella, Refining binomial confidence intervals, The Canadian Journal of Statistics 14 (1986) 113–129.

[9] A. Agresti, B. A. Coull, Approximate is better than "exact" for interval estimation of binomial proportions, The American Statistician 52 (2) (1998) 119–126.

[10] W. L. Stevens, Fiducial limits of the parameter of a discontinuous distribution, Biometrika 37 (1-2) (1950) 117–129.

[11] E. B. Wilson, Probable inference, the law of succession, and statistical inference, Journal of the American Statistical Association 22 (158) (1927) 209–212.

[12] F. Tuyl, R. Gerlach, K. Mengersen, A comparison of Bayes–Laplace, Jeffreys, and other priors: the case of zero events, The American Statistician 62 (1) (2008) 40–44.

[13] F. Tuyl, R. Gerlach, K. Mengersen, The Rule of Three, its variants and extensions, International Statistical Review 77 (2) (2009) 266–275.