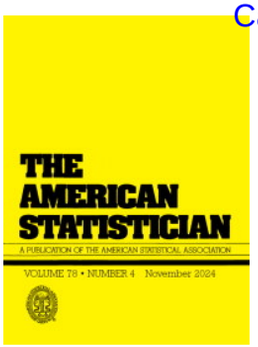
EXHIBIT A



The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: www.tandfonline.com/journals/utas20

Binomial Confidence Intervals for Rare Events: Importance of Defining Margin of Error Relative to Magnitude of Proportion

Owen McGrath & Kevin Burke

To cite this article: Owen McGrath & Kevin Burke (2024) Binomial Confidence Intervals for Rare Events: Importance of Defining Margin of Error Relative to Magnitude of Proportion, The American Statistician, 78:4, 437-449, DOI: 10.1080/00031305.2024.2350445

To link to this article: https://doi.org/10.1080/00031305.2024.2350445

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.

Published online: 24 May 2024.

Submit your article to this journal

Article views: 2381

View related articles

View Crossmark data

Citing articles: 2 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=utas20

THE AMERICAN STATISTICIAN
2024, VOL. 78, NO. 4, 437–449: Statistical Practice
https://doi.org/10.1080/00031305.2024.2350445

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Binomial Confidence Intervals for Rare Events: Importance of Defining Margin of Error Relative to Magnitude of Proportion

Owen McGrath and Kevin Burke

Department of Mathematics & Statistics, University of Limerick, Limerick, Ireland

ABSTRACT

Confidence interval performance is typically assessed in terms of two criteria: coverage probability and interval width (or margin of error). In this article, we assess the performance of four common proportion interval estimators: the Wald, Clopper-Pearson (exact), Wilson and Agresti-Coull, in the context of rare-event probabilities. We define the interval precision in terms of a relative margin of error which ensures consistency with the magnitude of the proportion. Thus, confidence interval estimators are assessed in terms of achieving a desired coverage probability whilst simultaneously satisfying the specified relative margin of error. We illustrate the importance of considering both coverage probability and relative margin of error when estimating rare-event proportions, and show that within this framework, all four interval estimators perform somewhat similarly for a given sample size and confidence level. We identify relative margin of error values that result in satisfactory coverage while being conservative in terms of sample size requirements, and hence suggest a range of values that can be adopted in practice. The proposed relative margin of error scheme is evaluated analytically, by simulation, and by application to a number of recent studies from the literature.

ARTICLE HISTORY
Received June 2023
Accepted April 2024

KEYWORDS
Binomial; Coverage; Confidence interval; Margin of error; Proportion; Rare event

1. Introduction

A fundamental problem in applied statistics is the construction of a confidence interval (CI) for a binomial proportion, p. In many applications, one deals with a large population within which an event of interest is rare. For example, in clinical statistics, p could represent the proportion of patients exhibiting treatment side effects; such a scenario arose in the context of COVID-19 vaccination (Polack et al. 2020). In manufacturing, the number of defective components is often very small relative to the large number of components produced. Indeed, many manufacturers now achieve a defect rate of 3.4 in one million (Evans and Lindsay 2015; Woodall and Montgomery 2014). In the aviation industry, strict regulations ensure that safety incidents are deemed as a rare occurrence, for example, Boeing (2022) shows that very few incidents occur within a large sample of flights. (Note: we revisit the COVID-19 and aviation examples as case studies in Section 8, along with an ADHD medication example.)

Ascertaining the order of magnitude of p is important in "large populations" such as the aforementioned. Indeed, with population sizes in the millions (or billions), there is a big practical difference between $p = 10^{-4}$ and $p = 10^{-6}$ (but such differences are much less important/detectable in smaller populations). For example, in high-throughput manufacturing, a difference in the order of magnitude in the failure rate has significant implications for the number of defects and/or product returns. From a purely pragmatic perspective, note that 10,000 observations are needed to obtain, on average, one event when $p = 10^{-4}$. However, while it is expected that relatively large samples will be required to adequately estimate the order of magnitude of a small proportion, p, practitioners will need more specific guidance on the sample size requirements; this is not well covered by the existing literature.

The problem of constructing a CI for p has a wide literature, including several comparative studies, for example, Gonçalves et al. (2012), Leemis and Trivedi (1996), Newcombe (1998), and Pires and Amado (2008). These works assess various proportion estimators, for example, Pires and Amado (2008) compare twenty different methods. However, these works, and the literature in general, focus primarily on situations where p is moderately large. As such, there is much less guidance in the existing literature regarding the scenario where p is small. Furthermore, there is little discussion of relative margin of error, which is needed in this small p setting. Whereas relative margin of error is not a prominent feature of CI assessment for moderately large proportions, it is essential that the margin of error scales with the magnitude of p for rare events. Therefore, we consider a valid CI estimator as one that achieves a desired coverage probability while also maintaining a specified relative margin of error, and, in contrast to much of the existing literature, we focus on the small p regime of $p \in [10^{-6}, 10^{-1}]$, where relative margin of error is especially important.

For our analysis, we consider the most widely used binomial confidence interval, the Wald interval, along with three other common intervals: Clopper-Pearson (exact), Wilson (score) and Agresti-Coull (adjusted Wald) (Wilson 1927; Clopper and Pear-

CONTACT Owen McGrath owen.mcgrath@ul.ie Department of Mathematics & Statistics, University of Limerick, Limerick, Ireland.
© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

son 1934; Agresti and Coull 1998). Despite its widespread use, the Wald interval is known to produce inadequate coverage when p is near 0 or 1, and/or the sample size, n, is small. It has also been well documented that this interval suffers from erratic coverage, even when p is moderate (Blyth and Still 1983; Vollset 1993; Böhning 1994; Agresti and Coull 1998). Brown, Cai, and DasGupta (2001) show that this coverage fluctuation also occurs for large n and recommend against using the Wald interval in practice. Newcombe (1998) also discourages the use of the Wald interval and suggests that its use be restricted to sample size planning. In recent work, Andersson (2023) discusses the deficiencies of the Wald interval and examines its coverage and noncoverage performance relative to the Wilson interval. Whilst the criticisms of the Wald interval can be justified, particularly when n is small, it is worth noting that the issue of erratic coverage is not unique to the Wald interval; this behavior is related to the binomial distribution and we illustrate (in Section 5) that it occurs for all four interval estimators.

A common approach in determining sample sizes is to set the (Wald) CI margin of error equal to a specified value, ϵ, and then solve for n. In order to maintain consistency between ϵ and p, we consider the relative margin of error, $\epsilon_R = \epsilon/p$, and obtain sample sizes by setting ϵ_R to a specified value and solving each interval equation for n. Lwanga and Lemeshow (1991) provide (Wald) sample size calculations for fixed margins of error and relative margins of error in the range $[0.01, 0.5]$ for $p \in [0.05, 0.95]$. However, in our work, we focus on the small-p regime of $p \in [10^{-6}, 10^{-1}]$ and provide computed coverage probabilities relating to $\epsilon_R \in [0.05, 0.75]$. In this regime, it is important to consider relative precision over fixed precision (fixed ϵ value). For example, $\epsilon = 0.1$ might be considered as reasonable precision for $p = 0.4$, but could equally be considered reasonable for $p = 0.2$. However, whereas $\epsilon = 0.05$ could be considered as a valid margin of error for $p = 10^{-1}$, it is far too large for a success probability of the order $p = 10^{-3}$. Ultimately, we find that $\epsilon_R \in [0.1, 0.5]$ yields a good compromise between estimation precision, coverage performance, and sample size requirements.

In this work, we illustrate the importance of using relative margin of error in a small p regime, recommend practical tolerances for both relative margin of error and coverage probability, and provide comparisons in terms of sample size requirements. We show that when CI performance is assessed in terms of both coverage probability and relative margin of error, the four CI estimators perform similarly in many cases. Although the differences between the estimators is less pronounced when considering the relative margin of error, we show that the Wilson (score) interval provides the best overall performance. In addition, we provide practical guidance on the sample sizes required to attain reasonable CI performance for various (small) values of p. We anticipate that this guidance will be useful to researchers working in real-world small-p applications.

The remainder of this article is organized as follows. In Section 2, we review some of the proportion estimators proposed in the literature, focusing, in particular, on the Wald, Clopper-Pearson, Wilson and Agresti-Coull intervals. Section 3 provides details of the CI evaluation criteria used in the work. Section 4 briefly discusses the initial estimation of p for sample size planning, then moves on to using relative margin of error in such planning, and CI performance evaluation. Sections 5 and 6 illustrate the importance of employing a relative margin of error scheme in CI assessment. Section 7 presents the challenge of estimating a rare-event proportion using a small sample size. In Section 8, we present a number of case studies to demonstrate the relative margin of error schemes in assessing the validity of estimated intervals. Finally, the article concludes in Section 9 with a discussion.

2. Binomial Proportion Interval Estimators

Several methods have been devised to estimate a binomial proportion, p, including the Wald, Clopper-Pearson, Wilson, Agresti-Coull, Jeffreys, arcsine transformation, Jeffreys' Prior and the likelihood ratio interval. A range of ensemble approaches have also been considered, for example, Kabaila, Welsh, and Abeysekera (2016), Park and Leemis (2019), and Turek and Fletcher (2012). In this work we assess the performance of the Wald, Clopper-Pearson, Wilson and Agresti-Coull intervals in standard form, that is, without modification or application of continuity correction.

2.1. Details of the Intervals Considered in Our Work

The Wald interval is included in this study as it is the most widely known and used estimator. We assess the Clopper and Pearson (1934) interval as it is an exact method, that is, unlike the other estimators used in this work, it is based directly on the cumulative probabilities of the binomial distribution. For this reason it is often regarded as the "gold standard" in binomial proportion estimation (Agresti and Coull 1998; Newcombe 1998; Gonçalves et al. 2012). Whereas the Wald interval is known to produce inadequate coverage for small n or p, the Clopper-Pearson interval is generally regarded as being overly conservative, unless n is quite large (Agresti and Coull 1998; Newcombe 1998; Brown, Cai, and DasGupta 2001; Thulin 2014). The intervals proposed by Wilson (1927) and Agresti and Coull (1998) both offer a compromise between the (liberal) Wald interval, and the (conservative) Clopper-Pearson interval.

We assess the performance of the Wilson and Agresti-Coull intervals in this work given their popularity within the literature. For example, Brown, Cai, and DasGupta (2001) recommend the Wilson or Jeffreys interval for small n. For larger n, they recommend the Wilson, Jeffreys or Agresti-Coull intervals, preferring the Agresti-Coull method for its simpler presentation. Agresti and Coull (1998) recommend the Wilson interval, and add that the 95% Wilson interval has similar performance to their method. Newcombe (1998) remarks on the *mid-p* method (an exact method closely related to the Clopper-Pearson method), and the Wilson method, noting the Wilson's advantage of having a simple closed form. Vollset (1993) too recommends the mid-p and Wilson (uncorrected and continuity corrected) intervals, along with the Clopper-Pearson interval, stating that those four intervals can be safely used at all times. Pires and Amado (2008) recommend the continuity corrected arcsine transformation or the Agresti-Coull method. Krishnamoorthy and Peng (2007) show that when controlling for Type I and Type II error

rates in one-sided hypotheses, the Wilson (score) and exact (Clopper-Pearson) tests require the same sample size. However, for two-sided hypothesis applications, and for constructing confidence intervals, they recommend the Wilson method.

Listed below are the Wald (W), Clopper-Pearson (CP), Wilson (WS) and Agresti-Coull (AC) interval formulas, where $z_{\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal distribution and x is the number of successes in a sample size of size n.

W: $\widehat{p} \pm z_{\alpha/2}\sqrt{\dfrac{\widehat{p}(1-\widehat{p})}{n}}$, where $\widehat{p} = x/n$.

CP: Lower: Beta $\left(\alpha/2; n\widehat{p}, n(1-\widehat{p})+1\right)$,
Upper: Beta $\left(1-\alpha/2; n\widehat{p}+1, n(1-\widehat{p})\right)$,
where Beta$(\cdot)$ is the quantile function of the Beta distribution.

WS: $$\frac{\widehat{p} + z^2_{\alpha/2}/2n \pm z_{\alpha/2}\sqrt{\widehat{p}(1-\widehat{p})/n + z^2_{\alpha/2}/4n^2}}{1 + z^2_{\alpha/2}/n}.$$

AC: $\widetilde{p} \pm z_{\alpha/2}\sqrt{\dfrac{\widetilde{p}(1-\widetilde{p})}{\widetilde{n}}}$, where $\widetilde{p} = \dfrac{n\widehat{p} + z^2_{\alpha/2}/2}{n + z^2_{\alpha/2}}$ and $\widetilde{n} = n + z^2_{\alpha/2}$.

2.2. Estimating Proportions when No Events are Observed

To provide an initial flavor of the challenging nature of estimating rare-event probabilities, we first consider the situation where no events are observed in the sample. Letting x denote the number of observed events, for very small p, it will be quite likely in small samples that $x = 0$, and, hence, $\widehat{p} = 0/n = 0$. For example, consider an event that occurs with probability $p = 10^{-3}$, and consider the case where the sample size is $n \leq 100$ such that $\Pr(X = 0) > 0.9$, that is, it is very likely that no events will be observed in this sample.

In this scenario where no events are observed, the resulting intervals are very conservative for small n. For example, with a sample of size $n = 100$, the following intervals are obtained: W: [0, 0], CP: [0, 0.0362], WS: [0, 0.0370], AC:[−0.0074, 0.0444]. Clearly the Wald interval is degenerate, and the remaining intervals are far too wide to be useful in small p settings where a good estimate of magnitude is required; the issue is more acute for smaller proportions, for example, the probability of observing no events is 0.99 if $p = 10^{-4}$ and 0.999 if $p = 10^{-5}$. As it is clear that a sample size of $n = 100$ will not suffice, further guidance is required in these scenarios, which we provide in the sequel. Of course, the required precision is analysis specific: the above intervals would be perfectly adequate if one was only interested in assessing if $p < 0.1$, for example. However, for situations where it is important to determine the order of magnitude, such as assessing the failure rate in high-volume manufacturing, then sufficiently large samples will be required to gain a reasonable estimate of the order of magnitude of p.

3. Evaluation Criteria

The most commonly used CI evaluation criteria are coverage probability and expected width (Gonçalves et al. 2012)—and these are the criteria that we consider in this work. However, other performance metrics have been proposed. For example, Vos and Hudson (2005) interpret an interval as the non-rejected parameter values in a hypothesis test and discuss the *p-confidence* and *p-bias* criteria; Newcombe (1998) presents a criterion using noncoverage as an indicator of location; and Park and Leemis (2019) adopt an ensemble approach and use root mean squared error and mean absolute deviation to measure CI performance.

3.1. Coverage Probability

The coverage probability can be interpreted as the computed interval's long-run percentage inclusion of the unknown parameter. Denoting L_x and U_x as the lower and upper CI bounds formed with x successes (suppressing the dependence on p and the significance level, α), the expected coverage probability, which we denote *CPr*, for a fixed parameter p, is given by

$$CPr(n,p) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} 1(L_x \leq p \leq U_x), \quad (1)$$

where $1(\cdot)$ is an indicator function that takes the value 1 when its argument is true, and 0 otherwise.

3.2. Expected Width

The expected width, which we denote *EW*, is given by

$$EW(n,p) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} (U_x - L_x), \quad (2)$$

and the expected margin of error, *EMoE*, is then given as $EMoE(n,p) = EW(n,p)/2$.

4. Calculating Sample Size

The first problem in CI estimation is determining the sample size required to achieve a desired estimation precision. There have been a range of sample size determination methods discussed in the literature, for example, Gonçalves et al. (2012), Korn (1986), and Liu and Bailey (2002), but here we adopt the common approach of deriving the sample size from the CI formula with fixed $\epsilon_R = \epsilon/p$. Used in conjunction with the Wald margin of error, one obtains

$$n = \left\lceil \frac{z^2_{\alpha/2}(1-p^*)}{\epsilon_R{}^2 p^*} \right\rceil, \quad (3)$$

where $\lceil\cdot\rceil$ denotes the ceiling function, and p^* denotes an anticipated value of p, that is, an initial estimate. (Sample size formulas for the Clopper-Pearson, Wilson and Agresti-Coull intervals are given in Appendix A.)

4.1. Initial Estimate of p

Selecting a value for p^* is required to make (3) operational, and this is an inherent practical challenge in any such sample size calculation. In some situations, it might be possible to overcome this problem by using subject matter knowledge or results from a previous study. If no previous information is available, a common approach is to consider the value of $p^* = 0.5$, but we do not adopt this approach here given that the focus of this work is on small/rare-event probabilities.

In some situations one might be able to gain a reasonable estimate of p. Consider a manufacturing environment where, for example, past experience or consultation with process experts could provide an analyst with information on the order of magnitude of p. One may be able to deduce that the true proportion is more likely to be of the order 1 in 10,000 rather than 1 in 1000. Such insight could be sufficient in setting a reasonable value for p^* and subsequently determining an appropriate sample size. Such initial estimation of an unknown parameter is a topic worthy of further discussion but is beyond the scope of this work. Here the focus is on the performance of the intervals after an initial estimate has been obtained.

4.2. Margin of Error Relative to p

The required precision is largely analysis dependent; what could be considered reasonable accuracy in one setting might be completely inappropriate in another. Although a relative margin of error scheme cannot be rigidly prescribed, we suggest a general scheme to avoid intervals that are too wide to be practically useful, or indeed intervals that are too narrow in the sense that a reasonable estimate could have been obtained using fewer resources.

To ensure that the margin of error is not larger than the order of magnitude of p^*, we impose $\epsilon_R \leq 1$. Thus, one could consider $0 \leq \epsilon_R \leq 1$ as a plausible margin of error scheme. However, considering ϵ_R values too close to the bound of 1 results in very wide intervals, whilst considering ϵ_R values too close to the bound of 0 results in very narrow intervals. As ϵ_R approaches zero, an increasingly large sample size is required, and such high precision is not likely to be required for many studies. Therefore, we suggest $\epsilon_R \in [0.1, 0.5]$ as a reasonable scheme. This scheme ensures that the interval is not impractically wide, nor excessively narrow (in terms of demanding very large sample sizes), and we show in Section 6 that acceptable coverage is achieved for this range of ϵ_R values.

A comparison of calculated sample sizes corresponding to $\epsilon_R = 0.4$ is provided in Table 1. (Sample size values in this and subsequent tables are rounded to 2 significant digits.) We can see from Table 1 that the Wald, Wilson and Agresti-Coull sample sizes are similar across the p^* range, but the Clopper-Pearson sample sizes are approximately 40% larger which is indicative of this method's conservatism. Note that Table 1 is only intended to provide an initial sense of the sample size requirements. However, final results can be found in Tables 10 and 11, and these account for the fact that empirical coverage for binomial proportion interval estimators is non-monotonic in the sample size (see Sections 5 and 6).

Table 1. Sample size comparison for $\epsilon_R = 0.4$.

p^*	ϵ	W	CP	WS	AC
10^{-1}	$4.0 \cdot 10^{-2}$	$2.2 \cdot 10^{2}$	$3.0 \cdot 10^{2}$	$2.2 \cdot 10^{2}$	$2.3 \cdot 10^{2}$
10^{-2}	$4.0 \cdot 10^{-3}$	$2.4 \cdot 10^{3}$	$3.3 \cdot 10^{3}$	$2.5 \cdot 10^{3}$	$2.5 \cdot 10^{3}$
10^{-3}	$4.0 \cdot 10^{-4}$	$2.4 \cdot 10^{4}$	$3.4 \cdot 10^{4}$	$2.5 \cdot 10^{4}$	$2.6 \cdot 10^{4}$
10^{-4}	$4.0 \cdot 10^{-5}$	$2.4 \cdot 10^{5}$	$3.4 \cdot 10^{5}$	$2.5 \cdot 10^{5}$	$2.6 \cdot 10^{5}$
10^{-5}	$4.0 \cdot 10^{-6}$	$2.4 \cdot 10^{6}$	$3.4 \cdot 10^{6}$	$2.5 \cdot 10^{6}$	$2.6 \cdot 10^{6}$
10^{-6}	$4.0 \cdot 10^{-7}$	$2.4 \cdot 10^{7}$	$3.4 \cdot 10^{7}$	$2.5 \cdot 10^{7}$	$2.6 \cdot 10^{7}$

Table 2. Wald-based sample size comparison – fixed ϵ.

p	Margin of error scheme			
	1	2	3	4
	$\epsilon = 4 \cdot 10^{-2}$ $p^* = p$	$\epsilon = 4 \cdot 10^{-2}$ $p^* = 0.5$	$\epsilon = 4 \cdot 10^{-4}$ $p^* = p$	$\epsilon = 4 \cdot 10^{-4}$ $p^* = 0.5$
10^{-1}	$2.2 \cdot 10^{2}$ (93.8)	$6.0 \cdot 10^{2}$ (94.5)	$2.2 \cdot 10^{6}$ (95.0)	$6.0 \cdot 10^{6}$ (95.0)
10^{-2}	$2.4 \cdot 10^{1}$ (21.4)	$6.0 \cdot 10^{2}$ (93.1)	$2.4 \cdot 10^{5}$ (95.0)	$6.0 \cdot 10^{6}$ (95.0)
10^{-3}	$3.0 \cdot 10^{0}$ (0.3)	$6.0 \cdot 10^{2}$ (45.2)	$2.4 \cdot 10^{4}$ (93.1)	$6.0 \cdot 10^{6}$ (95.0)
10^{-4}	$1.0 \cdot 10^{0}$ (0)	$6.0 \cdot 10^{2}$ (5.8)	$2.4 \cdot 10^{3}$ (21.3)	$6.0 \cdot 10^{6}$ (94.9)
10^{-5}	$1.0 \cdot 10^{0}$ (0)	$6.0 \cdot 10^{2}$ (0.6)	$2.4 \cdot 10^{2}$ (0.2)	$6.0 \cdot 10^{6}$ (94.9)
10^{-6}	$1.0 \cdot 10^{0}$ (0)	$6.0 \cdot 10^{2}$ (0.1)	$2.5 \cdot 10^{1}$ (0)	$6.0 \cdot 10^{6}$ (93.4)

Wald 95% CI coverage shown in parentheses.
Coverage computed with $p = p^*$

Given that ϵ_R is a function of p^*, the above calculated sample sizes are only applicable to each specific p^*. Even if the true proportion p equals p^* (i.e., the initial estimate is perfect), $\widehat{p}$ will of course vary from sample to sample and may not equal p; hence, the realized relative margin of error will typically differ from ϵ_R.

4.3. ϵ-p^* Compatibility

Next we illustrate the importance of defining the margin of error in relation to the magnitude of the proportion. Consider the following fixed margin of error schemes:

- Scheme 1: $\epsilon = 4 \cdot 10^{-2}$, $p^* = p$
- Scheme 2: $\epsilon = 4 \cdot 10^{-2}$, $p^* = 0.5$
- Scheme 3: $\epsilon = 4 \cdot 10^{-4}$, $p^* = p$
- Scheme 4: $\epsilon = 4 \cdot 10^{-4}$, $p^* = 0.5$

Table 2 displays the calculated Wald sample sizes and coverage probabilities corresponding to the above margin of error schemes. (A comparison of Wald, Clopper-Pearson, Wilson and Agresti-Coull coverage probabilities for the above ϵ schemes is given in Appendix B.) Referring to Table 2, fixing $\epsilon = 4 \cdot 10^{-2}$ and considering $p^* = p$ (*Scheme 1*) creates sample sizes that reduce dramatically as a function of p. This results in coverage probabilities that are completely inadequate for $p \leq 10^{-2}$. In *Scheme 2*, both ϵ and p^* are fixed and this creates a constant sample size of $n = 6 \cdot 10^{2}$. This sample size is reasonable for $p = 10^{-1}$, but is insufficient for the remaining p values, which is reflected in the poor coverage performance.

Table 3. Wald-based sample size comparison – variable ϵ.

p^*	ϵ_R 0.05	0.1	0.2	0.3	0.4	0.5	0.75
10^{-1}	$1.4 \cdot 10^4$ (94.9)	$3.5 \cdot 10^3$ (94.9)	$8.6 \cdot 10^2$ (95.0)	$3.8 \cdot 10^2$ (94.3)	$2.2 \cdot 10^2$ (93.8)	$1.4 \cdot 10^2$ (93.3)	$6.2 \cdot 10^1$ (94.7)
10^{-2}	$1.5 \cdot 10^5$ (94.9)	$3.8 \cdot 10^4$ (94.9)	$9.5 \cdot 10^3$ (95.2)	$4.2 \cdot 10^3$ (94.0)	$2.4 \cdot 10^3$ (95.2)	$1.5 \cdot 10^3$ (92.5)	$6.8 \cdot 10^2$ (90.2)
10^{-3}	$1.5 \cdot 10^6$ (94.9)	$3.8 \cdot 10^5$ (95.0)	$9.6 \cdot 10^4$ (95.0)	$4.3 \cdot 10^4$ (94.7)	$2.4 \cdot 10^4$ (93.1)	$1.5 \cdot 10^4$ (93.3)	$6.8 \cdot 10^3$ (90.4)
10^{-4}	$1.5 \cdot 10^7$ (95.0)	$3.8 \cdot 10^6$ (95.0)	$9.6 \cdot 10^5$ (95.0)	$4.3 \cdot 10^5$ (94.7)	$2.4 \cdot 10^5$ (93.1)	$1.5 \cdot 10^5$ (93.3)	$6.8 \cdot 10^4$ (90.4)
10^{-5}	$1.5 \cdot 10^8$ (95.0)	$3.8 \cdot 10^7$ (95.0)	$9.6 \cdot 10^6$ (95.0)	$4.3 \cdot 10^6$ (94.7)	$2.4 \cdot 10^6$ (93.1)	$1.5 \cdot 10^6$ (93.3)	$6.8 \cdot 10^5$ (90.4)
10^{-6}	$1.5 \cdot 10^9$ (95.0)	$3.8 \cdot 10^8$ (95.0)	$9.6 \cdot 10^7$ (95.0)	$4.3 \cdot 10^7$ (94.7)	$2.4 \cdot 10^7$ (93.1)	$1.5 \cdot 10^7$ (93.3)	$6.8 \cdot 10^6$ (90.4)

Wald 95% CI coverage shown in parentheses.

Scheme 3 is similar to *Scheme 1*, but here, ϵ is reduced to $4 \cdot 10^{-4}$. This ϵ-p combination produces sufficient coverage for $p \geq 10^{-2}$, but deteriorates for the smaller p values. In *Scheme 4*, p^* is fixed at 0.5 and $\epsilon = 4 \cdot 10^{-4}$, this results in a constant sample size of $n = 6 \cdot 10^6$, which produces good coverage throughout the p range, particularly for $p \geq 10^{-5}$. Whilst the coverage is satisfactory in this scheme, the magnitude of ϵ is not compatible with all p values, particularly $p = 10^{-1}$ and $p \leq 10^{-5}$. For $p = 10^{-1}$ the resulting interval is $[0.0996, 0.1004]$ which is too narrow in the sense that a reasonable interval could be obtained with a significantly reduced sample size. For $p = 10^{-6}$ the interval is truncated at $[0, 0.000401]$. Here, even though the coverage is reasonable, the interval is too wide to be practically useful since its width is two orders of magnitude larger than p.

The Wald sample sizes and coverage probabilities associated with the relative margin of error schemes: $\epsilon_R \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75\}$ are given in Table 3.

From Table 3, we see that by considering ϵ in relation to the magnitude of p, the coverage probabilities are reasonable across the p range, but now, the analyst must choose a scheme such that the resulting interval's width is appropriate. For example, consider $p^* = p = 10^{-1}$ and $\epsilon_R = 0.05$ where the resulting interval is $[0.095, 0.105]$. This interval is very narrow and the large sample size of $1.4 \cdot 10^4$ reflects this quite stringent margin of error. Moving to $\epsilon_R = 0.75$ has the advantage of significantly reducing the sample size, but, of course, the interval is significantly wider at $[0.025, 0.175]$. To obtain intervals that are neither too liberal nor too conservative, that are reasonable in terms of coverage performance, and which avoid excessively large sample sizes, we recommend $\epsilon_R \in [0.1, 0.5]$ as a reasonable scheme.

The coverage values for the Clopper-Pearson, Wilson and Agresti-Coull intervals are similar to those shown in Table 3 for $\epsilon_R \leq 0.5$. A comparison of coverage probabilities for $\epsilon_R = 0.75$ is given in Appendix B.

4.4. Suitability of ϵ_R Scheme

A range of qualifications/criteria are often used to check the validity of using approximate CI estimators. Fleiss, Levin, and Cho Paik (2003) state that the normal distribution provides excellent approximations to exact binomial procedures when $np \geq 5$ and $n(1-p) \geq 5$. Leemis and Trivedi (1996) also discuss the $np \geq 5$ (or 10) and $n(1-p) \geq 5$ (or 10) qualification.

Table 4. ϵ_R thresholds as per (4) for $a \in \{5, 10\}$.

p^*	$a = 5$ $\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$a = 10$ $\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
10^{-1}	0.698	0.832	1.093	0.493	0.588	0.773
10^{-2}	0.732	0.872	1.146	0.518	0.617	0.810
10^{-3}	0.735	0.876	1.151	0.520	0.619	0.814
10^{-4}	0.736	0.876	1.152	0.520	0.620	0.815
10^{-5}	0.736	0.877	1.152	0.520	0.620	0.815
10^{-6}	0.736	0.877	1.152	0.520	0.620	0.815

We examine the proposed ϵ_R scheme to assess its compatibility with the qualification $np^* \geq a$ and $n(1-p^*) \geq a$, where $a \in \{5, 10\}$, in relation to the Wald sample size equation. Multiplying (3) by p^* gives

$$np^* = \frac{z^2_{\alpha/2}(1-p^*)}{{\epsilon_R}^2} \geq a \implies \epsilon_R \leq \sqrt{\frac{z^2_{\alpha/2}(1-p^*)}{a}}. \quad (4)$$

For a given n, $np^* < n(1-p^*)$ when $p^* < 0.5$ and, therefore, (4) is sufficient in the small-p regime to ensure both np^* and $n(1-p^*)$ are greater than a. An evaluation of ϵ_R for $p^* \in [10^{-6}, 10^{-1}]$ and $\alpha \in \{0.1, 0.05, 0.01\}$ is provided in Table 4, which shows that our suggested relative margin of error scheme, $\epsilon_R \in [0.1, 0.5]$, lies below the threshold of (4). (The case of $\epsilon_R = 0.5$ negligibly exceeds the threshold of 0.493 for $p^* = 10^{-1}, \alpha = 0.1$.)

4.5. Tolerances for Assessing CI Performance

In this section we suggest suitable tolerances for assessing interval performance in terms of coverage probability and relative margin of error. In relation to achieving a desired coverage probability, one usually considers $(1-\alpha)100 \pm \epsilon^*\%$, where ϵ^* denotes a predefined coverage tolerance. The definition of such a tolerance is dependent on the individual researcher and particular study, and is thus difficult to quantify. In one study $(1-\alpha)100 \pm 4\%$ might be acceptable, whilst in another, one might require $(1-\alpha)100 \pm 0.5\%$. We suggest that $\epsilon^* \in \{1, 2, 3\}$ would be reasonable tolerances for most analyses, and as such, consider acceptable expected coverage probabilities as $CPr \in (1-\alpha)100 \pm 3\%$, where CPr is described in (1).

A tolerance is also necessary with regard to the relative margin of error. As with the coverage, the desired margin of error is dependent on the particular research question and hence can not be rigidly prescribed. However, as previously discussed, it is important that the magnitude of the margin of error reflect the magnitude of the estimated proportion.

Table 5 provides suggested tolerances for the assessment of ϵ_R and CPr for $(1-\alpha)100\%$ confidence intervals which could be considered reasonable in most settings.

5. Relative Margin of Error Central to Performance

Next we illustrate how the relative margin of error is fundamental to CI performance evaluation. We show that when a valid confidence interval is defined as achieving a desired coverage probability while simultaneously satisfying a minimum relative margin of error, the four interval estimators perform similarly for a given n-p^*-α combination.

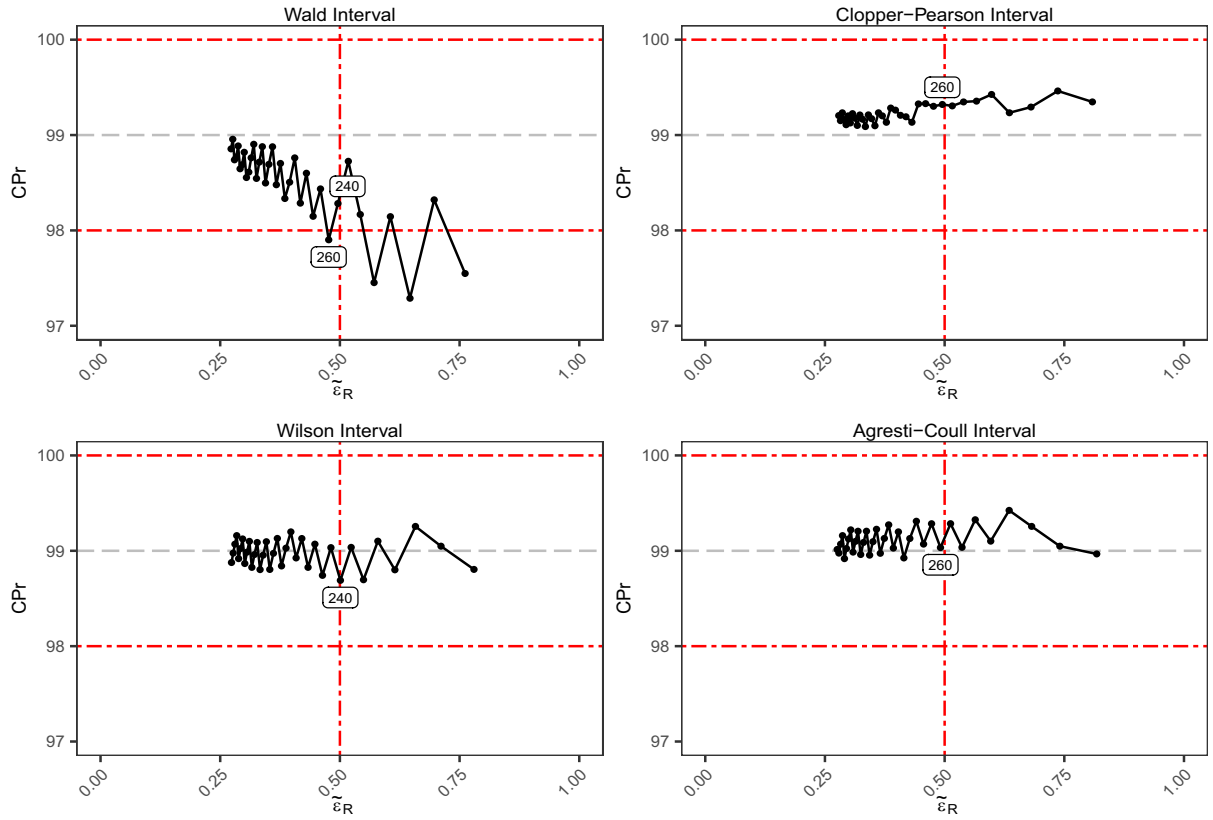


Figure 1. *CPr* versus $\tilde{\epsilon}_R$ for $p^* = p = 10^{-1}, \alpha = 0.01$. Dashed (gray) line represents the nominal *CPr* value. Dot-dashed (red) lines represent the target *CPr* and $\tilde{\epsilon}_R$ tolerances from Table 5. Sample size range is from $n = 100$, to $n = 800$, in steps of 20. Labels shown adjacent to data points depict the first n where both *CPr* and $\tilde{\epsilon}_R$ requirements are satisfied. Additional data labels are referred to within the text.

Table 5. Coverage & relative margin of error tolerances.

Tolerance	Coverage	Relative MoE
Target	$CPr \in (1-\alpha)100 \pm 1\%$	$\epsilon_R \leq 0.5$
Acceptable	$CPr \in (1-\alpha)100 \pm 2\%$	$\epsilon_R \leq 0.75$*
Minimally acceptable	$CPr \in (1-\alpha)100 \pm 3\%$	$\epsilon_R \leq 1$
Unacceptable	$CPr \notin (1-\alpha)100 \pm 3\%$	$\epsilon_R > 1$

*Considered *acceptable* as a compromise between $\epsilon_R \leq 0.5$ (desired) and $\epsilon_R \leq 1$ (limit).

To demonstrate CI performance we consider the *expected* relative margin of error, which we define as $\tilde{\epsilon}_R = EMoE/p$, where *EMoE* is half of the expected width (see (2)). As discussed in Section 4.2, a relative margin of error exceeding 1 is not acceptable from a practical perspective, and, as per Section 4.4, we suggest that it should not exceed 0.5.

Figure 1 provides a 99% CI performance comparison for $p^* = p = 10^{-1}$, and shows that the Wilson, Agresti-Coull and Clopper-Pearson intervals all achieve satisfactory coverage across the sample size range, whereas for $n \leq 260$, the coverage of the Wald interval oscillates around the lower limit of 98%. For example, the coverage is satisfactory at $n = 240$, but then falls below 98% at $n = 260$. This phenomenon of coverage oscillation relates to the discreteness of the binomial distribution and has been previously discussed in the literature, for example, Agresti and Coull (1998), Andersson (2023), Blyth and Still (1983), Brown, Cai, and DasGupta (2001), and Vollset (1993). For a given p, the empirical coverage does converge to the $(1-\alpha)100\%$ level with n as one would expect, but it does so in an oscillatory fashion for neighboring values of n. Figures 1 and 2 show that all four estimators suffer from this erratic behavior.

While the coverage performance of the Wald interval is inferior to the other three intervals for $n < 240$, none of the intervals satisfy the $\tilde{\epsilon}_R \leq 0.5$ requirement at these lower sample sizes. Thus, by stipulating a minimum requirement for $\tilde{\epsilon}_R$, the poor coverage at small n is rendered irrelevant and the performance of the Wald interval is more comparable to the other three intervals when $\tilde{\epsilon}_R \leq 0.5$.

A comparison of the performance of a 95% CI for $p^* = p = 10^{-2}$ is given in Figure 2 which shows that the Wald interval achieves $\tilde{\epsilon}_R \leq 0.5$ for $n \geq 1600$. For $n \geq 1600$ the Wald interval encounters five sample sizes where the coverage drops below the lower limit of 94%. The Clopper-Pearson interval requires a sample size of $n = 2000$ to satisfy both $\tilde{\epsilon}_R \leq 0.5$ and $CPr \in [94\%, 96\%]$, with the coverage exceeding the upper limit of 96% on seven occasions for $n > 2000$. The Agresti-Coull interval performs very well for $n \geq 1800$, with just one value ($n = 2600$), failing to satisfy both *CPr* and $\tilde{\epsilon}_R$ thereafter. The Wilson interval provides the best performance, satisfying both *CPr* and $\tilde{\epsilon}_R$ requirements $\forall\ n \geq 1600$.

Figures 1 and 2 highlight the similarities in performance when one considers $\tilde{\epsilon}_R$. In general, moderate-to-large sample sizes are required to satisfy both *CPr* and $\tilde{\epsilon}_R$ criteria, and at these sample sizes the performance across the four intervals is reasonably comparable.

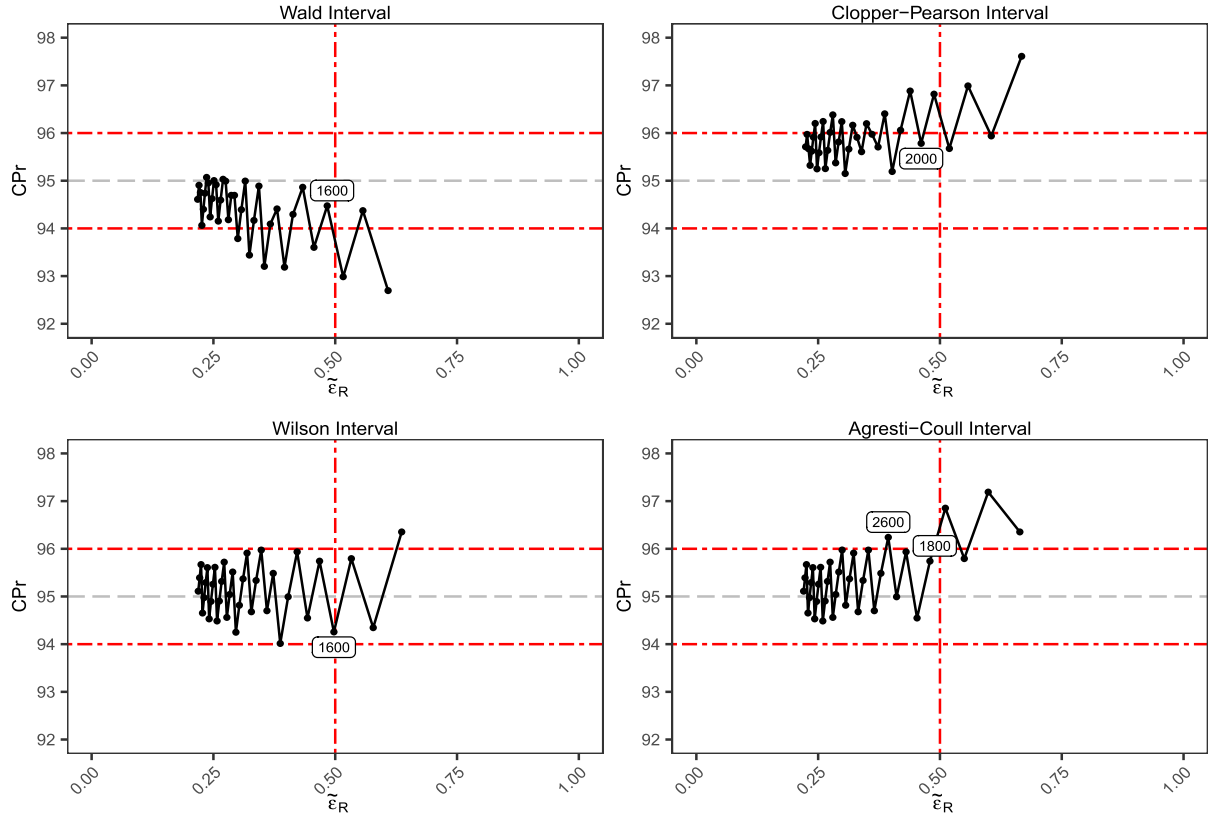


Figure 2. *CPr* versus $\tilde{\epsilon}_R$ for $p^* = p = 10^{-2}$, $\alpha = 0.05$. Dashed (gray) line represents the nominal *CPr* value. Dot-dashed (red) lines represent the target *CPr* and $\tilde{\epsilon}_R$ tolerances from Table 5. Sample size range is from $n = 1000$, to $n = 8000$, in steps of 200. Labels shown adjacent to data points depict the first n where both *CPr* and $\tilde{\epsilon}_R$ requirements are satisfied. Additional data labels are referred to within the text.

6. CI Performance Tables

Tables 6–9 provide a 95% CI comparison for $p^* = p = 10^{-1}$ and $p^* = p = 10^{-6}$, across a range of sample sizes and further illustrate the performance similarities among the estimators. The table cells are color coded according to the tolerances discussed in Table 5: target (green), acceptable (yellow), minimally acceptable (orange), and unacceptable (red).

We first consider $p^* = p = 10^{-1}$ and $10 \le n \le 140$, with Table 6 showing that none of the intervals satisfy the desired *CPr* and $\tilde{\epsilon}_R$ requirements simultaneously. The importance of considering the relative margin of error in CI evaluation is clearly evident. In several cases the coverage probability lies within the desired tolerance but the excessive relative margin of error renders the estimate impractical. For example, referring to the Wilson interval, $CPr(20, 10^{-1}) = 95.7\%$, however, $\tilde{\epsilon}_R = 1.32$ which is not acceptable.

We see from Table 7 that by increasing n, $\tilde{\epsilon}_R \le 0.5$ is satisfied (at $n = 150$, the Clopper-Pearson interval marginally exceeds 0.5, but is less than 0.5 thereafter). Each interval encounters sample sizes where *CPr* exceeds the bounds of $\pm 1\%$, but overall, *CPr* is satisfactory.

As per Table 8, none of the intervals satisfy $CPr \in [94\%, 96\%]$ and $\tilde{\epsilon}_R \le 0.5$ for $p^* = p = 10^{-6}$ and $n \le 14 \cdot 10^6$. The Wilson method performs best in this scheme, and if the tolerances of $CPr \in [93\%, 97\%]$ and $\tilde{\epsilon}_R \le 0.75$ were considered, it would produce a valid interval $\forall\ n$.

Referring to Table 9, for $15 \cdot 10^6 \le n \le 25 \cdot 10^6$ and $p^* = p = 10^{-6}$, the Wald interval has the worst coverage, with four *CPr* values exceeding $95 \pm 2\%$. The Wilson and Agresti-Coull intervals perform the best, but overall, all four intervals perform well in this large sample size scheme, particularly if the coverage tolerance was considered as $CPr \in 95 \pm 2\%$.

The *CPr* and $\tilde{\epsilon}_R$ performance across the p^* range is given in Tables 10 and 11. Shown is the performance of each CI estimator at a selection of sample sizes of interest. The Wald, Wilson and Agresti-Coull methods perform similarly for a given n-p^* combination, as shown in Table 10. The *CPr* and $\tilde{\epsilon}_R$ values of the Clopper-Pearson slightly exceed the desired limits, but overall, the performance is quite reasonable.

Table 6. 95% CI performance – $p^* = p = 10^{-1}$, "small" n.

n	W		CP		WS		AC	
	CPr	$\tilde{\epsilon}_R$	*CPr*	$\tilde{\epsilon}_R$	*CPr*	$\tilde{\epsilon}_R$	*CPr*	$\tilde{\epsilon}_R$
10	65.0	1.40	98.7	2.12	93.0	1.85	93.0	2.11
20	87.6	1.17	98.9	1.46	95.7	1.32	95.7	1.46
30	80.9	1.01	99.2	1.18	97.4	1.07	97.4	1.16
40	91.4	0.89	97.0	1.01	94.3	0.93	95.8	0.99
50	87.9	0.81	97.0	0.90	97.0	0.83	97.0	0.88
60	94.1	0.74	97.2	0.82	95.2	0.76	95.2	0.80
70	92.0	0.69	95.5	0.76	93.2	0.70	97.4	0.73
80	90.0	0.65	97.7	0.71	96.3	0.66	96.3	0.68
90	94.6	0.61	96.7	0.66	95.0	0.62	95.0	0.64
100	93.2	0.58	95.6	0.63	93.6	0.59	97.2	0.61
110	91.9	0.55	97.6	0.60	96.3	0.56	96.3	0.58
120	95.4	0.53	96.8	0.57	95.4	0.54	95.4	0.55
130	94.4	0.51	96.0	0.55	94.5	0.52	94.5	0.53
140	93.5	0.49	96.7	0.53	93.5	0.50	96.7	0.51

Table 7. 95% CI performance – $p^* = p = 10^{-1}$, "large" n.

n	W		CP		WS		AC	
	CPr	$\tilde{\epsilon}_R$	*CPr*	$\tilde{\epsilon}_R$	*CPr*	$\tilde{\epsilon}_R$	*CPr*	$\tilde{\epsilon}_R$
150	92.6	0.48	97.2	0.51	96.0	0.48	96.0	0.49
160	95.5	0.46	96.6	0.49	95.4	0.47	95.4	0.47
170	94.8	0.45	96.0	0.48	94.6	0.45	94.6	0.46
180	94.1	0.43	95.4	0.46	93.9	0.44	96.6	0.45
190	93.4	0.42	96.1	0.45	96.1	0.43	96.1	0.43
200	92.7	0.41	96.7	0.44	95.6	0.42	95.6	0.42
210	95.3	0.40	96.2	0.43	95.1	0.41	95.1	0.41
220	94.8	0.39	95.7	0.42	94.5	0.40	94.5	0.40
230	94.3	0.39	96.4	0.41	94.0	0.39	96.4	0.39
240	93.7	0.38	96.0	0.40	96.0	0.38	96.0	0.38
250	93.2	0.37	96.6	0.39	95.6	0.37	95.6	0.38
260	95.5	0.36	96.2	0.38	95.1	0.36	95.1	0.37
270	95.0	0.36	95.8	0.37	94.7	0.36	94.7	0.36
280	94.6	0.35	95.4	0.37	94.3	0.35	94.3	0.36
290	94.2	0.34	96.1	0.36	96.1	0.35	96.1	0.35
300	93.8	0.34	96.6	0.35	95.7	0.34	95.7	0.34
310	93.3	0.33	96.3	0.35	95.4	0.33	95.4	0.34
320	95.4	0.33	96.0	0.34	95.0	0.33	95.0	0.33
330	95.1	0.32	95.7	0.34	94.7	0.32	94.7	0.33
340	94.7	0.32	95.4	0.33	94.3	0.32	94.3	0.32
350	94.4	0.31	96.0	0.33	96.0	0.31	96.0	0.32

Table 8. 95% CI performance – $p^* = p = 10^{-6}$, "small" n.

n	W		CP		WS		AC	
	CPr	$\tilde{\epsilon}_R$	*CPr*	$\tilde{\epsilon}_R$	*CPr*	$\tilde{\epsilon}_R$	*CPr*	$\tilde{\epsilon}_R$
$7.5 \cdot 10^6$	93.6	0.70	95.8	0.79	93.7	0.75	97.4	0.79
$8.0 \cdot 10^6$	89.2	0.68	96.9	0.76	95.2	0.72	95.2	0.76
$8.5 \cdot 10^6$	91.9	0.66	97.7	0.73	96.3	0.70	96.3	0.74
$9.0 \cdot 10^6$	94.0	0.64	95.7	0.71	93.7	0.68	97.2	0.71
$9.5 \cdot 10^6$	90.3	0.63	96.7	0.69	95.2	0.66	95.2	0.69
$10.0 \cdot 10^6$	92.6	0.61	97.5	0.67	96.3	0.64	96.3	0.67
$10.5 \cdot 10^6$	94.4	0.60	95.7	0.65	93.9	0.63	97.1	0.65
$11.0 \cdot 10^6$	91.2	0.58	96.7	0.64	95.3	0.61	95.3	0.64
$11.5 \cdot 10^6$	93.2	0.57	97.5	0.62	96.3	0.60	96.3	0.62
$12.0 \cdot 10^6$	94.3	0.56	95.8	0.61	94.2	0.58	97.1	0.60
$12.5 \cdot 10^6$	92.1	0.55	96.8	0.60	95.5	0.57	95.5	0.59
$13.0 \cdot 10^6$	93.8	0.54	97.5	0.58	96.4	0.56	96.4	0.58
$13.5 \cdot 10^6$	94.7	0.53	96.0	0.57	94.6	0.55	94.6	0.57
$14.0 \cdot 10^6$	92.9	0.52	96.9	0.56	95.7	0.54	95.7	0.55

Table 9. 95% CI performance – $p^* = p = 10^{-6}$, "large" n.

n	W		CP		WS		AC	
	CPr	$\tilde{\epsilon}_R$	*CPr*	$\tilde{\epsilon}_R$	*CPr*	$\tilde{\epsilon}_R$	*CPr*	$\tilde{\epsilon}_R$
$15.0 \cdot 10^6$	91.9	0.50	96.3	0.54	94.9	0.52	94.9	0.53
$15.5 \cdot 10^6$	93.6	0.49	97.1	0.53	96.0	0.51	96.0	0.52
$16.0 \cdot 10^6$	94.4	0.49	95.6	0.52	94.1	0.50	96.8	0.52
$16.5 \cdot 10^6$	92.7	0.48	96.5	0.51	95.3	0.49	95.3	0.51
$17.0 \cdot 10^6$	94.2	0.47	97.2	0.51	96.2	0.49	96.2	0.50
$17.5 \cdot 10^6$	94.9	0.47	95.9	0.50	94.6	0.48	94.6	0.49
$18.0 \cdot 10^6$	93.5	0.46	96.7	0.49	95.6	0.47	95.6	0.48
$18.5 \cdot 10^6$	94.8	0.45	95.3	0.48	93.9	0.46	96.5	0.48
$19.0 \cdot 10^6$	92.8	0.45	96.2	0.48	95.0	0.46	95.0	0.47
$19.5 \cdot 10^6$	94.1	0.44	97.0	0.47	96.0	0.45	96.0	0.46
$20.0 \cdot 10^6$	94.8	0.44	95.7	0.46	94.4	0.45	94.4	0.46
$20.5 \cdot 10^6$	93.5	0.43	96.5	0.46	95.5	0.44	95.5	0.45
$21.0 \cdot 10^6$	94.7	0.43	95.1	0.45	93.8	0.43	96.3	0.44
$21.5 \cdot 10^6$	92.9	0.42	96.0	0.45	94.9	0.43	94.9	0.44
$22.0 \cdot 10^6$	94.2	0.42	96.8	0.44	95.8	0.42	95.8	0.43
$22.5 \cdot 10^6$	94.7	0.41	95.6	0.44	94.4	0.42	94.4	0.43
$23.0 \cdot 10^6$	93.6	0.41	96.4	0.43	95.4	0.41	95.4	0.42
$23.5 \cdot 10^6$	94.8	0.40	95.1	0.43	96.2	0.41	96.2	0.42
$24.0 \cdot 10^6$	93.1	0.40	96.0	0.42	94.9	0.41	94.9	0.41
$24.5 \cdot 10^6$	94.3	0.39	96.7	0.42	95.8	0.40	95.8	0.41
$25.0 \cdot 10^6$	94.8	0.39	95.5	0.41	94.4	0.40	94.4	0.40

Table 10. 95% CI performance comparison – n.

p^*	n	W		CP		WS		AC	
		CPr	$\tilde{\epsilon}_R$	*CPr*	$\tilde{\epsilon}_R$	*CPr*	$\tilde{\epsilon}_R$	*CPr*	$\tilde{\epsilon}_R$
10^{-1}	n_A: $1.4 \cdot 10^2$	93.3	0.49	96.8	0.53	95.2	0.50	95.2	0.51
	n_B: $1.6 \cdot 10^2$	94.2	0.47	95.6	0.50	94.1	0.47	95.4	0.48
10^{-2}	n_A: $1.6 \cdot 10^3$	94.0	0.49	97.1	0.52	95.9	0.50	95.9	0.52
	n_B: $1.8 \cdot 10^3$	94.1	0.45	95.4	0.48	95.4	0.46	95.4	0.47
10^{-3}	n_A: $1.6 \cdot 10^4$	94.0	0.49	97.0	0.53	95.7	0.50	95.7	0.52
	n_B: $1.7 \cdot 10^4$	94.7	0.47	95.9	0.50	94.7	0.48	95.9	0.49
10^{-4}	n_A: $1.6 \cdot 10^5$	94.0	0.49	96.9	0.53	95.7	0.51	95.7	0.52
	n_B: $1.7 \cdot 10^5$	94.7	0.47	95.9	0.50	94.7	0.48	95.8	0.49
10^{-5}	n_A: $1.6 \cdot 10^6$	94.0	0.49	96.9	0.53	95.7	0.51	95.7	0.52
	n_B: $1.7 \cdot 10^6$	94.7	0.47	95.9	0.50	94.7	0.48	95.9	0.49
10^{-6}	n_A: $1.6 \cdot 10^7$	94.0	0.49	96.9	0.53	95.7	0.51	95.7	0.52
	n_B: $1.7 \cdot 10^7$	94.7	0.47	95.9	0.50	94.7	0.48	95.8	0.49

n_A: First n where at least one interval satisfies $CPr \in 95 \pm 1\%$ and $\tilde{\epsilon}_R \leq 0.5$.
n_B: First n where all intervals satisfy $CPr \in 95 \pm 1\%$ and $\tilde{\epsilon}_R \leq 0.5$.

Table 11 gives the sample sizes required to maintain a desired level of *CPr* and $\tilde{\epsilon}_R$. Three performance schemes were investigated: (i) $CPr \in 95 \pm 3\%, \tilde{\epsilon}_R \leq 1$, (ii) $CPr \in 95 \pm 2\%, \tilde{\epsilon}_R \leq 0.75$ and (iii) $CPr = 95 \pm 1\%, \tilde{\epsilon}_R \leq 0.5$. For each scheme, the Wilson interval required the smallest sample size to achieve (and maintain) the desired performance, thus, providing further evidence of its overall superiority among the four estimators.

Table 12 shows the $\tilde{\epsilon}_R$ values pertaining to the sample sizes displayed in Table 11 (ϵ_R values were found to be very similar to the given $\tilde{\epsilon}_R$ values). It can be seen that in relation to a 95% CI, to ensure that the coverage remains within ±2%, the Wald interval requires $\tilde{\epsilon}_R \leq 0.37$, whereas the Wilson interval requires $\tilde{\epsilon}_R = 0.75$. The $\tilde{\epsilon}_R$ values corresponding to maintaining the coverage within $95 \pm 1\%$ (green table cells) are in close agreement with our recommendation to use $\tilde{\epsilon}_R \in [0.1, 0.5]$.

7. Estimating a Rare Event with a Small Sample Size

It is clear from the above results, that as expected a priori, quite large sample sizes are required to accurately estimate rare-event probabilities. Achieving accuracy on the order of magnitude for a small p is usually most relevant in large populations, where it will also be possible to collect large samples. For example, a quality engineer may have little problem in obtaining high-throughput process data of order $n = 10^6$ or greater, and, in this large-scale production setting, it will be critical to know whether the defect rate is, say, one in one thousand, or one in ten thousand. In Section 8, we assess three data-rich scenarios from the literature, that is, cases that involve estimating a small proportion while using large samples.

Notwithstanding the fact that accurate estimation of a small p is most important in large populations, an analyst may be faced with the challenge of estimating a small p with a limited sample size. We touched on this problem in Section 2.2, but now consider CI performance in more detail using the approach of Section 6.

Assume that the true proportion is of order 10^{-2}. As shown in Table 11, sample sizes of order $n = 10^3$ will be needed to accurately estimate p. However, here, we assume that the analyst

Table 11. 95% CI sample size comparison.

p^*	n	W	CP	WS	AC
10^{-1}	n_C	$1.5 \cdot 10^2$	$5.4 \cdot 10^1$	$\mathbf{3.6 \cdot 10^1}$	$4.2 \cdot 10^1$
	n_D	$2.9 \cdot 10^2$	$1.9 \cdot 10^2$	$\mathbf{6.2 \cdot 10^1}$	$1.0 \cdot 10^2$
	n_E	$6.5 \cdot 10^2$	$1.2 \cdot 10^3$	$\mathbf{3.5 \cdot 10^2}$	$\mathbf{3.5 \cdot 10^2}$
10^{-2}	n_C	$1.5 \cdot 10^3$	$7.1 \cdot 10^2$	$\mathbf{4.4 \cdot 10^2}$	$5.0 \cdot 10^2$
	n_D	$2.8 \cdot 10^3$	$2.0 \cdot 10^3$	$\mathbf{7.4 \cdot 10^2}$	$1.2 \cdot 10^3$
	n_E	$8.0 \cdot 10^3$	$1.0 \cdot 10^4$	$\mathbf{2.6 \cdot 10^3}$	$4.1 \cdot 10^3$
10^{-3}	n_C	$1.8 \cdot 10^4$	$7.0 \cdot 10^3$	$\mathbf{4.5 \cdot 10^3}$	$5.1 \cdot 10^3$
	n_D	$2.8 \cdot 10^4$	$1.7 \cdot 10^4$	$\mathbf{7.5 \cdot 10^3}$	$1.2 \cdot 10^4$
	n_E	$7.4 \cdot 10^4$	$1.0 \cdot 10^5$	$\mathbf{2.6 \cdot 10^4}$	$3.7 \cdot 10^4$
10^{-4}	n_C	$1.8 \cdot 10^5$	$7.0 \cdot 10^4$	$\mathbf{4.5 \cdot 10^4}$	$5.1 \cdot 10^4$
	n_D	$2.8 \cdot 10^5$	$1.6 \cdot 10^5$	$\mathbf{7.5 \cdot 10^4}$	$1.2 \cdot 10^5$
	n_E	$7.4 \cdot 10^5$	$1.0 \cdot 10^6$	$\mathbf{2.9 \cdot 10^5}$	$4.7 \cdot 10^5$
10^{-5}	n_C	$1.8 \cdot 10^6$	$7.0 \cdot 10^5$	$\mathbf{4.5 \cdot 10^5}$	$5.1 \cdot 10^5$
	n_D	$2.8 \cdot 10^6$	$1.7 \cdot 10^6$	$\mathbf{7.5 \cdot 10^5}$	$1.2 \cdot 10^6$
	n_E	$7.4 \cdot 10^6$	$1.0 \cdot 10^7$	$\mathbf{2.9 \cdot 10^6}$	$4.7 \cdot 10^6$
10^{-6}	n_C	$1.8 \cdot 10^7$	$7.0 \cdot 10^6$	$\mathbf{4.5 \cdot 10^6}$	$5.1 \cdot 10^6$
	n_D	$2.8 \cdot 10^7$	$1.7 \cdot 10^7$	$\mathbf{7.5 \cdot 10^6}$	$1.2 \cdot 10^7$
	n_E	$7.4 \cdot 10^7$	$1.0 \cdot 10^8$	$\mathbf{2.9 \cdot 10^7}$	$4.7 \cdot 10^7$

n_C: n_1 where $CPr \in 95 \pm 3\%$ and $\tilde{\epsilon}_R \leq 1 \ \forall \ n \geq n_1$
n_D: n_1 where $CPr \in 95 \pm 2\%$ and $\tilde{\epsilon}_R \leq 0.75 \ \forall \ n \geq n_1$
n_E: n_1 where $CPr \in 95 \pm 1\%$ and $\tilde{\epsilon}_R \leq 0.5 \ \forall \ n \geq n_1$
Smallest sample size for each scheme shown in bold.

Table 12. $\tilde{\epsilon}_R$ values corresponding to maintained coverage.

CPr	W	CP	WS	AC
$95 \pm 3\%$	0.46–0.50	0.80–0.87	0.98–1.00	0.97–1.00
$95 \pm 2\%$	0.34–0.37	0.45–0.52	0.75–0.75	0.60–0.61
$95 \pm 1\%$	0.22–0.23	0.17–0.20	0.31–0.39	0.29–0.33

$\tilde{\epsilon}_R$ values observed across the p^* range displayed as minimum–maximum.

Table 13. 95% CI performance – $p^* = p = 10^{-2}$, small n.

n	W		CP		WS		AC	
	CPr	$\tilde{\epsilon}_R$	CPr	$\tilde{\epsilon}_R$	CPr	$\tilde{\epsilon}_R$	CPr	$\tilde{\epsilon}_R$
20	18.2	1.80	98.3	9.19	98.3	8.70	98.3	11.36
40	33.1	1.73	99.3	5.22	93.9	5.08	99.3	6.57
60	45.2	1.65	97.8	3.80	97.8	3.72	97.8	4.72
80	55.2	1.57	99.1	3.07	95.3	3.00	99.1	3.74
100	63.3	1.50	98.2	2.62	92.1	2.55	98.2	3.13

is dealing with a hard-to-reach population where $n \leq 100$; the performance of the four intervals is displayed in Table 13.

It is clear that all four intervals perform quite poorly in this scenario both in terms of coverage and relative margin of error. The coverage of the Wilson interval is notably better than the others, and is reasonable for some sample sizes, albeit is still somewhat erratic. This interval does achieve excellent coverage for $n = 80$ for example, but the relative margin of error is $\tilde{\epsilon}_R \approx 3$, that is, the margin of error of 0.03 is much larger than $p = 0.01$. If the analyst only requires a rough estimate of p, for example, to answer the question of whether or not it is less than 0.1, then such a large margin of error will be acceptable. On the other hand, if the aim to is accurately estimate the order of magnitude of p, this will not be achievable for such small sample sizes (and, clearly, performance will degrade further for even smaller p). This again highlights the importance of considering relative margin of error in the small p setting, and our suggestion is to use $\tilde{\epsilon}_R \in [0.1, 0.5]$.

An anonymous reviewer advised us of two modern CI estimation approaches: an asymptotic method based on generalized fiducial inference (GFI) (Hannig 2009), and a recently-developed exact method known as the "repro samples" method (Xie and Wang 2022). The GFI method is an extension of the fiducial argument proposed by Fisher (1930), and the repro method is a simulation-based method that provides a finite sample CI coverage guarantee, which is particularly useful in small samples. We have tested both of these more modern methods (see Appendix C), and have found that they provide reasonable coverage (starting from a conservative position akin to the exact Clopper-Pearson method). However, when p and n are small, the methods experience the same issues as the classical methods we have considered; in particular, the relative margin of error is too large to be used in settings where the order of magnitude of a small p is of interest. (It is noteworthy, however, that the GFI and repro sampling methods are general inference procedures that provide good finite-sample performance in a wide range of problems beyond proportion estimation.) Ultimately, all of our work points to the fact that large samples are required in this small-p setting, and we have provided guidelines in Section 6.

8. Case Studies

We now consider the use of the relative margin of error in the estimation of small/rare-event proportions using data from the literature. More specifically, we consider: (i) a study on the prevalence of ADHD prescriptions in adolescents, (ii) a clinical trial relating to COVID-19 vaccine efficacy, and (iii) accident data from commercial jet aircraft records.

Using the values of n and $\widehat{p}$ reported in each of the aforementioned studies, we evaluate the validity of a 95% Wald CI in terms of the interval's relative margin of error. We discuss the Wald interval as it is the most commonly used interval estimator, and for each of these case studies, it produces similar results to the Clopper-Pearson, Wilson and Agresti-Coull intervals. We also refer to our sample size calculations/CI performance analyses to assess the suitability of the sample size in terms of achieving the desired coverage.

8.1. Assessing Prevalence of ADHD Medication

The first study we consider is a study conducted by Sawyer et al. (2017) to assess the prevalence of stimulant and antidepressant medication in Australian children and adolescents with symptoms of ADHD (Attention-Deficit/Hyperactivity Disorder) and major depressive disorder (MDD). A nationally representative sample of $n = 6310$ children between the age of 4 and 17 was obtained, which found that 13.7% of those with symptoms meeting the criteria of ADHD had used stimulant medications.

For a sample size of $n = 6310$, and an estimated proportion of magnitude $\widehat{p} = 0.137$, the 95% Wald CI is given as $[0.129, 0.145]$, with a realized relative margin of error of $\widehat{\epsilon_R} = \widehat{\epsilon}/\widehat{p} \approx 0.062$. Using the Delta method (see Appendix D), a 95% CI for ϵ_R is given as $[0.060, 0.064]$. (For each case study, a CI for $\tilde{\epsilon}_R$ was obtained using Monte Carlo simulation in conjunction with the Delta method, and each was found to be in agreement with

Table 14. 95% Wald CI – $\widehat{\epsilon_R}$ comparison.

$\widehat{p}$	n	$\widehat{\epsilon}$	$\widehat{\epsilon_R}$	$\widehat{p} \pm \widehat{\epsilon}$
	6310	0.008	0.06	[0.129, 0.145]
	605	0.027	0.20	[0.110, 0.164]
0.137	269	0.041	0.30	[0.096, 0.178]
	152	0.055	0.40	[0.082, 0.192]
	97	0.068	0.50	[0.069, 0.205]

the corresponding CI estimate for ϵ_R.) For this study, the ϵ_R CI values fall outside of our recommended range of $\epsilon_R \in [0.1, 0.5]$ meaning that the interval is somewhat narrower than what we recommend, that is, one could achieve an acceptable result with fewer observations. Indeed, Table 14 displays the sample sizes for a selection of $\widehat{\epsilon_R}$ values in this range, and, note, for example, that $\widehat{\epsilon_R} = 0.4$ leads to a sample size approximately forty times smaller than the sample size used in the study.

It can also be seen from Table 14 that a relative margin of error of $\widehat{\epsilon_R} = 0.2$ corresponds to a CI which is similar to that computed at $\widehat{\epsilon_R} = 0.06$, but uses a sample size that is approximately 10 times smaller. Had the order of magnitude of p been known in advance (e.g., if it was known that $p \approx 1/10$, rather than $p \approx 1/100$), then a smaller sample size would have sufficed. When p is of the order 10^{-2}, very good coverage is achieved for $n = 1800$ (see Table 10), hence the study sample size of $n = 6310$ is more than adequate for the scenarios where $p = 10^{-1}$ and $p = 10^{-2}$.

8.2. COVID-19 Vaccine Efficacy

An efficacy trial of the BNT162b2 mRNA COVID-19 vaccine was conducted by Polack et al. (2020). In this placebo-controlled, observer-blind trial, 43,548 participants were randomly assigned either the BNT162b2 vaccine or a placebo treatment. Of the 21,720 participants who received the vaccine, there were eight cases of COVID-19 after the second dose. This leads to an estimated proportion of $\widehat{p} = x/n = 8/21{,}720 \approx 3.7 \cdot 10^{-4}$, and, therefore, the 95% Wald CI is given by $[1.1 \cdot 10^{-4}, 6.2 \cdot 10^{-4}]$, with $\widehat{\epsilon_R} \approx 0.69$, and a 95% CI for ϵ_R of [0.421, 0.875]. As per Section 4.2 we recommend $\epsilon_R \in [0.1, 0.5]$, which is incorporated in the above interval. However, the CI lower bound is very close to our recommended ϵ_R upper bound, and thus, in our suggested scheme, the computed interval could be questioned with regard to its width.

Aside from having a somewhat large margin of error (relative to $\widehat{p}$), we need to consider the sample size in relation to the order of magnitude of $\widehat{p}$. That is, we must assess if the sample size is large enough to provide acceptable coverage. For example, if p were 10^{-4}, Table 11 indicates that a sample size of the order 10^5 is required, whereas, here, the sample size is of the order 10^4. Indeed, we have calculated that, with $p = 3.7 \cdot 10^{-4}$ and $n = 21{,}720$, the expected coverage is just 89.2%. The Clopper-Pearson, Wilson and Agresti-Coull intervals perform better in this n-p scheme, achieving coverage of 96.9%, 95.2%, and 95.2% (respectively). However, for this n-p combination, all three intervals have $\widehat{\epsilon_R} > 0.72$. Thus, to obtain a CI estimate where the margin of error is more consistent with $\widehat{p}$, and/or to enhance the coverage probability, a larger sample size would be required.

8.3. Commercial Aircraft Accidents

A summary of annual commercial jet aircraft flight hours, departures and accidents is provided by Boeing (2022). In the year 2021, there were 21.6 million aircraft departures with a total of 23 recorded incidents/accidents. Although *all* aircraft departures and accidents are recorded here, we may still view this as a sample from a larger population of flights that might have taken place (had demand been higher) or indeed for flights in upcoming years (provided that conditions such as aviation regulations and the composition of aircraft fleets remain similar). Therefore, it is still of interest to compute a confidence interval in this scenario, and, irrespective of the specific target population, the data still suffice for the purpose of demonstrating our proposed scheme.

For this data, $\widehat{p} \approx 1.1 \cdot 10^{-6}$, which gives a 95% Wald CI of $[0.66 \cdot 10^{-6}, 1.54 \cdot 10^{-6}]$, with $\widehat{\epsilon_R} \approx 0.40$ (95% CI for ϵ_R of [0.328, 0.494]). This relative margin of error is consistent with our recommendation of $\epsilon_R \in [0.1, 0.5]$, and as discussed in Section 4.3, provides a satisfactory estimator. Indeed, following the approach of Table 10, we have found that, when $p = 1.1 \cdot 10^{-6}$, all four estimators satisfy $CPr \in 95 \pm 1\%$ and $\tilde{\epsilon}_R \leq 0.5$ when $n = 21.6 \cdot 10^6$.

In the context of estimating the proportion of aircraft accidents, the analyst has no control over the sample size. That is to say, had a larger sample size been required, one would simply have to wait for more aircraft departures to occur. However, our analysis provides us with reassurance that our computed interval will perform satisfactorily.

9. Discussion

When constructing confidence intervals for small success probabilities it is important that the margin of error, ϵ, be considered relative to the magnitude of the proportion, p. Incompatibilities between ϵ and p can lead to completely unsatisfactory coverage or unnecessarily narrow intervals that require extremely large sample sizes. When dealing with moderate success probabilities, say $p \geq 0.2$, this is less important, but in the context of small or rare-event success probabilities, the consideration of ϵ relative to p is crucial to reduce the possibility of substantial mismatching between ϵ and p. For example, $\epsilon = 0.05$ might be considered as valid precision for $p = 10^{-1}$, but such a margin of error is not compatible with a proportion of the order $p = 10^{-3}$.

To ensure ϵ is compatible with the order of magnitude of p, we recommend using a relative margin of error scheme, ϵ_R. We suggest restricting the range of values to $\epsilon_R \in [0.1, 0.5]$ as higher values lead to imprecision and poor interval coverage, whereas lower values lead to sample sizes that are likely to be impractically large for many studies. Our recommendation of $\epsilon_R \in [0.1, 0.5]$ avoids intervals that are impractically wide or restrictively narrow in terms of sample size requirements, and we show that adequate performance is achieved within this range. In contrast to the existing literature, we have highlighted the importance of the relative margin of error, ϵ_R, in conjunction with the empirical coverage, when assessing CI performance in the small-p setting. When both criterion are considered simultaneously the Wald, Clopper-Pearson (exact), Wilson and Agresti-Coull

intervals perform similarly in many cases. In general, all four intervals fail to satisfy both criteria when the sample size is small, with improved performance at larger sample sizes as expected. For example, for a 95% confidence interval when $p = 10^{-1}$, none of the methods produce a satisfactory interval for $10 \leq n \leq 140$. Each interval achieves the nominal coverage of 95% at some (albeit not all) sample sizes in this range, but in each case the desired limit of $\epsilon_R \leq 0.5$ is exceeded. Once the sample size is increased ($n \geq 150$), and the ϵ_R requirement is satisfied, all four intervals perform well in terms of coverage.

The coverage probabilities of the Wald and Clopper-Pearson intervals for small n are generally poor, particularly in comparison to the Wilson and Agresti-Coull intervals. However, the considerable difference in coverage in such situations is rendered immaterial once the (we believe reasonable) requirement that $\epsilon_R \leq 0.5$ is considered. When satisfactory performance is defined as achieving a desired *CPr and* ϵ_R, the performance across these commonly-used intervals is much more comparable, particularly if one considers empirical coverage in the range $(1-\alpha)100 \pm 2\%$. In this relative margin of error framework the criticisms of inadequate coverage for the Wald interval, and excessive conservatism for the Clopper-Pearson interval, are somewhat alleviated, and all four intervals perform quite similarly. Although there are performance similarities, the Wilson and Agresti-Coull intervals are generally superior to the intervals of Wald and Clopper-Pearson. The Wilson and Agresti-Coull intervals achieve similar *CPr* and ϵ_R values for given n-p-α combinations, however, the Wilson interval is narrower and achieves favorable performance at lower sample sizes.

When the success probability is small, failure to consider the margin of error relative to the order of magnitude of the estimated proportion can result in poor coverage, and/or intervals that are unnecessarily narrow or excessively wide. As shown in the case studies presented in Section 8, the relative margin of error criterion provides a simple and effective assessment of the validity of an estimated interval in terms of its width/margin of error. For example, we have shown that all of the interval estimators considered in this article performed poorly for the COVID-19 study (Section 8.2) in terms of the relative margin or error, meaning that the confidence intervals were all impractically wide—and the Wald interval also had notably poor coverage. It is important to ensure that the interval precision is compatible with the order of magnitude of p. The relative margin of error serves as a useful evaluation criterion in this regard, and as such, we suggest that it should be considered when planning statistical studies.

Appendices

Appendix A. Sample Size Formulas

A.1. Clopper-Pearson Interval

Letting n_{LB} be the minimum n satisfying: $p^* - \text{Beta}\left(\alpha/2; np^*, n(1-p^*)+1\right) \leq \epsilon$ and letting n_{UB} be the minimum n satisfying: $\text{Beta}\left(1-\alpha/2; np^*+1, n(1-p^*)\right) - p^* \leq \epsilon$, the sample size is given by:

$$n = \lceil \max\{n_{LB}, n_{UB}\} \rceil$$

A.2. Wilson Interval

$$\epsilon = \frac{z_{\alpha/2}\sqrt{p(1-p)/n + z^2_{\alpha/2}/(4n^2)}}{1 + z^2_{\alpha/2}/n} \implies$$

$$n = \left\lceil \max\left\{ \frac{z^2_{\alpha/2}}{2\epsilon^2}\left(p^*(1-p^*) - 2\epsilon^2 \pm \sqrt{\epsilon^2(1-4p^*(1-p^*)) + (p^*(1-p^*))^2}\right)\right\}\right\rceil$$

A.3. Agresti-Coull Interval

Letting $\widetilde{p} = (np + z^2_{\alpha/2}/2)/(n + z^2_{\alpha/2})$ and $\widetilde{n} = n + z^2_{\alpha/2}$:

$$\epsilon = z_{\alpha/2}\sqrt{\frac{\widetilde{p}(1-\widetilde{p})}{\widetilde{n}}} \implies$$

$$n = \left\lceil \max\left\{ -\frac{1}{3a}\left(b + \xi^k C + \frac{\Delta_0}{\xi^k C}\right)\right\}\right\rceil;$$

$$k \in \{0,1,2\}$$

where

$a = 4\epsilon^2$

$b = 4z^2_{\alpha/2}(3\epsilon^2 - p^*(1-p^*))$

$\xi = (-1+\sqrt{-3})/2$

$C = \sqrt[3]{\left(\Delta_1 \pm \sqrt{\Delta_1^2 - 4\Delta_0^3}\right)/2}$

$\Delta_0 = 16z^4_{\alpha/2}(3\epsilon^2 - p^*(1-p^*))^2 - 24\epsilon^2 z^4_{\alpha/2}(6\epsilon^2 - 1)$

$\Delta_1 = 128z^6_{\alpha/2}(3\epsilon^2 - p^*(1-p^*))^3 + 432\epsilon^4 z^6_{\alpha/2}(4\epsilon^2 - 1) - 288\epsilon^2 z^6_{\alpha/2}(3\epsilon^2 - p^*(1-p^*))(6\epsilon^2 - 1)$

Appendix B. Fixed Margin of Error Schemes

Table 15 shows that for the fixed margin of error scheme: $\epsilon = 4 \cdot 10^{-4}, p^* = 0.5$, all four intervals produce coverage close to the nominal value. For the remaining three margin of error schemes the Wald interval produces very poor coverage in comparison to the other intervals. For example, in *Scheme 1* ($\epsilon = 4 \cdot 10^{-2}, p^* = p$), the Wald coverage for $p = 10^{-2}$ is just 21.4%, whereas the coverage of the Clopper-Pearson, Wilson and Agresti-Coull intervals is 97.7%.

Whilst the Clopper-Pearson, Wilson and Agresti-Coull intervals achieve better coverage, all four intervals fail to produce satisfactory coverage if the margin of error scheme is not compatible with the magnitude of p. For example, in *Scheme 3* ($\epsilon = 4 \cdot 10^{-4}, p^* = p$) the Wald interval produces a coverage of just 0.2% for $p = 10^{-5}$. However, the Clopper-Pearson, Wilson and Agresti-Coull intervals achieve coverage of 99.8%, 99.8%, and 100%, respectively. Whilst these coverage probabilities are much closer to the desired 95% coverage than the Wald's 0.2%, they are too far from the nominal value to be of practical use.

Referring to Table 16, for $\epsilon_R = 0.75$ and $p^* = p \leq 10^{-2}$, the Wald coverage is considerably poorer than the other three intervals. Both the Clopper-Pearson and Agresti-Coull intervals produce reasonable coverage at approximately 97%. The Wilson interval performs the best, producing coverage very close to the nominal 95% $\forall\, p$.

Appendix C. CI Estimation using GFI and Repro Samples Methods

To compare the performance of the Wald, Clopper-Pearson, Wilson and Agresti-Coull methods against the more modern GFI (Hannig

Table 15. Coverage comparison – fixed ϵ.

MoE Scheme	p	W	CP	WS	AC
1. $\epsilon = 4 \cdot 10^{-2}$, $p^* = p$	10^{-1}	93.8	96.9	94.7	95.9
	10^{-2}	21.4	97.6	97.6	97.6
	10^{-3}	00.3	99.7	99.7	99.7
	10^{-4}	00.0	100	100	100
	10^{-5}	00.0	100	100	100
	10^{-6}	00.0	100	100	100
2. $\epsilon = 4 \cdot 10^{-2}$, $p^* = 0.5$	10^{-1}	94.5	95.9	95.2	95.2
	10^{-2}	93.1	96.3	94.1	97.8
	10^{-3}	45.2	97.7	97.7	99.7
	10^{-4}	05.8	99.8	94.2	100
	10^{-5}	00.6	99.4	99.4	100
	10^{-6}	00.1	99.9	99.4	100
3. $\epsilon = 4 \cdot 10^{-4}$, $p^* = p$	10^{-1}	95.0	95.0	95.0	95.0
	10^{-2}	95.0	95.1	95.0	95.1
	10^{-3}	93.1	96.0	94.9	94.9
	10^{-4}	21.3	97.5	97.5	99.8
	10^{-5}	00.2	99.8	99.8	100
	10^{-6}	00.0	100	100	100
4. $\epsilon = 4 \cdot 10^{-4}$, $p^* = 0.5$	10^{-1}	95.0	95.0	95.0	95.0
	10^{-2}	95.0	95.0	95.0	95.0
	10^{-3}	95.0	95.1	95.0	95.0
	10^{-4}	94.9	95.2	95.0	95.0
	10^{-5}	94.9	96.1	95.5	95.5
	10^{-6}	93.4	96.3	94.0	97.7

Table 16. Coverage comparison – $\epsilon_R = 0.75$.

p^*	W	CP	WS	AC
10^{-1}	94.7	97.0	94.6	94.6
10^{-2}	90.2	97.0	94.9	97.0
10^{-3}	90.4	96.9	94.6	96.9
10^{-4}	90.4	96.9	94.6	96.9
10^{-5}	90.4	96.9	94.6	96.9
10^{-6}	90.4	96.9	94.6	96.9

2009) and repro sampling (Xie and Wang 2022) methods, we consider $p \in [10^{-1}, 10^{-4}]$, with sample sizes ranging from $n = 20$ to $n = 10^4$. This covers a range of scenarios from the main article, including larger samples as per Section 6 and smaller samples as per Section 7. As these more modern methods are simulation-based, this makes exact performance calculations more challenging (as were done in the main article). Therefore, here, we have carried out a simulation study with a large number of replicates (5000). The results are presented in Table 17.

We can see from Table 17, that, in terms of coverage, both of these methods converge toward the nominal 95% level with the sample size (starting from a conservative point akin to the exact Clopper-Pearson interval). Interestingly, in some of the more challenging scenarios presented here, for example, $p = 10^{-3}$ with $n = 10^3$ and $p = 10^{-4}$ and $n = 10^4$, one or both of these methods provide the best empirical coverage (albeit still being a 2–3 percentage points out from the nominal level in these scenarios). In any case, these methods do not appear to offer dramatic improvements on the other methods considered, and, in particular, the key issue of $\tilde{\epsilon}_R$ being large for small n is intrinsic to all of the methods. It is important to recognize, however, that the GFI and repro sampling methods are not limited only to proportion estimation, but, rather, are general inferential techniques that extend far beyond this to many other problems.

Table 17. Comparison of classical and modern CI methods.

p^*	n	W		CP		WS		AC		GFI		Repro	
		CPr	$\tilde{\epsilon}_R$	*CPr*	$\tilde{\epsilon}_R$	*CPr*	$\tilde{\epsilon}_R$	*CPr*	$\tilde{\epsilon}_R$	*CPr*	$\tilde{\epsilon}_R$	*CPr*	$\tilde{\epsilon}_R$
10^{-1}	20	87.6	1.17	98.9	1.46	95.7	1.32	95.7	1.46	97.9	1.31	98.2	1.42
	60	94.1	0.74	97.2	0.82	95.2	0.76	95.2	0.80	95.7	0.76	97.1	0.80
	10^2	93.2	0.58	95.6	0.63	93.6	0.59	97.2	0.61	96.1	0.59	95.8	0.61
	10^3	95.3	0.19	95.5	0.19	94.9	0.19	94.9	0.19	94.8	0.19	94.6	0.18
	10^4	95.1	0.06	95.1	0.06	94.9	0.06	94.9	0.06	94.9	0.06	94.9	0.06
10^{-2}	20	18.2	1.80	98.3	9.19	98.3	8.70	98.3	11.36	98.6	7.84	98.5	9.02
	60	45.2	1.65	97.8	3.80	97.8	3.72	97.8	4.72	97.5	3.26	98.3	3.73
	10^2	63.3	1.50	98.2	2.62	92.1	2.55	98.2	3.13	98.3	2.27	98.2	2.57
	10^3	92.7	0.61	97.6	0.67	96.4	0.64	96.4	0.66	94.4	0.62	96.2	0.65
	10^4	94.6	0.19	95.6	0.20	95.0	0.20	95.0	0.20	94.6	0.19	94.9	0.19
10^{-3}	20	02.0	1.90	98.0	85.01	98.0	81.22	100	>100	98.2	69.93	98.2	83.91
	60	05.8	1.91	99.8	30.69	94.2	30.84	99.8	42.46	98.8	25.05	99.7	30.26
	10^2	09.5	1.89	99.5	19.00	90.5	19.27	99.5	26.50	99.4	15.58	99.5	18.82
	10^3	63.2	1.51	98.1	2.67	91.9	2.65	98.1	3.29	97.9	2.35	97.9	2.63
	10^4	92.6	0.61	97.5	0.67	96.3	0.64	96.3	0.67	94.4	0.63	96.5	0.66
10^{-4}	20	00.2	1.91	99.8	>100	99.8	>100	100	>100	99.8	>100	99.8	>100
	60	00.6	1.94	99.4	>100	99.4	>100	100	>100	99.4	>100	99.4	>100
	10^2	01.0	1.94	99.0	>100	99.0	>100	100	>100	99.3	>100	98.9	>100
	10^3	09.5	1.90	99.5	19.33	90.5	19.94	100	27.66	99.6	15.83	99.4	19.11
	10^4	63.2	1.52	98.1	2.68	91.9	2.66	98.1	3.29	98.2	2.34	97.4	2.65

Appendix D. CI For ϵ_R

By the Delta method, $f(\widehat{p}) \to \mathrm{N}(f(p), f'(p)^2\sigma^2_{\widehat{p}})$ as $n \to \infty$. Given $f(\widehat{p}) = \widehat{\epsilon_R} = \widehat{\epsilon}/\widehat{p}$, and $\sigma_{\widehat{p}} = \sqrt{\widehat{p}(1-\widehat{p})/n}$, a $(1-\alpha)100\%$ Wald CI for ϵ_R is given by

$$\frac{\tilde{z}}{\sqrt{n}}\sqrt{\frac{1-\widehat{p}}{\widehat{p}}} \pm z_{\alpha/2}\frac{\tilde{z}}{\sqrt{4n\widehat{p}^4}}\sqrt{\frac{\widehat{p}}{1-\widehat{p}}}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},$$

where $\tilde{z}$ is the $(1-\tilde{\alpha}/2)$ quantile of the standard normal distribution, and $\tilde{\alpha}$ is the significance level pertaining to $\widehat{\epsilon}$.

Acknowledgments

The authors would like to thank two reviewers, an associate editor, and the editor for their useful comments and suggestions. The article was much improved as a result of their feedback.

Disclosure Statement

The authors report that there are no competing interests to declare.

Funding

This work was funded by the Technological University of the Shannon and Science Foundation Ireland (via the Confirm Smart Manufacturing Research Centre, grant number: 16/RC/3918).

ORCID

Owen McGrath http://orcid.org/0000-0002-8040-7390
Kevin Burke http://orcid.org/0000-0001-8724-809X

References

Agresti, A., and Coull, B. A. (1998), "Approximate is Better than 'Exact' for Interval Estimation of Binomial Proportions," *The American Statistician*, 52, 119–126. [438,442]

Andersson, P. G. (2023), "The Wald Confidence Interval for a Binomial p as an Illuminating "Bad" Example," *The American Statistician*, 74, 443–448. [438,442]

Blyth, C. R., and Still, H. A. (1983), "Binomial Confidence Intervals," *Journal of the American Statistical Association*, 78, 108–116. [438,442]

Boeing. (2022), "Statistical Summary of Commercial Jet Airplane Accidents," available at *https://www.boeing.com/resources/boeingdotcom/company/about_bca/pdf/statsum.pdf*. [Online] [Accessed on 04-December-2023]. [437,446]

Böhning, D. (1994), "Better Approximate Confidence Intervals for a Binomial Parameter," *The Canadian Journal of Statistics*, 22, 207–218. [438]

Brown, L. D., Cai, T. T., and DasGupta, A. (2001), "Interval Estimation for a Binomial Proportion," *Statistical Science*, 16, 101–133. [438,442]

Clopper, C. J., and Pearson, E. S. (1934), "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, 26, 404–413. [438]

Evans, J. R., and Lindsay, W. M. (2015), *An Introduction to Six Sigma & Process Improvement* (2nd ed.), Boston, MA: Cengage Learning. [437]

Fisher, R. (1930), "Inverse Probability," *Proceedings of the Cambridge Philosophical Society*, xxvi, 528–535. [445]

Fleiss, J. L., Levin, B., and Cho Paik, M. (2003), *Statistical Methods for Rates and Proportions* (3rd ed.), Hoboken, NJ: Wiley. [441]

Gonçalves, L., De Oliveira, M. R., Pascoal, C., and Pires, A. (2012), "Sample Size for Estimating a Binomial Proportion: Comparison of Different Methods," *Journal of Applied Statistics*, 39, 2453–2473. [437,438,439]

Hannig, J. (2009), "On Generalized Fiducial Inference," *Statistica Sinica*, 19, 491–544. [445,448]

Kabaila, P., Welsh, A. H., and Abeysekera, W. (2016), "Model-Averaged Confidence Intervals," *Scandinavian Journal of Statistics*, 43, 35–48. [438]

Korn, E. L. (1986), "Sample Size Tables for Bounding Small Proportions," *Biometrics*, 42, 213–216. [439]

Krishnamoorthy, K., and Peng, J. (2007), "Some Properties of the Exact and Score Methods for Binomial Proportion and Sample Size Calculation," *Communications in Statistics - Simulation and Computation*, 36, 1171–1186. [438]

Leemis, L. M., and Trivedi, K. S. (1996), "A Comparison of Approximate Interval Estimators for the Bernoulli Parameter," *The American Statistician*, 50, 63–68. [437,441]

Liu, W., and Bailey, B. J. R. (2002), "Sample Size Determination for Constructing a Constant Width Confidence Interval for a Binomial Success Probability," *Statistics & Probability Letters*, 56, 1–5. [439]

Lwanga, S. K., and Lemeshow, S. (1991), *Sample Size Determination in Health Studies: A Practical Manual*, Geneva: World Health Organisation. [438]

Newcombe, R. G. (1998), "Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods," *Statistics in Medicine*, 17, 857–872. [437,438,439]

Park, H., and Leemis, L. M. (2019), "Ensemble Confidence Intervals for Binomial Proportions," *Statistics in Medicine*, 38, 3460–3475. [438,439]

Pires, A. M., and Amado, C. (2008), "Interval Estimators for a Binomial Proportion: Comparison of Twenty Methods," *REVSTAT - Statistical Journal*, 6, 165–197. [437,438]

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Pérez Marc, G., Moreira, E. D., Zerbini, C., Bailey, R., Swanson, K. A., Roychoudhury, S., Koury, K., Li, P., Kalina, W. V., Cooper, D., Frenck, R. W., Hammitt, L. L., and …Gruber, W. C. (2020), "Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine," *The New England Journal of Medicine*, 383, 2603–2615. [437,446]

Sawyer, M. G., Reece, C. E., Sawyer, A. C., Johnson, S., Lawrence, D., and Zubrick, S. R. (2017), "The Prevalence of Stimulant and Antidepressant Use by Australian Children and Adolescents with Attention-Deficit/Hyperactivity Disorder and Major Depressive Disorder: A National Survey," *Journal of Child and Adolescent Psychopharmacology*, 27, 177–184. [445]

Thulin, M. (2014), "Coverage-Adjusted Confidence Intervals for a Binomial Proportion," *Scandinavian Journal of Statistics*, 41, 291–300. [438]

Turek, D., and Fletcher, D. (2012), "Model-averaged Wald Confidence Intervals," *Computational Statistics and Data Analysis*, 56, 2809–2815. [438]

Vollset, S. E. (1993), "Confidence Intervals for a Binomial Proportion," *Statistics in Medicine*, 12, 809–824. [438,442]

Vos, P., and Hudson, S. (2005), "Evaluation Criteria for Discrete Confidence Intervals: Beyond Coverage and Length," *The American Statistician*, 59, 137–142. [439]

Wilson, E. (1927), "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association*, 22, 209–212. [437,438]

Woodall, W. H., and Montgomery, D. C. (2014), "Some Current Directions in the Theory and Application of Statistical Process Monitoring," *Journal of Quality Technology*, 46, 78–94. [437]

Xie, M., and Wang, P. (2022), "Repro Samples Method for Finite- and Large-Sample Inferences," arXiv preprint arXiv:2206.06421. [445,448]