Academic Journal of Science and Technology, February, 2026, vol.8, No.1, p347-356

An Evaluation of Wald Confidence Intervals for the Binomial Proportion: Performance with and without Continuity Correction

Hassan M Rhoma[1*].

1. Department of Statistics, Faculty of Science, Tripoli University, Tripoli – Libya

ABSTRACT

This study provides a comprehensive evaluation of the performance of the standard Wald and Wald with continuity correction (Wald CC) confidence intervals for the binomial proportion, using simulation techniques across a wide range of population proportions (p) and sample sizes (n). Our evaluation is based on two key criteria: the probability of coverage and the expected interval width for each method. The results demonstrate the inadequacy of the standard Wald confidence interval, which fails to achieve the nominal 95% coverage across virtually all combinations of p and n. It is also highly unstable, even with large samples (e.g., n = 1500, 2000). In contrast, the Wald CC method performs significantly better. It achieves the nominal coverage (and exceeds it) for sample sizes $n \geq 200$ when the proportion p is not near the boundaries (i.e., for p between 0.1 and 0.9). Furthermore, its expected interval width is reasonable and decreases appropriately with increasing sample size. However, for extreme proportions ($p < 0.1$ or $p > 0.9$), both methods fail catastrophically, and their theoretical bounds often fall outside the [0,1] parameter space. We conclude with a strong recommendation against using the standard Wald interval in practice and advocate for the Wald CC method as a simple and effective alternative, particularly for medium and large samples with non-extreme proportions. For extreme proportions, more effective alternatives recommended.

KEYWORDS: Wald and Wald CC Interval; Proportion Confidence Interval; Coverage Probability; Expected Width.

Corresponding author: h.rhoma@uot.edu.ly

INTRODUCTION

The construction of a confidence interval (CI) for a binomial proportion, while seemingly straightforward, presents a classic and persistent challenge in statistical inference. Far from being a settled matter, the choice of an appropriate method has been the subject of extensive and ongoing debate for decades. An optimal confidence interval is ideally defined as the shortest set that maintains the nominal coverage probability. Ghosh (1979), in his seminal evaluation of binomial CIs, proposed three fundamental criteria for their assessment: coverage probability, interval width, and Neyman-shortness (also termed selectivity). This latter criterion measures the propensity of a CI to avoid covering incorrect parameter values, yet it is often overlooked in both methodological development and comparative evaluation.

The quest for a reliable interval has spurred significant research over the past half-century, resulting in a proliferation of alternative methods to the ubiquitous standard Wald (or asymptotic normal) interval. These alternatives have been the focus of numerous comparative studies. Vollset (1993) provided a detailed numerical comparison of eight different methods. Agresti and Coull (1998), noting the poor performance of the Wald interval, famously recommended their "adjusted Wald" interval as a simple and more accurate alternative for introductory teaching. In a more exhaustive review, Pires and Amado (2008) compiled and compared the properties of twenty distinct methods. These studies, among many others, have consistently demonstrated the severe shortcomings of the standard Wald interval, particularly its erratic coverage properties for small to moderate sample sizes and for proportions near the boundaries of 0 or 1.

Despite this well documented evidence, the standard Wald interval, defined as $\hat{p} \pm z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$, remains perilously widespread in applied research, textbook examples, and statistical software output, likely due to its compelling simplicity and direct analogy to intervals for a normal mean. A common but misguided justification for its use is the appeal to asymptotic theory, assuming it becomes adequate for "sufficiently large" samples. However, as Brown et al. (2001) rigorously argued in their comprehensive review, the performance of the Wald interval is "lamentably poor" and its path to the asymptotic limit is "shockingly erratic."

Given this persistent gap between statistical theory and common practice, there is a compelling need for focused evaluations that move beyond broad comparisons of many methods. Instead, targeted studies that assess the specific conditions under which commonly used but potentially flawed methods might or might not be acceptable are of great practical value. This study adopts this targeted approach. Its primary objective is not to conduct yet another broad comparison between the Wald interval and its many sophisticated alternatives but to perform a focused, in-depth evaluation of the performance of two closely related methods: the standard Wald interval and the Wald interval with continuity correction (Wald CC). The Wald CC, calculated as $\hat{p} \pm \left(z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} + \frac{1}{2n} \right)$,epresents a minimal modification intended to account for the discreteness of the binomial distribution. Its structural similarity to the standard Wald interval makes it a highly practical alternative for educators and practitioners. Yet, its specific operational characteristics precise coverage probabilities and expected widths across a comprehensive parameter space warrant a detailed and systematic investigation.

Therefore, this study aims to answer the following critical questions: Does the continuity correction genuinely remedy the fatal flaws of the standard Wald method? Under what specific conditions of sample size n and true population proportion p does the Wald CC achieve acceptable (near nominal) coverage? How does its expected width behave relative to its coverage performance? And finally, can clear, evidence-based guidelines be established for the application (or avoidance) of these two methods? By employing large scale computer simulation across a fine grid of parameters ($p = 0.01, 0.02, \ldots, 0.99$) , ($n = 20, 30, \ldots, 2000$), this research seeks to provide definitive answers, offering clear guidance to researchers and moving the discourse from theoretical criticism to actionable, data driven recommendations.

Binomial Probability Distribution

A sequence of n independent Bernoulli trials, each with success probability p, gives rise to the binomial distribution. This distribution describes the probability of a random variable X, defined as the total number of successes. For any x successes, the number of possible sequences is the binomial coefficient $\binom{n}{x}$, leading to the following probability distribution for

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

This can be denoted by $X \sim binomial(n, p)$, the mean and variance of the binomial distribution can be calculated as

$$E[X] = np$$
$$Var[X] = np(1-p)$$

Normal approximation to the binomial distribution

The Normal approximation to the Binomial distribution is one of the most celebrated and practical results in probability theory and statistics. Its origins lie in the Bernoulli Theorem, established by Jacob Bernoulli (1713). Bernoulli demonstrated that the relative frequency of success in a series of independent trials converges to the true probability of success. The pivotal step came with Abraham de Moivre (1667-1754). In 1733, de Moivre first derived the mathematical form of the Normal distribution as an approximation to the binomial probabilities for a large number of trials. He published this result in *The Doctrine of Chances* (1738), and it was later generalized by Pierre-Simon Laplace in his Analytical Theory of Probability (1812), which extended the result to a more general form. This generalization evolved into what we now know as the Central Limit Theorem (CLT), with the Binomial case being one of its earliest and most important specific applications.

According to the CLT, one obtains

$$\lim_{n\to\infty} \Pr\left(\frac{X-np}{\sqrt{np(1-p)}} \le x\right) = \Phi(x), \qquad for\ \forall x$$

If the number of trials in binomial distribution is sufficient large, one can obtain the "normal approximation to the binomial" as

$$F_X(k) = \Pr(X \le k) \approx \Phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right)$$

Wald Interval

The most prevalent application of confidence interval for the binomial proportion is the normal interval or the Wald approximation interval, where the actual distribution, is assumed to be a discrete Binomial distribution, but to obtain the interval we first approximate it to a continuous normal distribution, as depicted in Figure 1. The Wald interval results from inverting a Wald test (Agresti and Coull, 1998).

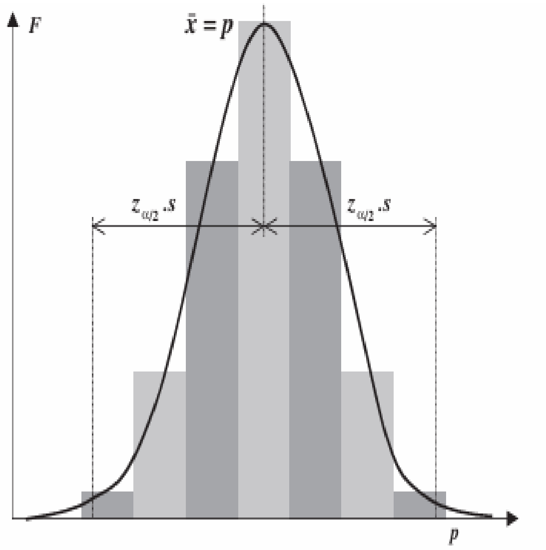


Figure 1. The Normal approximation to the Binomial

The Wald Test: $H_0: P = P_0$ v.s. $H_1: P \neq P_0$ (or 1-sided alternative)

Score Test uses the score statistic

$$Z_S = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$$

Wald Test uses the Wald statistic

$$Z_W = \frac{\hat{p}-p_0}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

A confidence interval is computed using a confidence level 1 – α. The confidence level represents the percentage of the computed confidence intervals containing the true value in the long run. Formally, let X be a random variable having a probability distribution with parameter θ. We have that $(l(X), u(X))$ is a 1 – α confidence interval for θ if $P(l(X) \le \theta \le u(X)) \ge 1-\alpha$, (Larsen and Marx, 2012).

We will derive the Wald interval. Let X, n, and p be as described in above, where $\hat{p} = X/n$ is the maximum likelihood estimate of p. As n gets large, then we have that

$$P\left(-z_{\alpha/2} \le \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \le z_{\alpha/2}\right) \approx$$
$$P\left(-z_{\alpha/2} \le \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \le z_{\alpha/2}\right) \approx 1-\alpha$$

then solve it with respect to p, which results in

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$
$$\approx 1-\alpha$$

So the $100(1-\alpha)\%$ bounds of the Wald interval is

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

The Wald method relies on certain conditions that must be satisfied. These conditions vary across different texts, so we have compiled the most common ones.

1. $np,\ n(1-p)$ are ≥ 5.
2. $n\hat{p}, n(1-\hat{p})$ are ≥ 10 .
3. $\hat{p}$ should not equal 0 or 1.

If $\hat{p} \approx 0$ or $\hat{p} \approx 1$, the confidence interval bounds can fall outside [0,1].

While the Wald method (normal approximation) involves only simple calculations, this approximation adversely affects its performance.

Wald Interval with continuity correction

A continuity correction has long been used when approximating a discrete probability distribution with a continuous one. Yates (1934) introduced an ad hoc rule of 'adding 0.5' to improve the accuracy of the normal approximation to a discrete distribution, a method now known as 'Yates' correction. Accordingly, the continuity corrected Wald (CCW) interval is obtained by subtracting $\frac{1}{2n}$ from the lower bound and adding $\frac{1}{2n}$ to the upper bound. This correction typically improves the interval's performance.

$$-Z_{\alpha/2} \le \frac{X+1/2-np}{\sqrt{np(1-p)}} \quad \text{and} \quad \frac{X-1/2-np}{\sqrt{np(1-p)}} \le Z_{\alpha/2}$$

We solve these equations using the notation $\hat{p} = \frac{X}{n}$. Thus, the equations we want to solve are

$$-Z_{\alpha/2} \le \frac{\hat{p}+1/(2n)-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \quad \text{and} \quad \frac{\hat{p}-1/(2n)-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \le Z_{\alpha/2}$$

with respect to p. The solution is

$$\left(\hat{p} - \frac{1}{2n} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\hat{p} + \frac{1}{2n} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

While the Wald interval with continuity correction improves upon the standard Wald interval, some argue it can lead to overcoverage, making the interval overly wide and conservative.

Evaluation Criteria

The most common criteria for evaluating confidence intervals are coverage probability and expected width, which are the primary metrics considered in this work. However, numerous other performance measures have been proposed in the literature.

A well performing confidence interval should satisfy two key properties: it should be narrow to maximize informative, and its actual coverage probability should be at least the nominal level while remaining as close to it as possible, avoiding excessive conservatism.

The Coverage Probability

Coverage probability is the probability that a confidence set contains the true population parameter. For a $(1-\alpha)\%$ confidence interval, we expect that approximately $(1-\alpha)\%$ of such intervals constructed from repeated sampling will contain the true parameter. Mathematically for a confidence interval procedure $CI(X)$ for θ:

Coverage $(\theta) = P(\theta \in CI(X) \mid \theta)$

For any confidence interval procedure for estimating p, the actual coverage probability at a fixed value of p is

$$\text{Coverage } (p) = \sum_{k=0}^{n} I(k,p)\binom{n}{k} p^{k}(1-p)^{n-k}$$

where $I(k,p) = 1$ if the interval contains p when X = k and $I(k,p) = 0$ if p is not contained in the interval. We summarize this, using the alternative description of performance, by averaging over the possible values that p can take, (Agresti & Coull

1998, p.120).

The expected width

Expected Width of a confidence interval is a crucial evaluation metric that balances precision (narrow intervals) with coverage probability (accuracy). The expected width is the average length of the confidence interval over repeated sampling.

The expected width is a critical metric for evaluating confidence intervals because it addresses three key issues:

1. The coverage-precision trade-off: Achieving 100% coverage requires an infinitely wide interval. The objective is to minimize the interval width while preserving the stated coverage probability.
2. Method comparison: For confidence interval procedures with equivalent coverage, the method yielding the smaller expected width is more efficient.
3. Sample size determination: Calculating the expected width is essential for estimating the sample size n needed to achieve a specified precision.

For a confidence interval of a parameter θ, if the interval is $[L, U]$, where U is the upper limit and L is the lower limit of the interval, then:

$$Width = U - L$$
$$Expected\ Width = E[U - L]$$

This implies

For a parameter θ with estimator $\hat{\theta}$ and standard error SE:

Confidence interval : $\hat{\theta} \pm z_{\alpha/2} . SE$

$$Width = 2. z_{\alpha/2} . SE$$

Thus

$$Expected\ Width = 2. z_{\alpha/2} . E[SE]$$

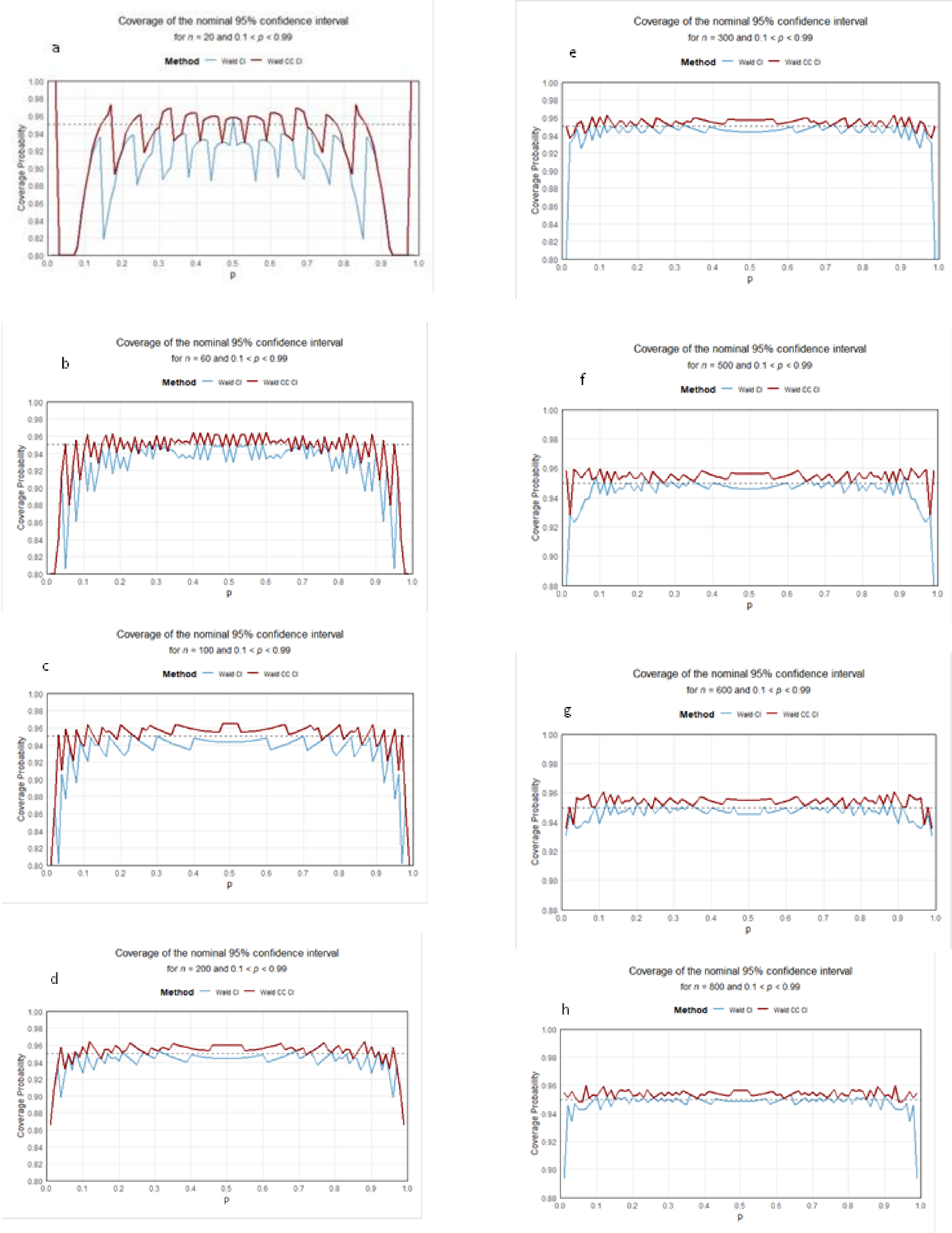
For Wald confidence interval: $\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$$Expected\ Width \approx 2. z_{\alpha/2} . E\left[\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

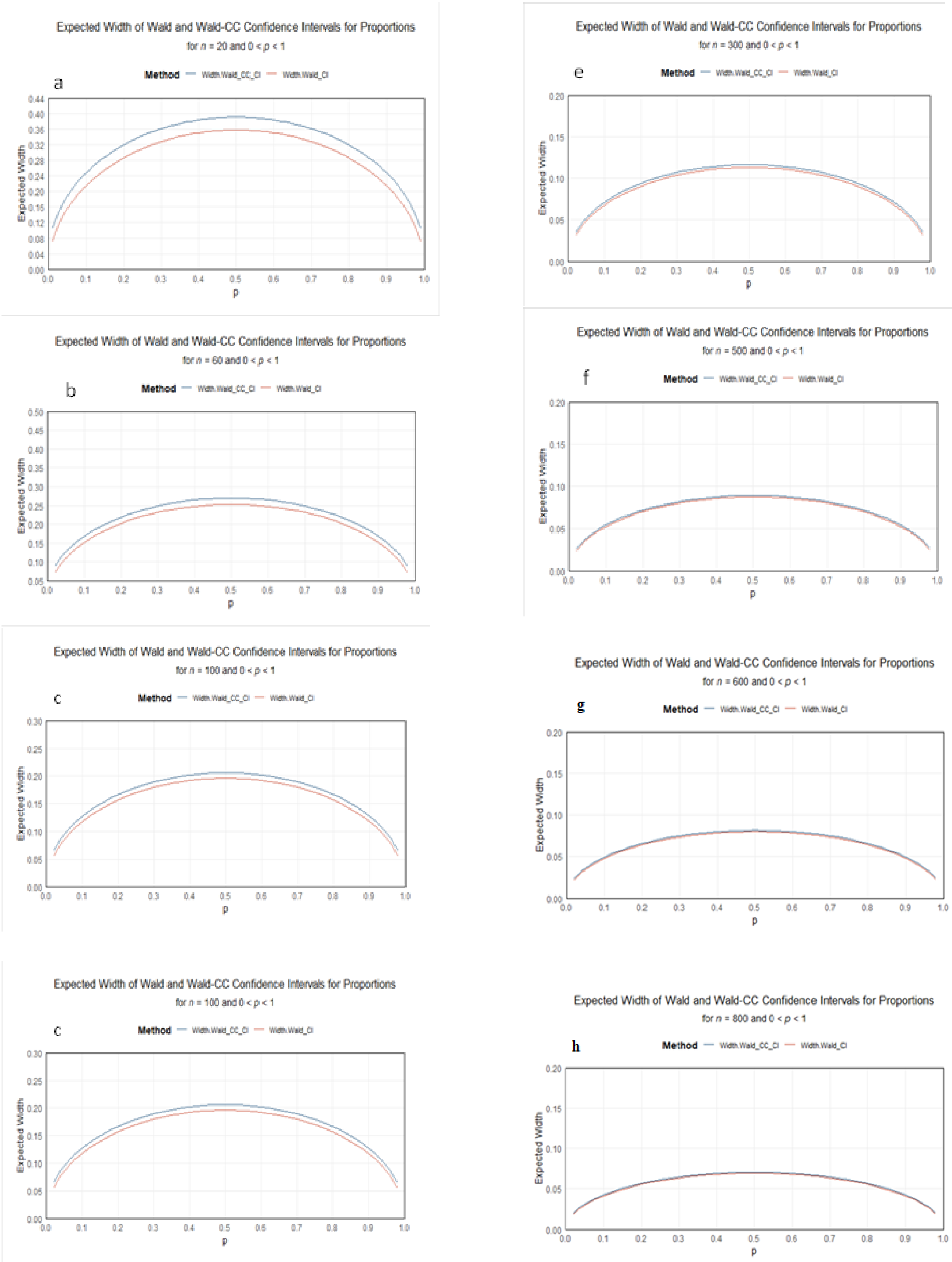
Simulation

A comprehensive evaluation was conducted to assess the coverage probability and expected width of 95% nominal confidence intervals generated by the standard Wald method and the Wald method with continuity correction (Wald CC). The study covered the complete parameter grid defined by proportions $p = 0.01, 0.02, \ldots, 0.99$ and sample sizes $n = 20, 30, \ldots, 2000$.

The coverage probabilities and expected widths for each method were computed using exact analytical calculation. For each combination of parameters (p, n), we computed the exact binomial probability mass function for all possible outcomes $k = 0, 1, \ldots, n$. For each outcome, we determined whether the resulting confidence interval contained the true parameter p, we used the theoretical expressions in Sections [The Coverage Probability, The expected width].



Figures 2: The coverage probability of the nominal 95% CI Standard Wald, and Wald with CC, for different values of n.



Figures 3: The expected Width of Standard Wald and Wald with CC, for different values of n.

Discussion

Figures 2 illustrate the coverage probability for both the Wald interval and the Wald interval with continuity correction Wald CC, demonstrating their performance across varying population proportions (p) and sample sizes $(n = 20,\ 60, 100, 300,\ 500,\ 600, 800)$. Figures 3 present the corresponding expected interval widths for both methods under the same range of p and n. The principal findings are as follows:

1. Standard Wald Interval: This method performed poorly across the entire parameter space. It consistently failed to achieve the nominal 95% coverage level, exhibiting highly unstable coverage probabilities that fluctuated dramatically with changes in p and n. Critically, this failure persisted even for very large sample sizes (e.g., n = 2000), contradicting the common perception that asymptotic adequacy is achieved with sufficiently large samples.

2. Wald with Continuity Correction (Wald CC): This method demonstrated substantially improved performance. Although slightly conservative for small sample sizes, it consistently met or exceeded the nominal 95% coverage for sample sizes $n \geq 200$, provided the true proportion p was not near the boundaries (specifically, for $0.1 \leq p \leq 0.9$). The expected interval width for Wald CC was reasonable and exhibited the theoretically desired behaviour, decreasing appropriately as sample size increased.

3. Performance at Extreme Proportions: For proportions near the boundaries ($p < 0.1\ or\ p > 0.9$), both methods exhibited catastrophic failure. Under these conditions, the theoretical confidence limits frequently fell outside the valid parameter space $[0,1]$, producing intervals that are not only imprecise but also logically invalid. The continuity correction, while beneficial for moderate proportions, did not adequately resolve this boundary problem.

Given its structural similarity to the widely known yet fundamentally flawed standard Wald interval, the Wald CC method presents a practical and easily implementable alternative for non-extreme proportions. However, for extreme proportions, researchers are advised to employ more reliable alternatives. This study provides clear, evidence-based guidance on the effective application domain of each method.

Conclusion

Based on the exhaustive simulation results presented in this study, we draw the following conclusions and offer the corresponding practical recommendations:

1. The standard Wald confidence interval for a binomial proportion consistently fails to maintain nominal coverage across virtually all combinations of p and n, even with large samples. Its erratic coverage properties and frequent production of invalid interval bounds render it unsuitable for statistical inference in any practical application. We recommendation that the standard Wald interval should be completely avoided in all practical applications, regardless of sample size.

2. The Wald interval with continuity correction method provides a significant enhancement in coverage properties with minimal added complexity. For sample sizes $n \geq 200$ and non-extreme proportions $(0.1 \leq p \leq 0.9)$, it consistently delivers near-nominal coverage without being excessively conservative. Its expected width is reasonable and decreases appropriately with increasing sample size. We recommendation that the Wald CC method is recommended as a viable and effective alternative for moderate to large samples $(n \geq 200)$ when the observed proportion is not near 0 or 1.

3. Both methods fail catastrophically at extreme proportions. For extreme proportions ($p < 0.1\ or\ p > 0.9$), neither the standard Wald nor the Wald CC method achieves acceptable performance. The theoretical confidence limits frequently fall outside $[0,1]$, rendering the intervals logically invalid. The continuity correction does not adequately resolve this boundary problem. We recommendation that When dealing with extreme proportions, researchers should avoid both Wald based methods and instead employ more sophisticated alternatives such as the Wilson score interval, the Agresti-Coull interval, or exact Clopper-Pearson methods. These alternatives, while more computationally involved, offer substantially greater reliability and maintain nominal coverage even near the boundaries.

4. The widespread use of the standard Wald interval is unjustified. Despite its simplicity and frequent appearance in textbooks and statistical software, the standard Wald interval's statistical properties do not justify its continued widespread use. Its persistence in practice likely stems from convenience and inertia rather than sound statistical reasoning. We recommendation that Practitioners, researchers, and educators should actively adopt better alternatives to ensure the validity of their statistical conclusions. The Wald CC interval represents a simple first step for non-extreme proportions, while more sophisticated methods should be employed for extreme proportions.

References

1. Agresti, A., & Coull, B. A. (1998). Approximate is better than exact for interval estimation of binomial proportions. *The American Statistician, 52*(2), 119-126.
2. Pires, A. M., & Amado, C. (2008). Interval estimators for a binomial proportion: Comparison of twenty methods. *REVSTAT–Statistical Journal, 6*(2), 165-197.
3. Brown, L. D., Cai, T. T., & Das Gupta, A. (2001). Interval estimation for a binomial proportion. *Statistical science, 16*(2), 101-117.
4. Bernoulli, J. (1713). *Ars conjectandi* [The art of conjecturing]. Thurneysen Brothers.
5. De Moivre, A. (1733). *Approximatio ad summam terminorum binomii $(a+b)^n$ in seriem expansi* [A method of approximating the sum of the terms of the binomial $(a+b)^n$ expanded into a series]. [Private circulation].
6. De Moivre, A. (1738). *The doctrine of chances: or, a method of calculating the probabilities of events in play*. Woodfall.
7. Ghosh, B. K. (1979). A comparison of some approximate confidence intervals for the binomial parameter. *Journal of the American Statistical Association, 74*(367), 894-900.
8. Laplace, P.-S. (1812). *Théorie analytique des probabilités* [Analytical theory of probabilities]. Courcier.
9. Vollset, S. E. (1993). Confidence intervals for a binomial proportion. *Statistics in medicine, 12*(9), 809-824.
10. Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society, 1*(2), 217–235.

Software and Reproducibility

All analyses were conducted using R version 4.5.0 ,no external packages were required for the core analyses

Reproducibility Statement:

To ensure full reproducibility of the results presented in this study, all code and data have been made publicly available. all results can be reproduced by running these scripts:

1. coverage_probability.R:

```
# (1) Calculate coverage probabilities for Wald
and Wald with continuity correction for binomial
proportions across various sample sizes and true
proportions
sample_sizes <- seq(20, 2000, by = 10)
p_values <- seq(0.01, 0.99, by = 0.01)

alpha <- 0.05
z <- qnorm(1 - alpha/2)

results <- data.frame( n = numeric(),p = numeric(),
  Wald = numeric(), Wald_CC = numeric())

# Calculate coverage probabilities
for (n in sample_sizes) {
  for (p in p_values) {
    k <- 0:n
    p_hat <- k / n
    probs <- dbinom(k, n, p)

se <- sqrt(p_hat * (1 - p_hat) / n) # Standard Wald
method
ci_lower <- pmax(0,p_hat - z * se)
ci_upper <- pmin(1,p_hat + z * se)
covered <- (ci_lower <= ) & (p <= ci_upper)
coverage_wald<-sum(probs[covered])

# Wald with continuity correction
     correction <- 0.5 / n
     ci_lower_cc <- pmax(0, p_hat - z * se - correct
  ion)
     ci_upper_cc <- pmin(1, p_hat + z * se + correc
  tion)
     covered_cc <- (ci_lower_cc <= p) & (p <= ci_
  upper_cc)
   coverage_wald_cc <- sum(probs[covered_cc])
```

```
# Store results
results<-rbind(results,data.frame(
    n = n,
    p = p,
    Wald = coverage_wald,
    Wald_CC = coverage_wald_cc ))
  }}

write.csv(results, "coverage_probability.csv", row.
names = FALSE)
```

2. expected_width.R:

(2) Calculate theoretical expected widths for Wald and Wald with continuity correction for binomial proportion confidence intervals

```
n_values <- seq(20, 2000, by = 10)
p_values <-seq(0.01,0.99,by =0.01)
conf_level <- 0.95

# Function to calculate theoretical expected width
theoretical_width <- function(p, n,conf_level =0.95
,method="wald") {
z <- qnorm(1 -(1 -conf_level) / 2)

  if (method == "wald") {
    se <- sqrt(p * (1 - p) / n)   # Theoretical Wald CI
width
    width <- 2 * z * se
  } else if (method == "wald_cc") {
# Theoretical Wald CI with continuity correction w
idth
    se <- sqrt(p * (1 - p) / n)
    width <- 2 * z * se + 1/n      # Added continuity c
orrection term
  }
  return(width)}

# Pre-allocate results data frame for better performa
nce

results <- data.frame(
  n = numeric(),
  p = numeric(),
  width_wald = numeric(),
  width_wald_cc = numeric())

# Calculate widths for all combinations
for (n in n_values) {
  for (p in p_values) {
    # Calculate theoretical widths
width_wald <- theoretical_width(p, n, method = "
wald")
width_wald_cc <- theoretical_width(p, n, method
= "wald_cc")

# Store results
results<-rbind(results,data.frame(
    n = n,

    p = p,
    width_wald = width_wald,
    width_wald_cc = width_wald_cc
  ))}}

# Save results to CSV file
write.csv(results, "Expected_Width.csv", row.nam
es = FALSE)
```

Supplementary Material

All numerical results presented in this study, including coverage probabilities and expected interval widths for each combination of sample size (n) and population proportion (p), are available in the accompanying Excel file (coverage_width.xlsx). The file contains 19701 rows of data covering $p = 0.01, 0.02, \ldots, 0.99$ and $n = 20, 30, \ldots, 2000$ The results can be accessed via the following QR code:

File Structure:
Column A: Sample size (n)
Column B: Population proportion (p)
Column C: Coverage probability (Standard Wald)
Column D: Coverage probability (Wald with CC)
Column E: Expected width (Standard Wald)
Column F: Expected width (Wald with CC)