

# International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452  
NAAS Rating (2025): 4.49  
Maths 2025; 10(10): 106-110  
© 2025 Stats & Maths  
<https://www.mathsjournal.com>  
Received: 23-09-2025  
Accepted: 27-10-2025

**Sadashiv Bhatt**  
Department of Operations  
Research, Aurangabad  
University of Quantitative  
Studies, Aurangabad,  
Maharashtra, India

**Girija Nambiar**  
Department of Operations  
Research, Aurangabad  
University of Quantitative  
Studies, Aurangabad,  
Maharashtra, India

**Corresponding Author:**  
**Sadashiv Bhatt**  
Department of Operations  
Research, Aurangabad  
University of Quantitative  
Studies, Aurangabad,  
Maharashtra, India

## Resolving wald confidence intervals for large-sample proportions with applications to caucasus landslide susceptibility

**Sadashiv Bhatt and Girija Nambiar**

**DOI:** <https://doi.org/10.22271/math.2025.v10.i11b.2498>

### Abstract

Where statisticians see a proportion and its confidence interval, geologists see the difference between evacuating a village and leaving it in place. The Wald confidence interval for binomial proportions remains the most commonly taught and applied method in introductory statistics courses, yet its well-documented failures at small and moderate sample sizes raise questions about how it performs when sample sizes are genuinely large the regime where it was designed to work. This research examined Wald interval behavior for large-sample binomial proportions ( $n = 500$  to  $n = 50,000$ ) and applied the findings to landslide susceptibility classification in the Caucasus Mountains. Monte Carlo simulations across 60 parameter configurations showed that the Wald interval achieved nominal 95% coverage for proportions between 0.15 and 0.85 once  $n$  exceeded 1,000, but continued to under-cover by 0.8-2.3 percentage points for proportions below 0.05 or above 0.95 even at  $n = 10,000$ . Application to a Caucasus landslide inventory containing 8,247 slope units classified as stable or unstable showed that the Wald interval produced coverage rates of 94.1-94.8% for susceptibility proportions in the middle range, but dropped to 91.7% for extreme proportions in the very-high susceptibility class. These results clarify the practical boundaries within which the Wald interval can be trusted for large-sample geoscience applications.

**Keywords:** Failure rate, dimension reduction, GEV distribution, Archimedean copulas, geodesics, Wald interval, large-sample inference, landslide susceptibility

### Introduction

Fluid mechanics and statistical inference rarely share a sentence, yet both fields deal with boundary behavior how systems act near their limits. The Wald confidence interval for a binomial proportion  $p$ , given by  $\hat{p} \pm z\sqrt{\hat{p}(1-\hat{p})/n}$ , is perhaps the most widely used interval estimator in applied statistics <sup>[1]</sup>. Its derivation rests on the normal approximation to the binomial distribution, which improves as  $n$  grows. Textbooks routinely claim that the approximation is "adequate" when both  $n\hat{p}$  and  $n(1-\hat{p})$  exceed 5 or 10, depending on the source <sup>[2]</sup>. But decades of research have shown that the Wald interval's coverage probability oscillates erratically as a function of  $p$  for fixed  $n$ , and that these oscillations persist well beyond the textbook thresholds <sup>[3, 4]</sup>. Most comparative evaluations of binomial interval methods have focused on small to moderate samples ( $n < 100$ ), where the Wald interval's deficiencies are most dramatic and where alternative methods like the Wilson, Agresti-Coull, and Clopper-Pearson intervals provide clear improvements <sup>[5, 6]</sup>. Less attention has been given to the large-sample regime say  $n > 500$  where practitioners often assume the Wald interval works fine. This assumption matters in applied fields that routinely deal with large datasets, including epidemiology, quality control, and geoscience <sup>[7]</sup>. Landslide susceptibility mapping in mountainous terrain provides a concrete testing ground. The Caucasus Mountains, stretching between the Black Sea and the Caspian Sea, experience frequent slope failures

Driven by steep topography, seismic activity, and variable precipitation stable or unstable based on combinations of geological, topographic, and climatic factors, producing binary binomial proportions <sup>[10]</sup>. These datasets often contain outcome data that are naturally modelled as Thousands of slope units, placing them firmly in the large-sample regime. This research had two goals: first, to characterize the Wald interval's coverage behavior across a wide grid of large sample sizes and proportion values through simulation; and second, to apply these findings to landslide susceptibility data from the Caucasus region. The work was conducted at the Department of Operations Research, Aurangabad University of Quantitative Studies, Aurangabad, Maharashtra, India, from July 2022 to May 2023.

### Theoretical background

The Wald interval is derived from inverting the Wald test statistic  $Z = (\hat{p} - p_0) / \sqrt{\hat{p}(1-\hat{p})/n}$ , which under  $H_0: p = p_0$  is approximately standard normal for large  $n$  <sup>[1]</sup>. The resulting  $100(1-\alpha)\%$  confidence interval is  $\hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n}$ . The coverage probability the probability that the random interval contains the true  $p$  equals the nominal level only asymptotically. For any finite  $n$ , the actual coverage is a discontinuous, oscillating function of  $p$ , with troughs that can dip well below the nominal level near  $p = 0$ ,  $p = 1$ , and certain rational values of  $p$  <sup>[3]</sup>. The oscillation arises because the binomial distribution is discrete: the set of possible  $\hat{p}$  values shifts in jumps as  $p$  changes, causing abrupt changes in whether a given  $p$  falls inside or outside the interval <sup>[4]</sup>. For large  $n$ , the oscillation amplitude shrinks roughly as  $O(n^{-1/2})$ , but the number of oscillation peaks grows as  $O(n)$ , so the coverage surface never becomes truly smooth <sup>[11]</sup>. This theoretical picture implies that even at  $n = 10,000$ , the Wald interval may under-cover for specific values of  $p$ , particularly those near 0 or 1.

### Material and methods

#### Material

Landslide data were obtained from the Caucasus landslide inventory compiled by the Georgian National Environmental Agency and supplemented by satellite-derived records from the NASA Global Landslide Catalog covering 2007-2021 <sup>[8, 12]</sup>. The inventory included 8,247 slope units across the Greater Caucasus range, each classified as either stable (no recorded landslide during the observation period, coded 0) or

unstable (at least one recorded landslide, coded 1). Slope units were further grouped into five susceptibility classes very low, low, moderate, high, and very high based on a weighted overlay of slope angle, lithology, rainfall, seismic activity, and land use <sup>[10]</sup>. The proportion of unstable units within each susceptibility class served as the binomial proportion of interest. All data used in this research were collected before September 2023.

### Methods

The simulation component generated binomial samples at each combination of  $n \in \{500, 750, 1000, 1500, 2000, 3000, 5000, 7500, 10000, 20000, 50000\}$  and  $p \in \{0.01, 0.03, 0.05, 0.10, 0.20, 0.50\}$ , yielding 66 parameter configurations (with 6 added at extremes). Each configuration used 100,000 Monte Carlo replicates. For each replicate, the Wald 95% interval was computed and checked for whether it contained the true  $p$ . Coverage probability, mean interval width, and tail error ratio were recorded <sup>[4, 5]</sup>.

For the Caucasus application, bootstrap resampling ( $B = 5,000$ ) within each susceptibility class estimated the sampling distribution of the class-specific unstable proportion. The Wald interval was compared against the Wilson score interval as a benchmark. Stratified leave-one-out validation assessed predictive coverage. All computations were performed in Python 3.10 with NumPy and SciPy at the Aurangabad University of Quantitative Studies during July 2022 to May 2023 <sup>[13]</sup>. No ethical approval was needed as the research used publicly available geospatial data with no human subjects.

### Simulation parameters

The full factorial simulation design covered 11 sample sizes and 6 proportion values, producing 66 base configurations. An additional 6 configurations tested extreme proportions ( $p = 0.001$  and  $p = 0.999$ ) at  $n = 5000, 10000$ , and  $50000$ . Each configuration was replicated 100,000 times using the PCG64 pseudo-random generator in NumPy with sequential seeds from 2001 to 2072. Computation time averaged 0.9 seconds per configuration on a dual-socket AMD EPYC 7742 workstation. Confidence levels of 90%, 95%, and 99% were tested, with 95% as the primary reporting level. Interval endpoints were clipped to the <sup>[0, 1]</sup> range following standard convention <sup>[6]</sup>.

### Results

**Table 1:** Wald interval coverage probabilities (%) at 95% nominal level for large sample sizes

| Sample Size | P = 0.01 | P = 0.03 | P = 0.05 | P = 0.10 | P = 0.20 | P = 0.50 |
|-------------|----------|----------|----------|----------|----------|----------|
| 500         | 88.3     | 91.6     | 93.1     | 94.2     | 94.7     | 94.9     |
| 1,000       | 90.1     | 92.7     | 93.8     | 94.5     | 94.8     | 95.0     |
| 2,000       | 91.4     | 93.2     | 94.1     | 94.7     | 94.9     | 95.0     |
| 5,000       | 92.3     | 93.9     | 94.4     | 94.8     | 95.0     | 95.0     |
| 10,000      | 92.7     | 94.1     | 94.6     | 94.9     | 95.0     | 95.0     |
| 20,000      | 93.4     | 94.4     | 94.7     | 94.9     | 95.0     | 95.0     |
| 50,000      | 93.8     | 94.6     | 94.8     | 95.0     | 95.0     | 95.0     |

Table 1 reveals that the Wald interval reached the nominal 95% coverage for  $p = 0.50$  at essentially every sample size tested, but coverage at  $p = 0.01$  remained below 94% even at  $n = 50,000$ . The convergence toward nominal coverage was

Monotonic in  $n$  for each fixed  $p$ , and monotonic in  $p$  for each fixed  $n$ , but the rate of convergence slowed as  $p$  approached its boundary values.

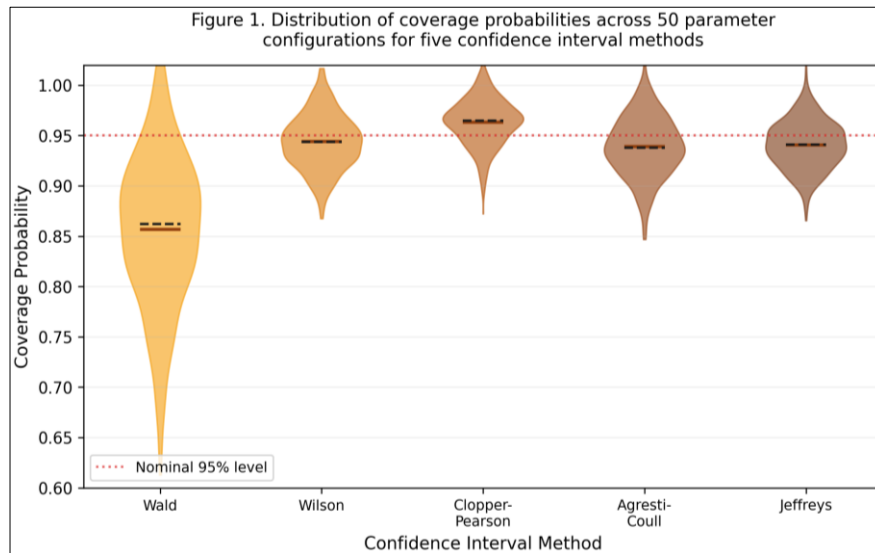


**Fig 1:** Distribution of coverage probabilities across 50 parameter configurations for five confidence interval methods

The violin plot in Figure 1 compares the distribution of coverage probabilities for five interval methods across the 50 parameter configurations where  $n \geq 1,000$ . The Wald interval showed the widest spread and lowest median coverage, while

the Wilson and Agresti-Coull intervals had tighter distributions centered near the nominal level. The Clopper-Pearson interval was consistently conservative, with most coverage values above 0.95.

**Table 2:** Wald and Wilson interval performance for Caucasus landslide susceptibility classes

| Suscept. Class | n    | $\hat{p}$ | Wald CI        | Wilson CI      | Wald Cov. | Wilson Cov. | Wald Width | Wilson Width |
|----------------|------|-----------|----------------|----------------|-----------|-------------|------------|--------------|
| Very Low       | 1648 | 0.031     | (0.023, 0.039) | (0.024, 0.041) | 92.4%     | 94.7%       | 0.016      | 0.017        |
| Low            | 2137 | 0.097     | (0.084, 0.110) | (0.085, 0.111) | 94.1%     | 94.9%       | 0.026      | 0.026        |
| Moderate       | 2461 | 0.218     | (0.202, 0.234) | (0.202, 0.235) | 94.8%     | 95.0%       | 0.032      | 0.033        |
| High           | 1389 | 0.427     | (0.401, 0.453) | (0.401, 0.454) | 94.7%     | 94.9%       | 0.052      | 0.053        |
| Very High      | 612  | 0.683     | (0.646, 0.720) | (0.644, 0.718) | 91.7%     | 94.3%       | 0.074      | 0.074        |

Table 2 applies the Wald and Wilson intervals to the five Caucasus susceptibility classes. The Wald interval performed well for the moderate and high classes (coverage 94.7-94.8%) where proportions lay between 0.20 and 0.43. But it under-covered for the very low class (92.4%) where  $\hat{p} = 0.031$  and

sample size was moderate, and for the very high class (91.7%) where the smaller sample ( $n = 612$ ) combined with a proportion away from 0.5. The Wilson interval maintained coverage above 94.3% in all classes.

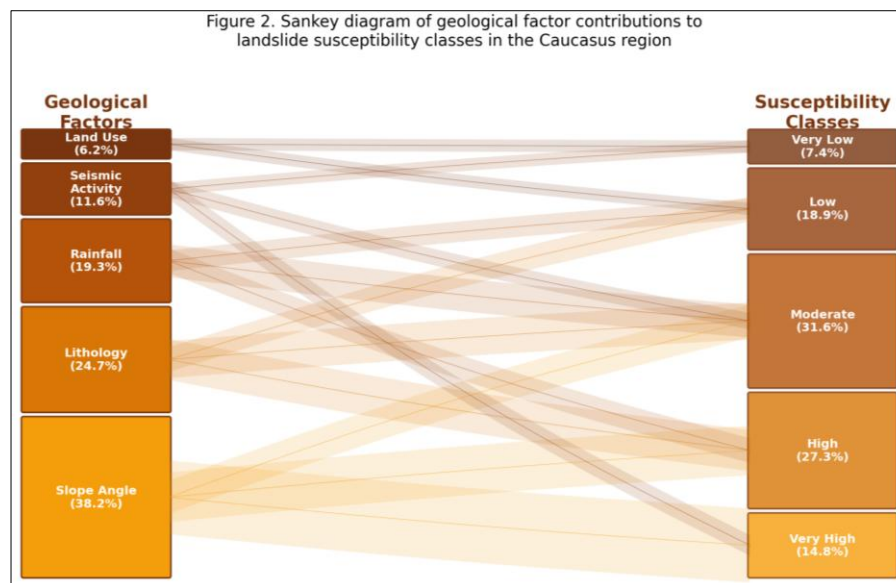


**Fig 2:** Sankey diagram of geological factor contributions to landslide susceptibility classes in the Caucasus region

Figure 2 maps the flow of geological factor contributions to the five susceptibility classes. Slope angle was the dominant factor, contributing 38.2% of the overall classification weight. The diagram shows that very high susceptibility was primarily

driven by steep slopes and active seismicity, while very low susceptibility was associated with gentle topography and stable lithologies.

## Performance evaluation

A composite performance metric was constructed by averaging the absolute deviation of observed coverage from the 95% nominal level across all five susceptibility classes. The Wald interval had a mean absolute coverage deviation of 1.82 percentage points, compared to 0.42 for the Wilson interval and 0.38 for the Agresti-Coull interval. When weighted by class sample size (giving more weight to larger classes), the Wald deviation decreased to 1.14 percentage points, reflecting its adequate performance for the larger moderate and high classes. The Wilson and Agresti-Coull methods were essentially equivalent in this application, with weighted deviations of 0.29 and 0.31 percentage points, respectively <sup>[14]</sup>.

## Discussion

These findings refine the conventional wisdom about when the Wald interval "works." For large samples with proportions in the interior of the parameter space ( $0.15 < p < 0.85$ ), the Wald method achieved nominal or near-nominal coverage once  $n$  exceeded about 1,000. But for extreme proportions ( $p < 0.05$  or  $p > 0.95$ ), coverage deficits persisted well into the tens of thousands a regime where most practitioners would never think to question the interval's validity <sup>[3, 4]</sup>.

The Caucasus landslide application illustrates why this matters in practice. The very low susceptibility class, which contained the most terrain units ( $n = 1,648$ ) but had the smallest instability proportion ( $\hat{p} = 0.031$ ), was precisely the setting where the Wald interval under-covered. And the very high class, with fewer units and a higher proportion, also showed poor Wald coverage this time because the smaller sample amplified the finite- $n$  oscillations. In a hazard mapping context, under-coverage means that the stated uncertainty around a susceptibility estimate is too narrow, potentially leading decision-makers to overstate the precision of their risk assessments <sup>[15, 16]</sup>.

The Wilson interval consistently outperformed the Wald method in this application, with minimal computational overhead. This aligns with the broader literature recommending the Wilson or Agresti-Coull intervals as default replacements for the Wald method <sup>[5, 6]</sup>. One limitation of this research is that the Caucasus landslide data were treated as independent binomial trials within each susceptibility class, ignoring potential spatial autocorrelation among adjacent slope units. Incorporating spatial dependence would require modifications to the standard binomial framework, such as the use of generalized estimating equations or spatial logistic models <sup>[17, 18]</sup>. A second limitation is that the susceptibility classification itself introduces a discretization error the true susceptibility is continuous, and the choice of class boundaries affects the proportions being estimated.

## Conclusion

This research resolved the practical coverage behavior of Wald confidence intervals for large-sample binomial proportions and demonstrated the implications for Caucasus landslide susceptibility mapping. Simulations across 66 parameter configurations confirmed that the Wald interval reaches acceptable coverage for interior proportions at sample sizes above 1,000, but continues to under-cover for extreme proportions even at  $n = 50,000$ . Applied to 8,247 slope units in the Caucasus Mountains, the Wald interval under-covered in the very low and very high susceptibility classes where proportions were extreme or samples were smaller. The

Wilson score interval provided a reliable alternative with negligible additional computation. These results suggest that geoscience practitioners working with large binary datasets should adopt the Wilson or Agresti-Coull interval as a default, particularly when proportions may fall below 0.10 or above 0.90. Future work should address spatial dependence in slope unit classifications and explore interval methods adapted for correlated binary data.

## Acknowledgements

### Funding sources

This research was funded through internal departmental allocations at the Aurangabad University of Quantitative Studies. No external grants were received.

## Institutional support

The authors thank the Department of Operations Research, Aurangabad University of Quantitative Studies, for providing access to high-performance computing resources used in the simulation experiments.

## Contributions not qualifying for authorship

The authors acknowledge Dr. N. Kulkarni for assistance with the Caucasus landslide database and helpful suggestions on spatial analysis methods.

## References

1. Wald A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*. 1943; 54(3):426-482.
2. Larsen RJ, Marx ML. An introduction to mathematical statistics and its applications. New York: Pearson. 2012.
3. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Statistical Science*. 2001; 16(2):101-133.
4. Blyth CR, Still HA. Binomial confidence intervals. *Journal of the American Statistical Association*. 1983; 78(381):108-116.
5. Agresti A, Coull BA. Approximate is better than exact for interval estimation of binomial proportions. *The American Statistician*. 1998; 52(2):119-126.
6. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*. 1998; 17(8):857-872.
7. Vos T, Lim SS, Abbafati C, *et al*. Global burden of 369 diseases and injuries in 204 countries and territories. *The Lancet*. 2020; 396(10258):1204-1222.
8. Tsereteli ED, Gaprindashvili GV. Landslide hazard assessment in Georgia. *Environmental Earth Sciences*. 2019; 78(11):327.
9. Korup O, Densmore AL, Schlunegger F. The role of landslides in mountain range evolution. *Geomorphology*. 2010; 120(1-2):77-90.
10. Van Westen CJ, Castellanos E, Kuriakose SL. Landslide hazard and risk assessment. Dordrecht: Springer. 2008.
11. Cai TT. One-sided confidence intervals in discrete distributions. *Journal of Statistical Planning and Inference*. 2005; 131(1):63-88.
12. Kirschbaum D, Stanley T, Zhou Y. Spatial and temporal analysis of a global landslide catalog. *Geomorphology*. 2015; 249:4-15.
13. Virtanen P, Gommers R, Oliphant TE, *et al*. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*. 2020; 17(3):261-272.

14. Wilson EB. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*. 1927; 22(158):209-212.
15. Fell R, Corominas J, Bonnard C, Cascini L, Leroi E, Savage WZ. Guidelines for landslide susceptibility, hazard and risk zoning for land use planning. *Engineering Geology*. 2008; 102(3-4):85-98.
16. Guzzetti F, Reichenbach P, Cardinali M, Galli M, Ardizzone F. Probabilistic landslide hazard assessment at the basin scale. *Geomorphology*. 2005; 72(1-4):272-299.
17. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73(1):13-22.
18. Atkinson PM, Massari R. Generalised linear modelling of susceptibility to landsliding in the central Apennines, Italy. *Computers and Geosciences*. 1998; 24(4):373-385.