2011

Were Clopper & Pearson (1934) too careful?

Frank Tuyl
University of Newcastle

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?

Abstract
The 'exact' interval due to Clopper & Pearson 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) in their article, Clopper & Pearson themselves changed between these two criteria, and (2) 'approximate' is indeed better than 'exact'.

Keywords
Exact inference, confidence interval, binomial distribution

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, Callaghan, NSW, 2308, AUSTRALIA

Abstract

The 'exact' interval due to Clopper & Pearson 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) in their article, Clopper & Pearson themselves changed between these two criteria, and (2) 'approximate' is indeed better than 'exact'.

Key words: Exact inference, confidence interval, binomial distribution

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 being calculated from

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

and the upper limit 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].) Put another way, compared with the Bayesian interval based on the (uniform) beta$(1,1)$ or Bayes-Laplace (B-L) prior, the C&P upper limit is based on subtracting a failure and the C&P lower limit on subtracting a success from the sample. 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 of the C&P interval is that it is 'exact': frequentist coverage, defined as $C(\theta) = \sum_{x=0}^{n} I(x,\theta)p(x|\theta)$, is at least equal to nominal for *any* value of θ, as shown in Figure 1. (Here $I(x,\theta)$ is the indicator function: it is 1 when the interval corresponding to outcome x covers θ and 0 otherwise.) In fact, the mid-P interval is based on including half of $p(x|\theta)$ in the two equations, leading to

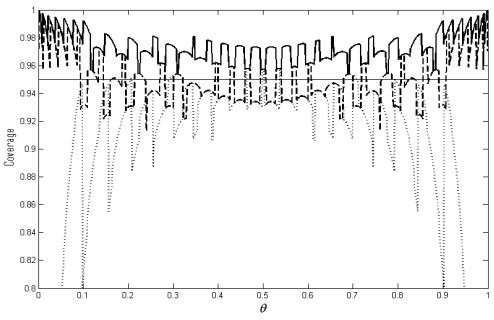

Figure 1: Coverage of the Binomial parameter for $n = 30$ and $\alpha = 0.05$: Clopper-Pearson (solid), Bayes-Laplace HPD (dashed) and Wald (dotted) methods.

an 'approximate' interval: this family of intervals aims for mean coverage to be close to nominal without compromising minimum coverage too much. The simplest example is the common Wald interval, based on the Normal approximation, which is quite unsatisfactory based on this criterion, as also shown in Figure 1.

Similar to the Wald interval, the central B-L interval has zero minimum coverage (near the extremes). This could be fixed by hybrid intervals that are one-sided for $x = 0\,(n)$, central otherwise [3], but a better interval is the one based on highest posterior density (HPD). 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 Clopper & Pearson's article, and in Section 3 we argue that exact intervals are altogether unnecessary.

2. "Statistical experience"

It appears that Clopper & Pearson [1] 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] seemed to change 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 by Blaker's method (see Figure 2). However, in the next section we argue that, from a practical point of view, there appears to be no need for *any* exact interval.

3. "Approximate is better than exact"

The title of this section is a reference to Agresti & Coull [4] 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

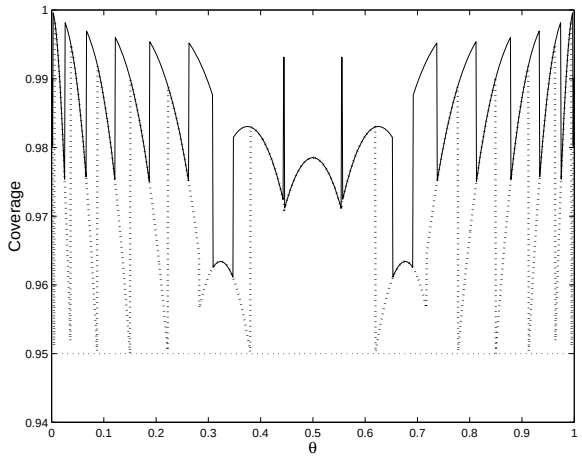

Figure 2: Coverage of the Binomial parameter for $n = 10$ and $\alpha = 0.05$: Clopper-Pearson (solid) and Blaker (dotted) methods.

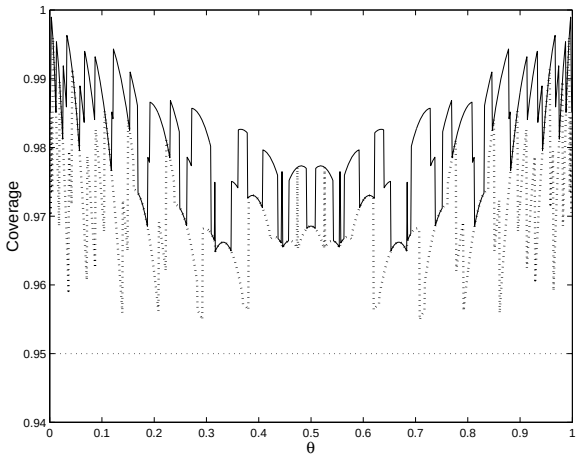

Figure 3: Coverage of the Binomial parameter based on $\alpha = 0.05$ and averaging $n = 10$ and $n = 20$: Clopper-Pearson (solid) and Blaker (dotted) methods.

term.", by which they meant that, from a practical point of view, it is better for a method to lead to "narrower intervals for which the actual coverage probability could be less than .95 but is usually quite *close* to .95". We now support this argument from another perspective.

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, an exact interval would still be preferable. However, this could only be the case if sample sizes had been kept constant over time, which we now illustrate.

Supposing that the practitioner did always want to apply $\alpha = 0.05$, if in fact they considered a varying sample size (under hypothetical 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 different 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 [5], 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 average coverage probability, for any value of θ, equals the nominal confidence level. His approach leads to serious practical problems, however, beyond the scope of this short note.

4. Conclusion

We conclude that based on their original intention, which was to allow for "many values of x and n", Clopper & Pearson'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 strictly above-nominal under varying n. Approximate methods seem more aligned with C&P's original wording, and from this family the Bayes-Laplace method based on HPD appears preferable, with its nominal mean coverage and minimum coverage that would seem acceptable for most practical purposes.

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

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

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