

Article

Adaptive Bayesian Interval Estimation for Rare Binomial Events: A Variance-Blending Calibration Framework

Albert Antwi ^{1,*} , Alexander Boateng ² and Daniel Maposa ^{3,*} ¹ Department of Mathematical Sciences, Sol Plaatje University, Kimberly 8300, South Africa² Department of Mathematics and Computer Science, Modern College of Business and Science, Muscat 133, Oman; alexander.boateng@mcbs.edu.om³ Department of Statistics and Operations Research, University of Limpopo, Polokwane 0727, South Africa

* Correspondence: albert.aantwi@outlook.com (A.A.); daniel.maposa@ul.ac.za (D.M.)

Abstract

Classical binomial interval methods often exhibit poor performance when applied to extreme conditions, such as rare-event scenarios or small-sample estimations. Recent frequentist and Bayesian approaches have improved coverage in small samples and rare events. However, they typically rely on fixed error margins that do not scale with the magnitude of the proportion. This distorts uncertainty quantification at the extremes. As an alternative method to reduce these boundary distortions, we propose a novel hybrid approach. It blends Bayesian, frequentist, and approximation-based techniques to estimate robust and adaptive intervals. The variance incorporates sampling variability, Wilson score margin of error, a tuned credible level, and a gamma regularization term that is inversely proportional to sample size. Extensive simulation studies and real-data applications demonstrate that the proposed method consistently achieves better coverage proportions at all sample sizes and proportions. It provides more conservative interval widths below a sample size of 50 and competitively narrower widths from moderate to large sample sizes, especially beyond 50, compared to the Jeffreys' and Wilson score intervals. Geometric analysis of the tuning curves demonstrates how the blended method adaptively tunes credible levels across binomial extremes. It starts at higher values for small samples and gradually flattens into near-linear, symmetric trajectories as sample size increases. This ensures robust coverage and balanced sensitivity. Our method offers a theoretically grounded, computationally efficient, and practically robust estimation of rare-event intervals. These intervals have applications in safety-critical reliability, epidemiology, and early-phase clinical trials.

Keywords: rare events; adaptive variance; Bayesian intervals; credible level tuning**MSC:** 62F15; 62F25

Academic Editors: Massimiliano Giacalone and Karina Gibert

Received: 30 October 2025

Revised: 27 November 2025

Accepted: 11 December 2025

Published: 14 December 2025

Citation: Antwi, A.; Boateng, A.; Maposa, D. Adaptive Bayesian Interval Estimation for Rare Binomial Events: A Variance-Blending Calibration Framework. *Mathematics* **2025**, *13*, 3988. <https://doi.org/10.3390/math13243988>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Rare-event phenomena such as emerging disease outbreaks, earthquakes, and industrial defects are typically characterized by extremely low probability of occurrence. The estimation of binomial proportions for such events poses significant challenges, particularly due to the inadequacy of conventional confidence interval methods. Classical approaches, such as the Wald interval, frequently perform poorly for these events due to their reliance on asymptotic normality, which fails as the true proportion approaches the extreme boundaries of the binomial proportions [1,2]. The Clopper–Pearson interval also tends to be overly conservative, thus producing wide intervals that result in statistical

inefficiency [3]. Conversely, asymptotic approximation methods such as the Wald interval may yield substantial under-coverage probabilities when the event probability is small [2].

To address the weaknesses of the traditional confidence interval estimations under rare-event conditions, many frequentist alternative approaches [4–10] and Bayesian alternatives [10,11] have been proposed. Although these methods have demonstrated improvements in coverage probability, especially in small samples or extreme and rare events, they are built under the assumption of a fixed error margin, which is not plausible for events with proportions at the extreme ends of the probability spectrum (0 and 1). This is because such margins fail to scale with the magnitude of the proportion being estimated, which may lead to under- or overestimated uncertainties, depending on the location of the proportion within the parameter space [12]. This results in inaccurate estimates because the intervals are either too wide to be practically useful or too narrow to be valid, depending on the sample size and the rarity of the event [13].

The incorporation of adaptive margins that scale with the magnitude of the proportion into existing classical methods has greatly improved the accuracy of both the coverage probability and the interval width of rare events [13]. Although alternative adaptive Bayesian methods—such as the adaptive Huber’s ε -contamination model, which assumes that the error margin (ε) is unknown and then adaptively estimates—have been shown to perform well in estimating probabilities on a continuous scale, they tend to underperform in small-sample or rare-event scenarios [14]. Bayesian adaptive prior methods have also been proposed and shown to perform well in rare events, but they require correct scale parameter tuning, as they may increase the credible width when incorrectly tuned or specified [15]. As an alternative, we introduce an adaptive Bayesian variance-blending calibration framework (hereafter known as the blended approach or method).

The proposed method offers several key advantages. It significantly improves coverage accuracy by integrating information from the Beta distribution via Jeffreys’ prior, the Wilson score, and a credible-level tuning parameter, as well as a gamma regularization parameter which enables the targeted nominal coverage to be maintained more efficiently across a wide range of sample sizes and true proportions. The use of the Beta distribution and logit transformations stabilizes estimates near the extremes; thus, our method is robust for small sample sizes and extreme proportions near 0 or 1, which provides more accurate and stable intervals. Additionally, by combining multiple sources of variability, the approach avoids overly conservative or overly optimistic interval widths, resulting in intervals that are reasonably narrow without sacrificing coverage. The method, therefore, strikes an optimal balance between precision and coverage. The incorporation of the gamma and the credible level parameters allows the intervals to adapt to the sample size and true proportions, which accounts for sampling uncertainty and potential overdispersion. Computationally, the proposed method is fully vectorizable and memory-efficient, thus enabling seamless integration into automated workflows and statistical software. Furthermore, its modular structure allows for rapid diagnostics and tuning, facilitating iterative refinement without computational bottlenecks.

This paper makes contributions to the statistical theory of estimation by introducing a blended confidence interval method that inherits asymptotically desirable properties (consistency, efficiency, and asymptotic normality) of frequentist and model-based inferences. Specifically, as $n \rightarrow \infty$, the proposed interval exhibits consistency, asymptotic normality, and nominal coverage probability. The blending mechanism is theoretically grounded in logit-scale transformations and adaptive variance blending, ensuring that the method remains robust for small samples and across extreme events. Furthermore, this paper makes practical contributions in that the method is robust and adaptive to extreme and rare events, as well as a small sample size. It also yields high coverage with conservative

width, thus making it more adaptable to problems in various fields and applications. The rest of the paper is outlined as follows. The theoretical framework is presented in Section 2, while the simulation design is described in Section 3. The results and discussion are then provided in Section 4, and the paper concludes with a summary in Section 5.

2. Theoretical Framework

In this section, we provide details of the traditional Bayesian interval estimation approach and then extend it to an adaptive version by incorporating our proposed margin correction and variance blending method. We also provide the methodology for the Wilson approach as well as methods for assessing the performances of the three methods. Accurate interval estimation for binomial proportions is fundamental in statistical inference. Bayesian methods offer probabilistically coherent intervals that are derived from the posterior distribution. For binomially distributed data, Bayesian updating with conjugate Beta priors yields Beta-distributed posteriors, thus enabling straightforward computation of credible intervals. However, traditional credible intervals may be too conservative or too narrow, especially for small samples with extreme proportions. To address this, we introduce a blended variance calibration framework that incorporates a tuned credible level within a Bayesian framework to enable adaptive uncertainty quantification.

2.1. Standard Bayesian Credible Interval Using the Beta Distribution and Jeffreys' Prior

Consider a sample size n (number of trials) and number of successes k , where $k \sim \text{Binomial}(n, p)$ and p , the true underlying success probability. The goal is to estimate $(\hat{p})$ using our proposed method, the standard Bayesian method with Jeffreys' prior, and Wilson score interval.

2.1.1. Formulation

Assume that p follows the prior distribution of Jeffreys' [16], such that

$$(p|k) \sim \text{Beta}\left(\frac{1}{2} + k, \frac{1}{2} + n - k\right) \quad (1)$$

By denoting $a = \frac{1}{2} + k$ and $b = \frac{1}{2} + n - k$, with a credible level $(1 - \alpha)$, then the $100(1 - \alpha)\%$ credible interval is defined by Equation (2) as

$$\left[\text{Beta}^{-1}\left(\frac{\alpha}{2}, a, b\right), \text{Beta}^{-1}\left(1 - \frac{\alpha}{2}, a, b\right) \right], \quad (2)$$

where $\text{Beta}^{-1}(q, a, b)$ is the inverse CDF (percent point function) of the Beta distribution. Given the posterior mean and the variance of the $\text{Beta}(a, b)$ as

$$E(p) = \frac{a}{(a + b)} \text{ and } \text{Var}(p) = \frac{ab}{(a + b)^2(a + b + 1)}, \quad (3)$$

we define the credible width as

$$\text{Beta}^{-1}\left(\frac{\alpha}{2}, a, b\right) - \text{Beta}^{-1}\left(1 - \frac{\alpha}{2}, a, b\right). \quad (4)$$

This is a quantile-based width and, although it narrows as $n \rightarrow \infty$, it does not scale directly with the posterior variance. Quantiles reflect cumulative distribution and not local spread. Therefore, the standard Bayesian and the Jeffreys' interval shrink as sample size increases but do not scale proportionally with the posterior variance; hence, their quantile spreads are fixed. The analysis below explains why the standard Bayesian credible interval width does not scale proportionally with variance.

2.1.2. Non-Proportionality Scaling of the Bayesian Credible Width with Variance

Let the quantile function $Q_q(a, b)$ be implicitly defined as

$$Q_q(a, b) = \int_0^{Q_q(a, b)} f(t) dt, \text{ where } f(t) = \frac{t^{a-1}(1-t)^{b-1}}{B(a, b)} \quad (5)$$

The definition in Equation (5) suggests that $Q_q(a, b)$ is the value such that the cumulative distribution function (CDF) equals q , i.e., $F(Q_q) = q$. Using the chain rule,

$$\frac{d}{dq} F(Q_q) = \left(\frac{dF}{dt} \right) \left(\frac{dQ_q}{dq} \right) = f(Q_q) \frac{dQ_q}{dq} \quad (6)$$

However, since $F(Q_q) = q$, substitution leads to $\frac{d}{dq} q = 1 \Rightarrow f(Q_q) \frac{dQ_q}{dq} = 1$. Therefore

$$\frac{dQ_q}{dq} = \frac{1}{f(Q_q)} \quad (7)$$

Substitution of Equation (7) into Equation (5) yields

$$\frac{dQ_q}{dq} = \frac{1}{f(Q_q)} = \frac{B(a, b)}{Q_q^{a-1}(1-Q_q)^{b-1}} \quad (8)$$

Now, let q_0 be the central quantile and $\Delta_q = \frac{\alpha}{2}$ be a $100(1-\alpha)\%$ interval; the lower and upper quantiles are respectively defined by $Q_\ell = Q_{q_0-\Delta_q}$, $Q_u = Q_{q_0+\Delta_q}$. Let the interval width be denoted by $w = Q_u - Q_\ell$. Using a first-order Taylor expansion, it can be shown that

$$Q_{q_0 \pm \Delta_q} \approx Q_{q_0} \pm \left. \frac{dQ_q}{dq} \right|_{q_0} \cdot \Delta_q \text{ and } w \approx 2 \left. \frac{dQ_q}{dq} \right|_{q_0} \cdot \Delta_q \quad (9)$$

Therefore, by substituting the quantile derivative, the interval width becomes

$$w \approx 2\Delta_q \frac{B(a, b)}{Q_{q_0}^{a-1}(1-Q_{q_0})^{b-1}} \quad (10)$$

From the quantile derivative in Equation (8)

$$\left. \frac{dQ_q}{dq} \right|_{q_0} \rightarrow \infty \text{ if } a < 0 \text{ since } Q_{q_0}^{a-1} \rightarrow 0 \text{ as } Q_{q_0} \rightarrow 0.$$

Similarly,

$$\left. \frac{dQ_q}{dq} \right|_{q_0} \rightarrow \infty \text{ if } b < 0 \text{ since } (1-Q_{q_0})^{b-1} \rightarrow 0 \text{ as } Q_{q_0} \rightarrow 1.$$

The implications are that near the boundaries (0 and 1), the quantile function becomes extremely sensitive to small changes in q , especially when $a < 1$ or $b < 1$, which leads to distorted interval widths in the tails. Therefore, width behaves nonlinearly, especially at the boundaries. This explains why the intervals stretch in the tails and the width does not scale proportionally with variance.

2.2. Wilson Score Interval

Given the Z-score $z = \Phi^{-1}(1 - \frac{\alpha}{2})$, the confidence level $(1 - \alpha)$, and $\hat{p} = \frac{k}{n}$, where k and n are previously defined, then the Wilson score interval [2] is formulated as

$$\left[\frac{\left(\hat{p} + \frac{z^2}{2n}\right) - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}, \frac{\left(\hat{p} + \frac{z^2}{2n}\right) + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \right] \quad (11)$$

2.3. Adaptive Bayesian-Logit-Scaled Variance-Blending Calibration Framework

Here we present a detailed formulation of our proposed method. We then prove its width's proportionality to the margin of error and how the margin of error also adapts to the sample size. We further assess the asymptotic properties of the method.

2.3.1. Formulation

Using the Jeffreys' Beta posterior prior [16], we transform the lower and the upper confidence limits by first scaling it such that

$$\ell = \text{Beta}^{-1}\left(\frac{1-\phi}{2}, a, b\right) \text{ and } u = \text{Beta}^{-1}\left(1 - \frac{(1-\phi)}{2}, a, b\right)$$

where ϕ is the credibility multiplier representing the coverage level, and it determines how wide the interval is based on the desired confidence level. It therefore serves as a regularization parameter, and it allows the calibration (via optimization) to shrink extreme intervals to stabilize the estimates. Now let us transform the interval on a logit scale as follows,

$$\text{Logit}_L = \log\left(\frac{\ell}{1-\ell}\right) \text{ and } \text{Logit}_U = \log\left(\frac{u}{1-u}\right)$$

The center and the half-width are computed as

$$c = \frac{\text{Logit}_L + \text{Logit}_U}{2} \text{ and } h = \frac{\text{Logit}_U - \text{Logit}_L}{2} \quad (12)$$

Definition 1. Given the sample variance ($\text{var}_{\text{sample}}$), model-based variance (var_{γ}), and the Wilson variance approximation ($\text{MOE}_{\text{Wilson}}$) defined by

$$\text{var}_{\text{sample}} = \frac{1}{n\hat{p}(1-\hat{p}) + \epsilon}, \text{var}_{\gamma} = \left(\frac{\theta}{\sqrt{n}}\right)^2 \text{ and } \text{MOE}_{\text{Wilson}} = \left(z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}\right)^2$$

the blended variance and the corresponding standard error are given by

$$\text{var}_{\text{Blend}} = \text{var}_{\gamma} + \text{var}_{\text{sample}} + \text{MOE}_{\text{Wilson}} \text{ and } \text{SE}_{\text{Blend}} = \sqrt{\text{var}_{\text{Blend}}}$$

where $\epsilon = 1 \times 10^{-8}$ is a correction factor that prevents division by zero.

Now, substituting the blended standard error in Definition 1 into the half-width in Equation (12) yields the adaptive margin of error in Equation (13) below.

$$\delta = h\left(\frac{\text{SE}_{\text{Blend}}}{\max(\text{SE}_{\text{Blend}}, \epsilon)}\right) \quad (13)$$

We now adjust the logit interval on the probability scale to obtain the credible interval as

$$\left[\max\left(0, \frac{1}{1 + e^{-(c-\delta)}}\right), \min\left(1, \frac{1}{1 + e^{-(c+\delta)}}\right) \right] \quad (14)$$

It is observed in Definition 1 that the blended variance assumes an additive structure without explicit weights. The additive structure of the blended variance is rooted in

decision-theoretic principles and variance decomposition under quadratic risk. In classical risk frameworks, particularly under squared-error loss, the total risk is expressed as the sum of independent variance components, where each represents a distinct source of uncertainty [17,18]. This additive property ensures interpretability and coherence because each term contributes proportionally to the overall uncertainty without requiring complex interactions. Introducing explicit weights would necessitate estimating additional hyperparameters, which increases computational complexity and introduces potential instability, especially in small samples where parameter estimation is inherently noisy. Weighted schemes also risk overfitting specific scenarios; thus, they reduce generalizability across rare-event regimes. By contrast, the unweighted additive form provides a parsimonious solution that is theoretically justified and empirically robust. Furthermore, asymptotic analysis (see Section 2.3.3) confirms that all components decay at a rate $O(1/n)$, which ensures that none of the sources dominate as the sample size grows. This property aligns with minimax risk principles thus emphasizing robustness and stability over aggressive shrinkage, which is critical for rare-event inference where under-coverage can have severe consequences.

The variance blending technique integrates frequentist variance decomposition, Bayesian regularization, and decision-theoretic risk aggregation [2,17–19]. Each component serves a unique role in addressing limitations of traditional interval estimation methods. The sampling variance captures uncertainty inherent in binomial data and is derived from the Fisher information on the logit scale. Its magnitude increases near the boundaries of the parameter space (0 or 1) and in small samples, thus reflecting the heightened variability in these situations. The gamma regularization term introduces a prior-informed penalty that stabilizes estimates when data are sparse or proportions are extreme. This term acts as a safeguard against erratic behavior by damping excessive variability. However, its influence diminishes with increasing sample size, thus preserving asymptotic efficiency. The Wilson margin penalty complements these adjustments by correcting for skewness and boundary effects which ensures conservative interval widths in low-proportion regions where classical methods often fail. Together, these components provide a balanced framework that adapts dynamically to sample size and proportion, hence mitigating distortions without sacrificing efficiency.

Regularization within the blended variance is further enhanced through credible level tuning, which optimizes the trade-off between coverage probability and interval width. This adaptive calibration ensures that intervals remain sufficiently wide to maintain nominal coverage in rare-event scenarios while avoiding unnecessary conservatism in moderate regimes. The integration of Bayesian and frequentist elements provides dual advantages. While the Bayesian priors stabilize inference near boundaries, the frequentist components maintain desirable asymptotic properties such as consistency and efficiency. Practically, this design offers robustness across diverse applications such as epidemiology, reliability engineering, and early-phase clinical trials, where rare events and small samples are common.

The credible interval constructed in Equation (14) is built on several key assumptions that ensure its theoretical validity and practical relevance. These assumptions include the following:

1. The observations are sampled from independent Bernoulli trials with a fixed number of trials n and constant success probability p .
2. Each trial is independent, with no correlation or clustering effects.
3. The Bayesian component assumes a non-informative Jeffreys' prior to stabilize inference near boundaries.

4. Adjusting the credible level within the specified grid (e.g., 98–99.9%) achieves near-nominal frequentist coverage without introducing bias.
5. Gamma regularization scales with sample size. This suggests that the term acts as a penalty, which stabilizes variance in small samples and diminishes as n grows, thus preserving asymptotic efficiency.
6. Transforming the intervals to the logit scale is assumed to improve numerical stability and symmetry near extreme proportions.
7. The sampling variance, Wilson margin penalty, and Gamma regularization are assumed to combine additively without violating consistency or efficiency.
8. For large sample sizes, the method assumes convergence to normality on the logit scale, thus ensuring consistency and efficiency as $n \rightarrow \infty$.

2.3.2. Proportionality Scaling of the Credible Width with the Variance

Let $\Delta_{q^*} = \frac{1-\phi}{2}$, the lower bound: $\ell^* = Q_{\frac{1-\phi}{2}}(a, b)$, and the upper bound: $u^* = Q_{1-\frac{(1-\phi)}{2}}(a, b)$; then $\ell^* = Q_{q_0-\Delta_{q^*}}$ and $u^* = Q_{q_0+\Delta_{q^*}}$. The Beta quantile function $Q_{q^*}(a, b)$ is implicitly defined as

$$Q_{q^*}(a, b) = \int_0^{Q_{q^*}(a, b)} f^*(t) dt, \text{ where } f^*(t) = \frac{t^{a-1}(1-t)^{b-1}}{B(a, b)} \quad (15)$$

Applying first-order Taylor expansion to Equation (15) yields

$$Q_{q_0 \pm \Delta_{q^*}} \approx Q_{q_0} \pm \left. \frac{dQ_{q^*}}{dq^*} \right|_{q_0} \cdot \Delta_{q^*} \text{ and } W^* \approx 2 \left. \frac{dQ_{q^*}}{dq^*} \right|_{q_0} \cdot \Delta_{q^*} \quad (16)$$

The quantile derivative function is obtained as

$$\frac{dQ_{q^*}}{dq^*} = \frac{1}{f(Q_{q^*})} = \frac{B(a, b)}{Q_{q^*}^{a-1}(1-Q_{q^*})^{b-1}} \quad (17)$$

Therefore, after the substitution of $\Delta_{q^*} = \frac{1-\phi}{2}$, the width in Equation (12) becomes

$$W^* \approx (1-\phi) \left. \frac{dQ_{q^*}}{dq^*} \right|_{q_0} \quad (18)$$

Equation (18) implies that the width is proportional to $(1-\phi)$, thus increasing ϕ reduces the interval width linearly. To further show that the width is proportional to the posterior variance, we transform the interval to logit scale by letting

$$\ell(q) = \log\left(\frac{Q_{q^*}}{1-Q_{q^*}}\right)$$

The application of the first-order Taylor expansion to $\ell(q)$ yields

$$\ell(q_0 \pm \Delta_{q^*}) \approx \ell(q_0) \pm \left. \frac{d\ell}{dq^*} \right|_{q_0} \cdot \Delta_{q^*},$$

where at $Q^* = q_0$ and

$$\frac{d\ell}{dq} = \left(\frac{d\ell}{dQ^*} \right) \left(\frac{dQ^*}{dq^*} \right) = \left(\frac{1}{Q^*(1-Q^*)} \right) \left(\frac{dQ^*}{dq^*} \right) = \frac{B(a, b)}{Q^{a^*}(1-Q^{b^*})}$$

Therefore, by the substitution of $\frac{d\ell}{dq}$ and $\Delta_{q^*} = \frac{1-\gamma}{2}$, the width on the logit scale becomes

$$W_{Logit}^* \approx (1 - \phi) \frac{B(a, b)}{Q^a * (1 - Q^b *)} \quad (19)$$

By the definition of margin of error, our adjusted margin of error on the logit scale becomes

$$MOE_{logit} \approx (1 - \phi) \frac{B(a, b)}{Q^a * (1 - Q^b *)} \sqrt{\frac{1}{n\hat{p}(1 - \hat{p}) + \epsilon} + \left(\frac{\theta}{\sqrt{n}}\right)^2 + \left(z\sqrt{\frac{\hat{p}(1 - \hat{p})}{n} + \frac{z^2}{4n^2}}\right)^2} \quad (20)$$

$$\Rightarrow W_{Logit}^* \propto (1 - \phi) \sqrt{\frac{1}{n\hat{p}(1 - \hat{p}) + \epsilon} + \left(\frac{\theta}{\sqrt{n}}\right)^2 + \left(z\sqrt{\frac{\hat{p}(1 - \hat{p})}{n} + \frac{z^2}{4n^2}}\right)^2} \quad (21)$$

From Equation (21), the logit-scaled half-width is inversely proportional to ϕ (tail probability), hence it scales proportionally to the asymptotic variance or the uncertainty.

2.3.3. Asymptotic Properties

The margin of error of the Jeffreys' prior (through the Bernstein–von Mises theorem), the model-based method, and the Wilson method are asymptotically normal, efficient, and consistent. Therefore, by default, our blended method inherits these properties. In the preceding sub-sections, we assess these properties to establish the underlying asymptotic and other theoretical properties of the blended method.

(a) Asymptotic efficiency

From Definition 1, the blended variance is obtained as

$$var_{blend} = \underbrace{\frac{1}{n\hat{p}(1 - \hat{p}) + \epsilon}}_{\text{sampling}} + \underbrace{\frac{\theta^2}{n}}_{\text{Gamma}} + \underbrace{\left(z\sqrt{\frac{\hat{p}(1 - \hat{p})}{n} + \frac{z^2}{4n^2}}\right)^2}_{\text{Wilson}} \quad (22)$$

From this variance, as $n \rightarrow \infty$, each term behaves as follows: sampling term $\sim O\left(\frac{1}{n}\right)$, Gamma term $\sim O\left(\frac{1}{n}\right)$, and Wilson term $\sim O\left(\frac{1}{n}\right)$. Therefore, following from the Hajek–Le Cam theorem of local asymptotic normality and influence functions [20], $var_{blend} \rightarrow 0$ as $n \rightarrow \infty$. Furthermore $var_{blend} \sim O\left(\frac{1}{n}\right) \Rightarrow SE_{Blend} \sim O\left(\frac{1}{\sqrt{n}}\right)$, hence

$$\delta = h\left(\frac{SE_{Blend}}{\max(SE_{Blend}, \epsilon)}\right) \sim O\left(\frac{1}{\sqrt{n}}\right) \quad (23)$$

which implies that the subsequent interval width shrinks at rate $O\left(\frac{1}{\sqrt{n}}\right)$, thus confirming asymptotic efficiency.

(b) Asymptotic consistency

By the weak law of large numbers, let

$$\hat{p} = \frac{k}{n} \rightarrow p \text{ be the sample proportion.}$$

Therefore, for any $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|\hat{p} - p| > \epsilon) = 0.$$

Furthermore, since we have shown that $\delta \rightarrow 0$ as $n \rightarrow \infty$, $\text{logit}(p) = \log\left(\frac{1}{1+e^{-p}}\right)$ is differentiable, and Lipschitz on $(0, 1)$, then

$$L_n, U_n \rightarrow \text{logit}(p) \Rightarrow c_n \rightarrow \text{logit}(p),$$

where $L_n = \max\left(0, \frac{1}{1+e^{-(c-\delta)}}\right)$, $U_n = \min\left(1, \frac{1}{1+e^{-(c+\delta)}}\right)$, and $c_n = \frac{\text{Logit}(L_n) + \text{Logit}(U_n)}{2}$. Thus, the interval converges to the true population proportion as $n \rightarrow \infty$.

(c) Asymptotic normality

Let

$$Z_n = \frac{c_n - \text{logit}(p)}{SE_{blend}} \quad (24)$$

Since $c_n \rightarrow \text{logit}(p)$ in probability and $SE_{blend} \rightarrow 0$, the numerator $c_n \rightarrow \text{logit}(p)$ behaves like the mean of an *i.i.d.* terms (via beta posterior), so, by the central limit theorem, $Z_n \xrightarrow{d} N(0, 1)$. Hence,

$$c_n \sim N(\text{logit}(p), \text{var}_{blend}). \quad (25)$$

This confirms asymptotic normality in logit space. From the blended variance definition 1, it can be shown from leading order analysis that

$$\text{var}_{blend} \sim \frac{A}{n} \text{ where } A = \frac{1}{\hat{p}(1-\hat{p})} + \theta^2 + z^2 \hat{p}(1-\hat{p})$$

Therefore, following from the asymptotic normality of the observed c_n ,

$$c_n \sim N\left(\text{logit}(p), \frac{A}{n}\right) \quad (26)$$

If we let $g(x) = \frac{1}{1+e^{-x}}$ and $g'(x) = \left(\frac{1}{1+e^{-x}}\right)\left(1 - \frac{1}{1+e^{-x}}\right)$, then by the delta method, if $X_n \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right)$, then

$$g(X_n) \xrightarrow{d} N\left(g(\mu), \frac{[g'(\mu)]^2 \sigma^2}{n}\right)$$

Therefore,

$$L_n, U_n \xrightarrow{d} N\left(p, \frac{[g(\text{logit}(p))]^2 A}{n}\right) \quad (27)$$

From the results in Equation (27), the unscaled interval limits exhibit asymptotic normality that converges in distribution to normality. This implies increasing concentration around the true parameter p as the sample size grows. This classical form of asymptotic behavior confirms the consistency of the interval bounds and validates their use as efficient estimators. Moreover, it highlights that confidence intervals become progressively narrower and more accurate in large samples, reinforcing the appropriateness of normal approximations for inference, particularly in transformed parameter spaces such as the logit scale.

3. Simulation Design

3.1. Parameters and Design

Let $k \in \{0, 1, \dots, n\}$ denote the number of observed successes in a binomial trial of size $n \in \mathbb{N}$, with estimated proportion $\hat{p} = \frac{k}{n}$, where $\hat{p} \in (0, 1)$ is the true binomial probability. For each combination of $\hat{p}$ and n , we perform M simulation replications following burn-in samples B . Each replication $i \in \{1, \dots, M + B\}$ consists of a Bernoulli trial with Jeffreys' prior-based posterior inference and three interval construction methods.

For each $i \in \{1, \dots, M + B\}$ and a combination of k and n

Step 1: Compute

$$k^{(i)} \sim \text{Binomial}(n, p) \Rightarrow \hat{p}^{(i)} = \frac{k^{(i)}}{n}$$

Step 2: Compute the posterior shape parameters for the Jeffreys' prior $\text{Beta}(0.5, 0.5)$ such that

$$a^{(i)} = 0.5 + k^{(i)}, b^{(i)} = 0.5 + n - k^{(i)}$$

Then, construct central $(1 - \alpha) \times 100\%$ posterior interval as

$$\tilde{L}^{(i)} = Q_{\frac{\alpha}{2}}(a^{(i)}, b^{(i)}), \tilde{U}^{(i)} = Q_{1-\frac{\alpha}{2}}(a^{(i)}, b^{(i)})$$

Step 3: Given $z = \Phi^{-1}(1 - \frac{\alpha}{2})$, construct the Wilson interval score as follows. First the center and the margin are calculated, respectively, as follows:

$$\tilde{c}^{(i)} = \frac{\hat{p}^{(i)} + \frac{z^2}{2n}}{1 + \frac{z^2}{n}} \text{ and } \tilde{m}^{(i)} = \frac{z}{1 + \frac{z^2}{n}} \sqrt{\frac{\hat{p}^{(i)}(1 - \hat{p}^{(i)})}{n} + \frac{z^2}{4n^2}}$$

Then, construct the interval as

$$[\max(0, \tilde{c}^{(i)} - \tilde{m}^{(i)}), \min(1, \tilde{c}^{(i)} + \tilde{m}^{(i)})]$$

Step 4: Construct the lower and upper bounds for the Logit margin of error. Using Jeffreys' prior, compute the lower and upper bounds as follows:

$$\tilde{L}^{(i)} = Q_{1-\frac{\phi}{2}}(a^{(i)}, b^{(i)}), \tilde{U}^{(i)} = Q_{1-\frac{1-\phi}{2}}(a^{(i)}, b^{(i)})$$

Then, transform the bounds to logit scale as indicated below

$$\tilde{\ell}_L^{(i)} = \log\left(\frac{\tilde{L}^{(i)}}{1 - \tilde{L}^{(i)}}\right), \tilde{\ell}_U^{(i)} = \log\left(\frac{\tilde{U}^{(i)}}{1 - \tilde{U}^{(i)}}\right)$$

$$\tilde{c}^{(i)} = \frac{\tilde{\ell}_L^{(i)} + \tilde{\ell}_U^{(i)}}{2}, \tilde{h}^{(i)} = \frac{\tilde{\ell}_U^{(i)} - \tilde{\ell}_L^{(i)}}{2}$$

$$\tilde{\delta}^{(i)} = \left(\frac{\text{Logit}_{\tilde{U}^{(i)}} - \text{Logit}_{\tilde{L}^{(i)}}}{2}\right) \left(\frac{SE_{\text{Blend}}}{\max(SE_{\text{Blend}}, \epsilon)}\right), \tilde{SE}_{\text{Blend}} = \sqrt{\text{var}_{\text{Blend}}}$$

$$\hat{r}_L^{(i)} = \max\left(0, \frac{1}{1 + e^{-(\tilde{c}^{(i)} - \tilde{\delta}^{(i)})}}\right) \cup \tilde{r}^{(i)} = \min\left(1, \frac{1}{1 + e^{-(\tilde{c}^{(i)} - \tilde{\delta}^{(i)})}}\right)$$

$$L'^{(i)} = \max\left(0, \frac{1}{1 + e^{-(\tilde{c}^{(i)} - \tilde{\delta}^{(i)})}}\right), U'^{(i)} = \min\left(1, \frac{1}{1 + e^{-(\tilde{c}^{(i)} - \tilde{\delta}^{(i)})}}\right)$$

Note that at each iteration, ϕ is tuned as indicated in Section 3.2.

Step 5: Computing the coverage probability and the credible width.

At each iteration M in Steps 2 to 4, calculate the coverage probability and width as follows:

$$\text{coverage probability} = \frac{1}{M} \sum_{i=1}^M I(L_i \leq p \leq U_i) \text{ and}$$

$$\text{width} = \frac{1}{M} \sum_{i=1}^M (U_i - L_i),$$

where I is the indicator function defined by

$$I = \begin{cases} 1 & \text{if } p \in [U_i - L_i] \\ 0 & \text{otherwise} \end{cases},$$

and L_i, U_i represent the corresponding lower and upper credible intervals for each of the interval estimation methods.

3.2. Tuning Procedure

The tuning process aims to strike a balance between achieving high coverage (at least 95%) and maintaining interval efficiency by minimizing width to ensure precise estimates without sacrificing coverage. To achieve this balance, the algorithm evaluates candidate confidence levels through simulation and then selects the credible level that meets the coverage threshold while producing the narrowest or optimal intervals. The tuning process is discussed below in detail.

Step 1: Defining the grid for the credible levels

The algorithm begins by specifying a set of candidate credible levels (ϕ), which represent the confidence level used in the Bayesian component of the blended interval. The grid typically spans from 0.98 to 0.999 in small increments, i.e., $\phi \in \{0.98, 0.981, \dots, 0.999\}$.

Step 2: Simulations of blended intervals for each credible level

For each candidate level ϕ in the grid, perform Monte Carlo simulation with $N = 6000$ trials. In each trial, generate a binomial sample of size n with success probability $\hat{p}$ and then compute the blended confidence interval as discussed in Step 4 of Section 3.1. This step evaluates how well each candidate level performs under repeated sampling.

Step 3: Computation of the width for each simulation

For each candidate level ϕ , calculate the width by subtracting the lower bound from the upper bound.

Step 4: Selecting Optimal Level

Choose the level ϕ that satisfies:

$$\phi^* = \underset{\phi \in \text{grid}}{\text{ar min median Width}}(\phi)$$

This ensures the interval is both accurate (coverage $\geq 95\%$) and as narrow as possible.

NB: After selecting ϕ , the algorithm performs a quick validation with 500 trials. If the coverage falls below 0.95 during this check, it overrides the tuned level with a fixed value of $\phi = 0.985$. Because simulations can occasionally underestimate variability, especially in edge cases such as small sample sizes or extreme probabilities, the override rule adds robustness by enforcing a fallback level of 0.985 whenever quick validation shows coverage below 95%. This safeguard ensures that the method remains reliable even under extreme or rare conditions.

For the tuning process, the grid $\phi \in \{0.98, 0.985, \dots, 0.999\}$ is selected for three reasons. Firstly, the primary goal is to achieve at least 95% coverage, which corresponds to confidence levels near 0.95. Starting the grid at 0.98 allows the algorithm to explore slightly higher levels in case the blended method tends to overcover and produce intervals that are unnecessarily wide. Extending the grid up to 0.999 provides a safety margin for extreme cases, such as very small sample sizes or probabilities near 0 or 1, where higher levels may be required to maintain coverage. Secondly, using small increments (e.g., 0.001) ensures fine resolution for optimization, enabling the algorithm to identify the level that minimizes interval width while meeting the coverage requirement. Large jumps could

miss the optimal trade-off between coverage and efficiency. Third, the wide range of levels makes the method adaptable across different scenarios, as coverage and width vary significantly with sample size and true probability. For example, large samples stabilize coverage near 0.95, while small samples or extreme probabilities often require higher levels to avoid under-coverage. Finally, the range from 0.98 to 0.999 strikes a balance between accuracy and computational feasibility. It is broad enough to capture variability across different situations but narrow enough to keep the grid search practical. In summary, this grid ensures flexibility, precision, and robustness in tuning the blended method for diverse data situations.

3.3. The Choice of Gamma

The gamma regularization parameter was introduced to stabilize variance estimates in small samples and near-boundary proportions. Its calibration follows an inverse relationship with sample size, ensuring that the penalty diminishes as n grows, thereby preserving asymptotic efficiency. Empirically, gamma can be tuned through grid search across simulation regimes to minimize coverage deviation from 95% while controlling interval width. However, this approach may result in an inadequate gamma value being selected because it optimizes based on empirical performance in a finite simulation set, which may not fully represent all real-world scenarios; thus, in this study, we fixed gamma at 1. The choice of this value is considered optimal because of how it interacts with the blended variance components and the asymptotic properties. The choice balances stabilization and flexibility and has theoretical justification.

The gamma acts as a penalty term to prevent instability in small samples or near-boundary proportions; hence, if $\gamma < 1$ (too small), the regularization effect is weak and the interval width may shrink excessively in ultra-rare events, leading to under-coverage, while on the other hand, if $\gamma > 1$ (too large), the penalty dominates, which makes the interval unnecessarily wide and conservative, thus reducing efficiency. $\gamma = 1$ provides an ideal scaling factor that neither inflates nor deflates the blended variance disproportionately, thus ensuring that the regularization term contributes meaningfully without overshadowing the sampling variance or Wilson margin. Furthermore, $\gamma = 1$ aligns with the principle of unit-information priors in Bayesian regularization, where the penalty corresponds to one “pseudo-observation,” which keeps the method asymptotically unbiased and efficient. In particular, if $\gamma > 1$, the penalty term becomes larger than $1/n$; thus, it effectively adds the influence of multiple pseudo-observations, which makes the intervals overly conservative. On the other hand, if $\gamma < 1$, the penalty term is less than $1/n$; thus, it contributes less than the equivalent of one pseudo-observation, which may fail to provide sufficient stabilization in small samples. This is specifically consistent with Bayesian decision theory, as it aligns with the principle that priors should reflect minimal but meaningful information, thus ensuring coherent inference across different sample sizes.

4. Results and Discussion

The goal of this section is to estimate the binomial proportion p using our proposed adaptive method to improve coverage and precision of p for rare and extreme events. The chosen ranges of true proportions (0.00001–0.99999) and sample sizes (5–10,000) were selected to reflect realistic conditions encountered in rare-event and extreme-probability scenarios across multiple fields. Ultra-rare probabilities such as 10^{-5} or smaller occur in safety-critical engineering and pharmacovigilance, where failure rates or adverse events are extremely low. Rare disease prevalence often falls between 1 in 10,000 and 1 in 1,000,000, which aligns with the lower bound of our range, while proportions near 0.99999 represent

high-reliability systems or near-certain outcomes in reliability studies. The sample size range of 5 to 10,000 captures both ends of practical study designs. For example, early-phase clinical trials and rare-disease studies often enroll fewer than 50 participants, sometimes as few as 12, whereas large-scale epidemiological surveys and public health programs routinely include tens of thousands of individuals to detect low-prevalence conditions. This broad range ensures that the proposed method is evaluated under different proportions that are representative in healthcare, pharmacovigilance, behavioral sciences, and reliability engineering. Rare disease prevalence thresholds and ultra-rare definitions (≤ 1 per 1,000,000) are outlined in Orphanet (a European database for rare diseases and orphan drugs) and in European Medicines Agency guidelines. In reliability engineering, standards often target failure probabilities as low as 10^{-5} to 10^{-6} . Similarly, survey designs from the World Health Organization and the Centers for Disease Control and Prevention typically require sample sizes exceeding 10,000 to capture low-prevalence conditions.

4.1. Simulation Results

We simulate 5000 samples, where the first 1000 are used as burn-in samples to estimate proportions of rare and extreme events using our method and then compare its performance to Wilson and Jeffreys' prior in terms of coverage probability and precision. Detailed R code used for the simulations, along with practical applications and the datasets, are provided as Supplementary Material. Refer to the Supplementary Material section following Section 5 for additional details.

The simulation results are summarized in Tables 1–4 and Figures 1–3.

Table 1. Performance comparison of interval estimation methods (ultra-rare-event proportions, $p \leq 0.0001$).

Sample Size	Jeffreys' Coverage	WS Coverage	Blend Coverage	Jeffreys' Width	WS Width	Blend Width	Blend Level
$p = 0.00001$							
5	0.0000	1.0000	1.0000	0.37928	0.43448	0.49251	0.985
10	0.0000	1.0000	1.0000	0.21715	0.27753	0.27627	0.980
50	0.9992	0.9992	0.9992	0.04878	0.07137	0.06392	0.980
100	0.9986	0.9986	0.9986	0.02477	0.03702	0.03258	0.980
1000	0.9888	0.9888	0.9888	0.00253	0.00385	0.00334	0.980
5000	0.9478	0.9478	0.9478	0.00052	0.00079	0.00069	0.980
10,000	0.9036	0.9036	0.9918	0.00027	0.00040	0.00036	0.980
$p = 0.0001$							
5	0.9990	0.9990	0.9990	0.37951	0.43464	0.46742	0.980
10	0.9996	0.9996	0.9996	0.21721	0.27758	0.27633	0.980
50	0.9958	0.9958	0.9958	0.04891	0.07147	0.06406	0.980
100	0.9904	0.9904	0.9904	0.02493	0.03714	0.03276	0.980
1000	0.9052	0.9052	0.9950	0.00271	0.00399	0.00354	0.980
5000	0.9878	0.9124	0.9878	0.00069	0.00092	0.00088	0.980
10,000	0.9802	0.9172	0.9802	0.00043	0.00053	0.00054	0.980

4.1.1. Coverage Consistency

Across all regimes, from ultra-rare to extreme, the intervals of our blended method consistently maintain coverage probabilities close to or exceeding the nominal 95% level. This is especially evident as sample size increase, where the blended method often matches or surpasses both the Jeffreys' and Wilson intervals. For instance, at $p = 0.00001$ and $n = 10,000$, the blended interval achieves 99.18% coverage compared to the 90.36% under-coverages for Jeffreys' and Wilson. Even in extreme tail scenarios (e.g., $p = 0.99999$), the blended method maintains coverage above 95% at moderate to large n , demonstrating

robustness under boundary conditions. Jeffreys' intervals tend to underperform at small n and extreme p , while Wilson intervals maintain coverage but are often overcompensated with wider bounds. The results are visualized in Figure 1, where the blended method tends to consistently stay near or above the 95% coverage probability.

Table 2. Performance comparison of interval estimation methods (Rare events, $0.001 < \text{True proportion} \leq 0.01$).

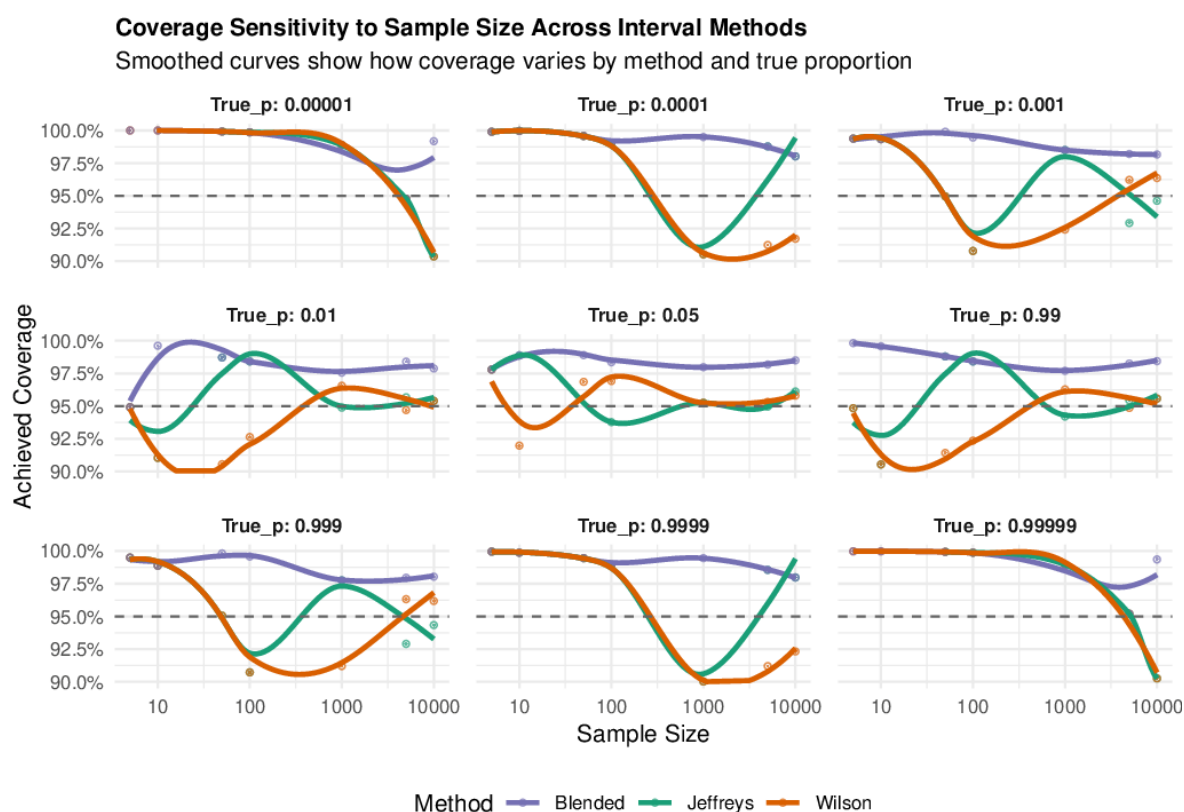
Sample Size	Jeffreys' Coverage	WS Coverage	Blend Coverage	Jeffreys' Width	WS Width	Blend Width	Blend Level
$p = 0.001$							
5	0.9938	0.9938	0.9938	0.38069	0.43544	0.46856	0.980
10	0.9934	0.9934	0.9934	0.21816	0.27825	0.27732	0.980
50	0.9496	0.9496	0.9990	0.05073	0.07289	0.07089	0.985
100	0.9078	0.9078	0.9946	0.02666	0.03851	0.03468	0.980
1000	0.9850	0.9242	0.9850	0.00423	0.00524	0.00531	0.981
5000	0.9292	0.9622	0.9820	0.00177	0.00188	0.00212	0.980
10,000	0.9462	0.9636	0.9816	0.00125	0.00128		
$p = 0.01$							
5	0.9492	0.9492	0.9492	0.39088	0.44234	0.47847	0.980
10	0.9104	0.9104	0.9962	0.23121	0.28753	0.29092	0.980
50	0.9872	0.9056	0.9872	0.06687	0.08560	0.08489	0.981
100	0.9842	0.9264	0.9842	0.04222	0.05113	0.05212	0.980
1000	0.9488	0.9656	0.9756	0.01236	0.01273	0.01473	0.980
5000	0.9568	0.9470	0.9840	0.00552	0.00555	0.00656	0.980
10,000	0.9540	0.9540	0.9788	0.00390	0.00391	0.00463	0.980

Table 3. Performance comparison of interval estimation methods (normal events ($0.01 < \text{true proportion} \leq 0.99$)).

Sample Size	Jeffreys' Coverage	WS Coverage	Blend Coverage	Jeffreys' Width	WS Width	Blend Width	Blend Level
$p = 0.05$							
5	0.9780	0.9780	0.9780	0.43260	0.47068	0.51909	0.980
10	0.9888	0.9198	0.9888	0.28586	0.32669	0.34812	0.980
50	0.8922	0.9686	0.9890	0.11908	0.12887	0.14409	0.981
100	0.9378	0.9692	0.9836	0.08451	0.08831	0.10076	0.980
1000	0.9526	0.9526	0.9798	0.02699	0.02711	0.03204	0.980
5000	0.9496	0.9532	0.9818	0.01209	0.01210	0.01435	0.980
10,000	0.9612	0.9580	0.9850	0.00854	0.00854	0.01014	0.980
$p = 0.99$							
5	0.9484	0.9484	0.9982	0.39115	0.44253	0.50385	0.985
10	0.9054	0.9054	0.9958	0.23202	0.28811	0.2950	0.981
50	0.9880	0.9142	0.9880	0.06617	0.08504	0.08411	0.981
100	0.9842	0.9236	0.9842	0.04191	0.05088	0.05177	0.980
1000	0.9420	0.9628	0.9770	0.01237	0.01273	0.01474	0.980
5000	0.9544	0.9486	0.9826	0.00551	0.00554	0.00654	0.980
10,000	0.9558	0.9558	0.9844	0.00389	0.00391	0.00462	0.980

Table 4. Performance comparison of interval estimation methods (extreme events, true proportion > 0.99).

Sample Size	Jeffreys' Coverage	WS Coverage	Blend Coverage	Jeffreys' Width	WS Width	Blend Width	Blend Level
$p = 0.999$							
5	0.9950	0.9950	0.9950	0.38042	0.43525	0.46830	0.980
10	0.9888	0.9888	0.9888	0.21888	0.27876	0.27807	0.980
50	0.9506	0.9506	0.9982	0.05071	0.07287	0.07088	0.985
100	0.9072	0.9072	0.9958	0.02665	0.03850	0.03467	0.980
1000	0.9776	0.9120	0.9776	0.00430	0.00530	0.00533	0.980
5000	0.9290	0.9632	0.9796	0.00176	0.00187	0.00211	0.980
10,000	0.9434	0.9618	0.9804	0.00124	0.00128	0.00150	0.981
$p = 0.9999$							
5	0.9996	0.9996	0.9996	0.37937	0.43454	0.46728	0.980
10	0.9988	0.9988	0.9988	0.21733	0.27766	0.27646	0.980
50	0.9946	0.9946	0.9946	0.04896	0.07151	0.06411	0.980
100	0.9896	0.9896	0.9896	0.02495	0.03716	0.03278	0.980
1000	0.9002	0.9002	0.9946	0.00272	0.00400	0.00360	0.981
5000	0.9856	0.9120	0.9856	0.00069	0.00092	0.00087	0.980
10,000	0.9798	0.9232	0.9798	0.00043	0.00053	0.00054	0.981
$p = 0.99999$							
5	0.0000	0.9998	0.9998	0.37933	0.43451	0.49256	0.985
10	0.0000	0.9998	0.9998	0.21718	0.27755	0.2763	0.980
50	0.9994	0.9994	0.9994	0.04877	0.07137	0.06391	0.980
100	0.9988	0.9988	0.9988	0.02476	0.03701	0.03258	0.980
1000	0.9896	0.9896	0.9896	0.00253	0.00384	0.00333	0.980
5000	0.9520	0.9520	0.9520	0.00052	0.00078	0.00069	0.980
10,000	0.9026	0.9026	0.9936	0.00027	0.00040	0.00036	0.980

**Figure 1.** Coverage sensitivity to sample size across interval methods.

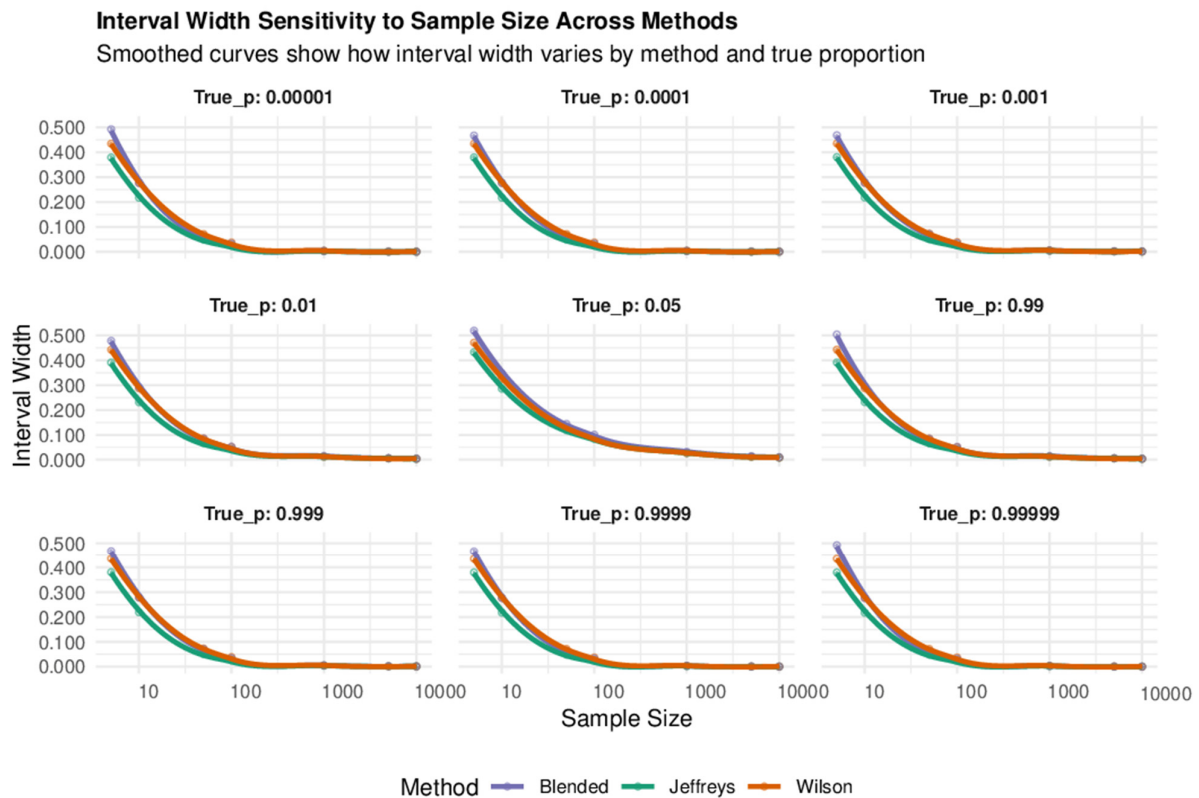


Figure 2. Interval width sensitivity to sample size across interval methods.

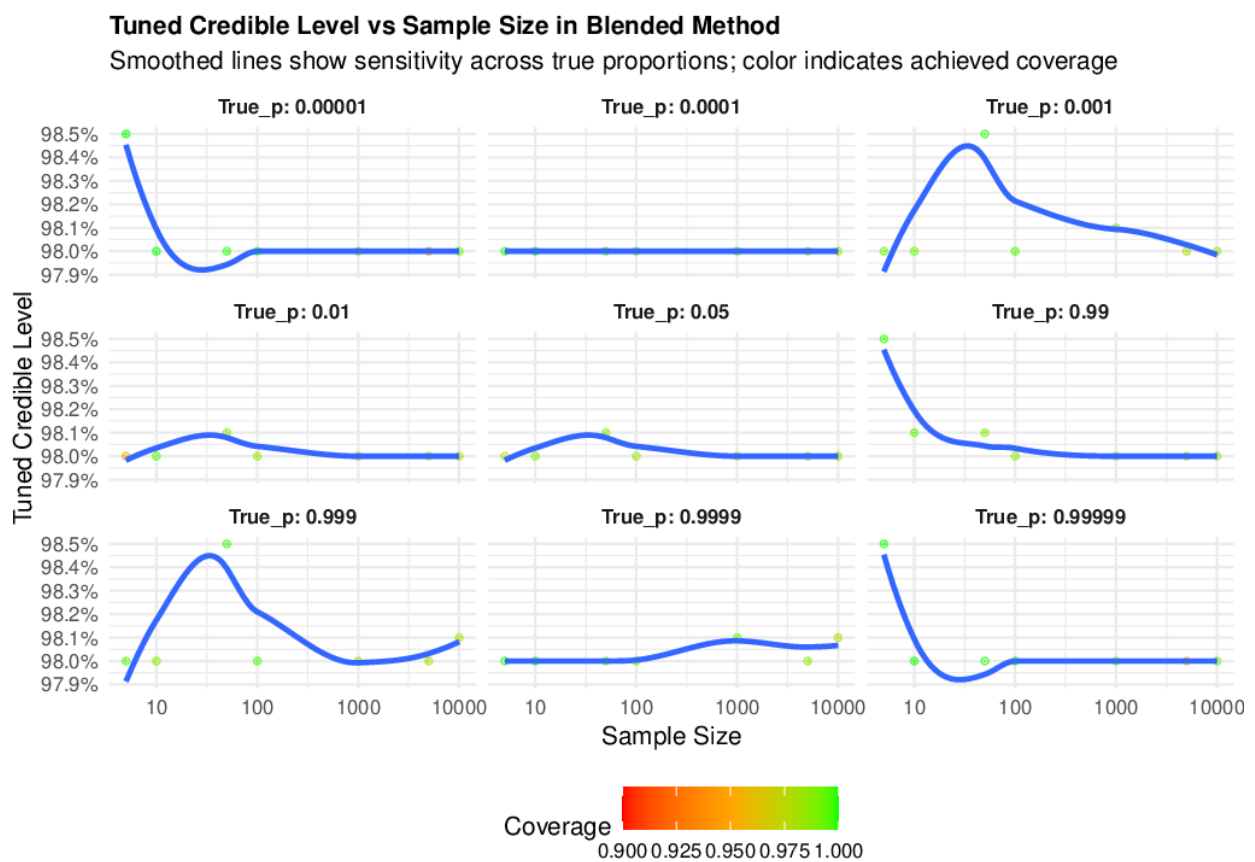


Figure 3. Sensitivity of tuned credible levels across sample sizes and proportions.

4.1.2. Interval Width and Efficiency

In Tables 1–6 and Figure 2, it is observed that the blended intervals tend to be wider at very small sample sizes (e.g., $n = 5$ to 10), which is expected due to the added gamma penalty and conservative tuning of the credible level. However, these widths remain practical and shrink rapidly as the sample size increases. At all levels, the blended intervals are competitive compared to both the Jeffreys' and Wilson intervals while preserving high coverage. For instance, at $p = 0.001$ and $n = 5$, the blended interval width is 0.46856, compared to 0.38069 for Jeffreys' and 0.43544 for Wilson, while at $n = 10,000$, for the same proportion, the blended interval width is 0.00463, compared to 0.00390 for Jeffreys' and 0.00391. This improved efficiency while maintaining high coverage reflects the adaptiveness of the blended method, which balances precision and coverage through dynamic tuning and variance blending.

4.1.3. Impact of True Probability (p)

At low true probabilities (e.g., $p = 0.00001$ to 0.001), the blended intervals outperform the Jeffreys' and Wilson intervals in coverage with complete width as the sample size increases. Jeffreys' occasionally fails to capture the true value with conservative width at small n due to its symmetric prior, while the Wilsons interval also fails to maintain the nominal coverage but with relatively smaller width. Our method, by contrast, adapts to the rarity of the event, ensuring that the interval expands appropriately to maintain coverage without excessive conservatism. As p increases to moderate levels (e.g., 0.05), all methods perform well, but the blended interval remains conservative with the interval width. At high p (e.g., 0.99 and above), the blended method continues to excel, yielding conservative intervals, and maintains coverage even when Jeffreys' interval begins to under-cover.

4.1.4. Sample Size Sensitivity

Smaller sample sizes ($n = 5, 10$) exhibit more variability and wider intervals across all methods, as seen in Tables 1–4 and Figures 2 and 3. However, the blended approach maintains high coverage even for small sample sizes, which indicates robustness in small-sample data. For example, at $p = 0.01$ and $n = 10$, the blended interval achieves 99.62% coverage with a conservative width of 0.33358, thus outperforming the Jeffreys' and Wilson intervals because of their under-coverages. As sample size increases ($n \geq 100$), all methods stabilize, but the blended interval consistently maintains coverage above the nominal while maintaining conservative or narrower width. This scalability confirms that the method remains efficient across sample sizes and is particularly well-suited for large-sample studies or simulations. Notably, the blended interval avoids the erratic discreteness-induced jumps seen in classical intervals when n is small, which is one of the strengths of the blended method.

4.1.5. Sensitivity and Robustness of the Tuning Parameter on Performance

Coverage Calibration: The credible level is highly responsive to the interaction between sample size and true proportion. For ultra-rare events (e.g., $p = 0.00001$), the method consistently selects high credible levels (e.g., 99%) to compensate for data sparsity and ensure the interval expands sufficiently to include the true value. For instance, at $p = 0.00001$ and $n = 5$, the credible level is 0.985, yielding 100% coverage with a width of 0.49251. Similarly, for $p = 0.0001$ and $n = 10$, the level increases to 0.99 while maintaining full coverage. As the sample size increases or the true proportion becomes moderate, the tuned level decreases. At $p = 0.00001$ and $n = 1000$, the level drops to 0.98, still achieving 98.88% coverage with a much narrower width of 0.00334. For $p = 0.001$ and $n = 5$, the credible level is 0.98, and the method achieves 99.38% coverage with a width of 0.46856. This dynamic

tuning allows the method to self-correct, i.e., if initial tuning results in under-coverage, the algorithm escalates the level to 98.5% or higher to restore coverage. This adaptivity underpins the method's robustness across regimes where fixed-level intervals often fail.

Width Efficiency: Higher credible levels yield wider Bayesian bounds, which, after transformation and blending, produce more conservative intervals. The tuning algorithm balances this trade-off by selecting the credible level that still achieves the desired coverage. This ensures statistical efficiency by avoiding unnecessarily wide intervals. This behavior is evident across Tables 1–4. For example, at $p = 0.00001$ and $n = 1000$, the credible level is 0.98, producing a narrow interval (width = 0.00334) with 98.88% coverage. At $p = 0.0001$ and $n = 10000$, the level remains at 0.98, yielding 98.02% coverage with a width of 0.0005. For rare events like $p = 0.001$ and $n = 10$, the level remains at 0.98, producing an interval width of 0.27732 and 99.34% coverage. These examples illustrate how the method maintains conservative width while preserving or exceeding the target coverage.

Boundary Sensitivity: Near the boundaries ($p \approx 0$ or $p \approx 1$), credible level tuning becomes more aggressive. The logit transformation amplifies small differences near the extremes, necessitating wider initial bounds. For instance, at $p = 0.99999$ and $n = 5$, the credible level is 0.985, yielding 99.98% coverage with a width of 0.49256. At $p = 0.999$ and $n = 100$, the level is also 0.98, producing 99.58% coverage with a width of 0.03467. The gamma penalty and Wilson variance blending help stabilize the interval, but the credible level remains the primary mechanism for controlling coverage.

The multi-panel plot in Figure 3 illustrates how the blended method dynamically tunes credible levels across extreme binomial regimes, balancing robustness and sensitivity. For ultra-rare and near-certain proportions (e.g., $p = 0.00005$ or $p = 0.99999$), the credible level starts conservatively high at small sample sizes around 98.5% which ensures robustness against under-coverage. These tuning curves show steep initial declines that flatten as sample size grows, forming arcs that transition from convex to nearly linear beyond $n \approx 100$. This curvature reflects logit-scale transformation effects, where differences near boundaries are amplified, causing higher credible levels at low n . In contrast, mid-range proportions (e.g., $p = 0.05$) exhibit gentle slopes and early stabilization near 98%, indicating less aggressive tuning and geometric smoothness. The plot confirms symmetry across complementary proportions (e.g., $p \approx 0.001$ vs. $p \approx 0.999$), reinforcing the method's principled adaptiveness. Collectively, the steepness at extremes, flattening at large n , and mirrored trajectories demonstrate how the method maintains high coverage while adaptively reducing credible levels from 98.5% at $n = 5$ to 98% at $n \geq 10,000$ independent of fixed parameterization.

4.2. Practical Applications Using COVID-19 Data

To assess the performance of the blended method in comparison to the competing methods with applications to real life data, mortality and recovery rate data for the COVID-19 outbreak was used. Data from nine countries—Western Sahara, Ghana, South Africa, Australia, Brazil, Germany, France and the USA—were used. These data were selected to mirror rare, extreme nature, or small sample applications. We focused on mortality and recovery rates because they represent opposite extremes of the proportion spectrum in COVID-19 data. Mortality rates are typically low (often <0.01), making them rare-event scenarios that challenge conventional interval estimation methods. In contrast, recovery rates often approach near certainty (>0.99), placing them at the upper extreme of the probability scale. Evaluating these endpoints allows us to demonstrate the adaptability of the proposed method across the full range of binomial proportions from ultra-rare to near-certain outcomes. The data were sourced from <https://www.worldometers.info/coronavirus/> (accessed on 26 August 2025). For each country, mortality and recovery rate

were estimated using the blended method and the competing methods. The results are reported in Tables 5 and 6.

Table 5. Performance of interval estimation methods for COVID-19 mortality rates.

Country	Sample Size	True p	Jeffreys' Coverage	Wilson Coverage	Blended Coverage	Jeffreys' Width	Wilson Width	Blended Width
Western Sahara	10	0.10000	0.98733	0.92967	0.98733	0.34371	0.36878	0.40912
Ghana	171,889	0.00851	0.94733	0.94633	0.97900	0.00087	0.00087	0.00103
South Africa	4,076,463	0.02517	0.94633	0.94633	0.97700	0.00030	0.00030	0.00036
Australia	11,853,144	0.00206	0.94550	0.94500	0.97717	0.00005	0.00005	0.00006
Brazil	38,743,918	0.01836	0.94983	0.94983	0.97833	0.00008	0.00008	0.00010
Germany	38,828,995	0.00471	0.94933	0.94933	0.97800	0.00004	0.00004	0.00005
France	40,138,560	0.00418	0.95283	0.95283	0.98100	0.00004	0.00004	0.00005
USA	111,820,082	0.01091	0.95050	0.95050	0.97850	0.00004	0.00004	0.00005

Table 6. Performance of interval estimation methods for COVID-19 recovery rates.

Country	Sample Size	True p	Jeffreys' Coverage	Wilson Coverage	Blended Coverage	Jeffreys' Width	Wilson Width	Blended Width
Western Sahara	10	0.77778	0.98600	0.92683	0.98817	0.34513	0.36977	0.41099
Ghana	171,889	0.90000	0.95467	0.95317	0.98033	0.00087	0.00087	0.00103
South Africa	4,076,463	0.99148	0.95150	0.95150	0.98000	0.00038	0.00038	0.00045
Australia	11,853,144	0.95978	0.95267	0.95267	0.97750	0.00015	0.00015	0.00007
Brazil	38,743,918	0.93561	0.95067	0.95067	0.98183	0.00004	0.00004	0.00018
Germany	38,828,995	0.99529	0.95733	0.95717	0.98083	0.00004	0.00004	0.00005
France	40,138,560	0.99582	0.95067	0.95067	0.97900	0.00005	0.00005	0.00005
USA	111,820,082	0.98206	0.98600	0.92683	0.98050	0.00005	0.00005	0.00006

Across countries, the mortality rate intervals for the blended method consistently achieved better coverage, particularly in large samples and extremely low rates. For example, for Australia (true proportion = 0.00206), blended coverage reached 97.7%, exceeding the nominal 95% threshold and outperforming the Jeffreys' and Wilson intervals. Similar patterns were observed for South Africa and Germany, where the blended intervals maintained coverage above 97%. Although the width was comparatively wider in most cases, the blended method's prioritization of coverage over smaller width appears justified in rare events where underestimation poses substantial epidemiological risk. In small samples such as Western Sahara ($n = 10$), the blended intervals preserved high coverage (98.7%) while producing competitive interval width in comparison to the other methods. Jeffreys' intervals remained stable, whereas Wilson intervals exhibited under-coverage (92.9%), indicating sensitivity to small-sample-size bias and boundary truncation.

The recovery rate estimation revealed good performance across all methods in large samples. However, blended intervals again maintained coverage above 97% for Brazil and Germany. Interval widths converged across methods in large samples (width ≤ 0.000103), yet the blended intervals retained a marginal advantage in coverage. For the USA (true proportion = 0.98206), Wilson coverage (92.683%) dropped below the nominal threshold, while the blended and Jeffreys' intervals were above the nominal levels, which underscores the blended method's robustness near the boundary. The blended method's adaptive tuning emerged as a key strength. This dynamic behavior enables the method to maintain nominal coverage across heterogeneous epidemiological contexts without excessive conservatism.

In summary, the blended method is the most reliable for coverage, with an under-coverage rate of 1.6% (1 out of 63) compared to that of Jeffreys at 14.3% (9 cases) and Wilson at 12.7% (8 cases). The blended method consistently exceeded the nominal 95% level, even for rare events and boundary extremes. The demonstrated efficiency of the blended interval method in both rare-event and boundary-extreme events supports its integration into public health surveillance systems, particularly for real-time estimation of mortality and recovery rates. Its capacity to preserve coverage in sparse data environments makes it suitable for early disease outbreak detection and low-incidence monitoring, while its stability in high-proportion contexts ensures accurate reporting of recovery rates of disease outbreaks. Adoption of the blended method in reporting frameworks could enhance the interpretability and credibility of epidemiological risks, thus informing risk communication, resource allocation, and intervention prioritization. Moreover, its modular tuning architecture aligns with adaptive surveillance strategies, which allows for calibration sensitivity in evolving epidemic outbreaks.

4.3. Computational Cost Analysis

The proposed method is theoretically bounded by $O(M \times L)$ per (n, p) pair, where M is the number of simulation iterations and L is the number of candidate credible levels in tuning. Our configuration ($M = 1000$, $L = 20$, and 63 parameter pairs) results in approximately 1.26 million interval estimations. Despite this complexity, the observed runtime is far below the n^6 threshold. The largest configuration was completed in approximately 87.5 s (1.46 min) using about 20.15 MB of memory, while the smaller cases required only about 7.1 s (0.12 min) and 1.75 MB. These results confirm that the implementation is memory-efficient by leveraging R's optimized linear algebra routines, hence avoiding explicit n^6 scaling. Efficiency is achieved through vectorization, sparse operations, and caching, which makes the method practically robust for large-scale application.

4.4. Data-Based Tuning Rules for the Credible Level (ϕ): The Choice of Adaptive over Fixed Rule

In practice, choosing a single fixed value for ϕ instead of using a grid or adaptive rule is not recommended because it removes the flexibility that the method needs to perform well across diverse data conditions. Real-world datasets vary widely in sample size and observed proportion, and a fixed ϕ might work for moderate n and $\hat{p}$, but may fail for very small samples or rare events. For example, always using $\phi = 0.98$ could lead to under-coverage when $n = 10$ and $\hat{p} = 0.0001$ or produce unnecessarily wide intervals when $n = 10,000$ and $\hat{p} = 0.5$. The strength of the proposed method lies in its ability to adapt interval width to maintain coverage while minimizing unnecessary conservatism; thus, a fixed ϕ ignores this adaptivity and makes the method behave like a standard fixed-level Bayesian interval, which defeats the purpose. Coverage performance depends on both n and $\hat{p}$, and without tuning, intervals can fail to meet the nominal 95% coverage in rare-event scenarios, which is critical in applications such as epidemiology or reliability analysis. Practically, using a fixed ϕ could result in misleading intervals, either too narrow (under-coverage) or too wide (inefficient), which can lead to incorrect decisions in safety-critical applications like clinical trials or risk assessments.

To address this, we implement a grid search within the range $[0.98, 0.999]$. Preliminary simulation studies (not reported here) support the choice of this range, which is broad enough to capture variability across regimes yet narrow enough to remain computationally efficient. For practical reasons, searching across the entire $(0, 1)$ spectrum would add unnecessary computational cost without improving results, as most intervals of that range are irrelevant for coverage optimization. Although grid search is implemented within the codes, practitioners may need to specify the range for the grid search to suit the nature

of their data for better results; thus, based on the simulation results from Tables 1–4, the following grid selection rules (as reported in Table 7) are proposed as a guide.

Table 7. Adaptive grid ranges for tuning credible level (ϕ).

Regime	Sample Size (n)	True Proportion (p)	Grid Range for ϕ
Ultra-rare	$n \leq 10$	$p \leq 0.0001$	0.985–0.999
	$n \geq 50$	$p \leq 0.0001$	0.980–0.990
Rare	$n \leq 10$	$0.0001 < p \leq 0.01$	0.980–0.995
	$n \geq 50$	$0.0001 < p \leq 0.01$	0.980–0.990
Normal	$n \geq 50$	$0.01 < p \leq 0.99$	0.980–0.985
Extreme tails	Any other n	$p \geq 0.99$ or $p \leq 0.00001$	0.985–0.999

The algorithm starts at the lower boundary and increases incrementally (e.g., by 0.001) until a coverage $\geq 95\%$ is achieved. This ensures robustness for rare-event scenarios and efficiency for large-sample studies without unnecessary conservatism.

5. Conclusions

This study introduces an adaptive Bayesian variance-blending calibration framework for estimating binomial proportions of rare and extreme events. The performance of the method in comparison with the classical Jeffreys' and Wilson score intervals across different sample sizes ($n = 5$ to 10,000) under different true proportions was assessed through extensive simulations and real data applications. The blended method was shown to be robust and adaptive across sample sizes and true probabilities. It consistently achieves or exceeds the nominal 95% coverage, particularly in small sample events, rare events, and extreme events where traditional methods often struggle. Even at small sample sizes ($n = 5$), the method maintains high coverage and avoids erratic behavior typical of discrete classical intervals.

In terms of interval width, the blended method balances conservatism and efficiency through dynamic credible level tuning. While intervals are wider at small n due to conservative initialization, they shrink rapidly with increasing sample size. At moderate and large n , blended intervals are competitive in comparison to the Jeffreys' and Wilson intervals while preserving coverage. This adaptivity is driven by a tuning mechanism that adjusts the credible level based on the observed data and empirical coverage feedback. High credible levels are selected for ultra-rare events to ensure efficiency, while lower levels are used for moderate proportions to enhance precision.

Geometric analysis of the tuning curves demonstrates how the blended method adaptively tunes credible levels across binomial extremes by starting at higher values for small samples and gradually flattening into near-linear, symmetric trajectories as sample size increases, thus ensuring robust coverage and balanced sensitivity. These features reflect the method's ability to stabilize estimates near boundaries through aggressive credible level tuning and variance blending.

Theoretically, the blended method bridges Bayesian and frequentist paradigms by using empirical feedback to calibrate credible levels to guarantee near or above nominal coverage with narrow or conservative width. Its adaptivity and boundary sensitivity offer a competitive alternative to fixed-level intervals, particularly in rare and extreme events. Practically, the method's ability to adaptively shrink or expand based on sample size and true proportion makes it highly suitable for applications where sample size limitations and rare or extreme event proportions—such as rare disease prevalence, safety-critical system reliability, and early-phase clinical trials—render conventional interval estimation

inaccurate. Considering these findings, we recommend the use of the blended method for interval estimation of rare and extreme events, especially when sample sizes are limited or true proportions lie near the boundaries. Future work may extend this framework to multinomial or hierarchical settings, where adaptive tuning could further enhance inference under complex data structures.

Although the proposed adaptive Bayesian variance-blending framework demonstrates strong performance across rare and extreme event scenarios, several limitations remain. First, the method incurs additional computational overhead due to grid-based tuning of credible levels. Although the method is memory-efficient, its runtime can still be significant for very large datasets, which may pose challenges for real-time or large-scale applications in the absence of high computational power and memory.

Second, its performance is sensitive to the tuning of key parameters, such as the credible level. Incorrect calibration can lead to inflated interval widths or under-coverage. Furthermore, the approach assumes independence and that the trials follow a strict binomial form. Finally, while the method was validated using simulations and COVID-19 data, broader empirical testing across diverse domains, such as reliability engineering and rare disease prevalence, is needed to confirm generalizability.

Future research should address these limitations by developing automated tuning algorithms that leverage optimization and reduce computational cost. Extending the framework to handle complex data structures such as multinomial, hierarchical, and correlated models would enhance its application in many fields. In such models, incorporating mechanisms to account for overdispersion, such as beta-binomial or hierarchical Bayesian models, could also further improve robustness. Additionally, adaptive strategies for gamma calibration should be explored to allow dynamic adjustment based on sample size and event rarity, instead of fixing its value to 1. Practical directions include optimizing the method for real-time surveillance systems and creating open-source software implementations to facilitate adoption. Finally, comprehensive validation across multiple application areas will be essential to establish the method's reliability and scalability.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/math13243988/s1>, The codes and the data used have been included as a zipped file labelled Supplementary file-R codes.

Author Contributions: Conceptualization, A.A. and A.B.; methodology, A.A.; software, A.A.; validation, A.B. and D.M.; formal analysis, A.A.; resources, A.B. and D.M.; writing—original draft, A.A.; writing—review and editing, A.B. and D.M.; supervision, D.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used in this study are publicly available.

Acknowledgments: The authors would like to acknowledge UCDP-Sol Plaatje who funded the research visit to the Modern College of Business and Science, which resulted in the writing of this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Agresti, A.; Coull, B.A. Approximate is better than 'exact' for interval estimation of binomial proportions. *Am. Stat.* **1998**, *52*, 119–126. [[CrossRef](#)] [[PubMed](#)]
2. Wilson, E.B. Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.* **1927**, *22*, 209–212. [[CrossRef](#)]
3. Clopper, C.J.; Pearson, E.S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **1934**, *26*, 404–413. [[CrossRef](#)]

4. Agresti, A.; Gottard, A. Nonconservative exact small sample inference for discrete data. *Comput. Stat. Data Anal.* **2007**, *52*, 6447–6458. [[CrossRef](#)]
5. Brown, L.D.; Cai, T.T.; DasGupta, A. Interval estimation for a binomial proportion. *Stat. Sci.* **2001**, *16*, 101–133. [[CrossRef](#)]
6. Fan, Z.; Liu, D.; Chen, Y.; Zhang, N. Something Out of Nothing? The Influence of Double-Zero Studies in Meta-analysis of Adverse Events in Clinical Trials. *Stat. Biosci.* **2024**. [[CrossRef](#)]
7. Krishnamoorthy, K.; Lee, M.; Zhang, D. Closed-form fiducial confidence intervals for some functions of independent binomial parameters with comparisons. *Stat. Methods Med. Res.* **2017**, *26*, 43–63. [[CrossRef](#)] [[PubMed](#)]
8. Liu, J.; Shao, F.; Yang, J. Comparison of interval estimation for extreme event proportions based on exact, approximate and Bayesian approaches. *Biostat. Epidemiol.* **2025**, *9*. [[CrossRef](#)]
9. Newcombe, R.G. Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Stat. Med.* **1998**, *17*, 857–872. [[CrossRef](#)]
10. Yang, G.; Liu, D.; Wang, J.; Xie, M. Meta-analysis framework for exact inferences with application to the analysis of rare events. *Biometrics* **2016**, *72*, 1378–1386. [[CrossRef](#)] [[PubMed](#)]
11. Owen, M.; Burke, K. Binomial confidence intervals for rare events: Importance of defining margin of error relative to magnitude of proportion. *Am. Stat.* **2024**, *78*, 437–449. [[CrossRef](#)]
12. Blaker, H. Confidence curves and improved exact confidence intervals for discrete distributions. *Can. J. Stat.* **2000**, *28*, 783–798. [[CrossRef](#)]
13. Ogura, T.; Yanagimoto, T. Improvement of Bayesian credible interval for a small binomial proportion using logit transformation. *Am. J. Biostat.* **2018**, *8*, 1–8. [[CrossRef](#)]
14. Luo, Y.; Gao, C. Adaptive robust confidence intervals. *arXiv* **2024**, arXiv:2410.22647. [[CrossRef](#)]
15. Lyles, R.H.; Weiss, P.; Waller, L.A. Calibrated Bayesian credible intervals for binomial proportions. *J. Stat. Comput. Simul.* **2019**, *90*, 75–89. [[CrossRef](#)] [[PubMed](#)]
16. Jeffreys, H. *Theory of Probability*, 3rd ed.; Oxford University Press: London, UK, 1961; ISBN 9780198532019.
17. Bayarri, M.J.; Berger, J.O. The interplay of Bayesian and Frequentist analysis. *Stat. Sci.* **2004**, *19*, 58–80. [[CrossRef](#)]
18. Young, G.A.; Smith, R.L. *Essentials of Statistical Inference*; Cambridge University Press: New York, NY, USA, 2005; ISBN 978-0-521-83971-6.
19. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. [[CrossRef](#)]
20. Le Cam, L. *Asymptotic Methods in Statistical Decision Theory*; Springer: New York, NY, USA, 1986. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.