Finding Reliable Confidence Interval for the Odds Ratio through Simulation and Visualization

Group Members: Jiyoung Cheong, Phuong Dinh, Reagan Billingsley

Abstract

There are different ways to compute the confidence interval for the odds ratio (OR) with common methods suggested by Woolf, Agresti, and Gart. This work focuses on finding the best method for calculating the OR using simulation and visualization in different ways, focusing on the analysis of the three common methods with and without Welch's adjustment. Parameters were chosen ranging from moderate values to extreme ones, with 1760 unique combinations in total. Monte Carlo simulation was used for examination. According to the output, tw, ta, and zw performed well with ta showing the most success. The only case in which ta failed was when both population sample sizes were extremely small. In addition, further exploration was applied to a real-world example of smoking status and lung cancer diagnosis. Other simulation studies with 90% and 99% confidence levels were also performed.

Research Background

Odds ratios (ORs) are widely used in statistical analysis to quantify the strength of association between two binary variables. In the context of a 2×2 contingency table, the OR measures how much more likely an event (e.g., success or failure) is to occur in one group compared to another. For example, in medical research, ORs are frequently used to assess the relationship between exposure (e.g., smoking) and an outcome (e.g., lung cancer). Despite their simplicity, ORs provide valuable insights into relationships that can inform decision-making in various fields.

The theoretical probabilities for success and failure in two groups are given by p_1 and $1-p_1$ for Group 1, and p_2 and $1-p_2$ for Group 2. The odds ratio is defined as

$$\theta = \frac{p1/(1-p1)}{p2/(1-p2)}$$

where p_1, p_2 are the probability of success for Group 1 and Group 2, respectively. In practice, these probabilities are unknown and must be estimated using observed data. The maximum likelihood estimates (MLEs) of p_1 and p_2 are computed as $\widehat{p_1} = \frac{n_{11}}{n_{1+}}$ and $\widehat{p_2} = \frac{n_{21}}{n_{2+}}$, where n_{ij} represents the cell counts in the 2×2 contingency table. The MLE of the odds ratio is then calculated as

$$\widehat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

To account for sampling variability, confidence intervals (CIs) for the odds ratio are constructed. One of the earliest and most widely used methods is the Woolf CI (Woolf, 1955), which is based on the delta method. The Woolf CI is expressed on the log scale as

$$log\,\widehat{\theta} \pm z_{\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

and is then exponentiated to obtain the CI for θ. However, the Woolf method suffers from poor performance when sample sizes are small or when cell counts include zeros, leading to overly conservative coverage rates (i.e., higher than the nominal level).

To address these limitations, Gart proposed an adjustment that adds 0.5 to each cell count before calculating the CI (Gart, 1966). This adjustment, also known as the Haldane-Anscombe correction, improves performance for small sample sizes but may still exhibit conservatism in certain scenarios. Agresti introduced an independent-smoothed adjustment, which adds a more sophisticated correction term (c_{ij}) to each cell count (Agresti, 1999). These adjustments have been shown to perform well in specific contexts, particularly when odds ratios fall within a moderate range (e.g., 0.1 to 10).

Despite these advancements, challenges remain in constructing reliable confidence intervals for odds ratios. For instance, Fagerland et al. (2011) and Lawson (2004) recommend the Gart method for small sample sizes but note its limitations for extreme odds ratios or highly imbalanced sample sizes. Furthermore, the application of Welch's adjustment—commonly used in two-sample t-tests to reduce the assumption of equal variances.

Odds ratio confidence intervals are critical in fields such as medicine, biology, epidemiology, and social sciences. For example, in clinical trials, OR CIs help determine the effectiveness of

treatments. In ecological studies, they are used to analyze associations between environmental factors and species distributions. However, the reliability of these intervals depends heavily on the choice of method and the underlying data characteristics, such as sample size and odds ratio magnitude.

This project aims to systematically assess the performance of three popular OR confidence interval methods—Woolf, Gart, and Agresti—with and without Welch's adjustment. By exploring a wide range of parameter combinations, including small and unequal sample sizes, extreme odds ratios, and varying success probabilities, we seek to identify scenarios where these methods perform reliably or unreliably. Such insights will enhance our understanding of OR confidence intervals and guide their appropriate use in practical applications.

Simulation Study

1. Method

In this study, we conducted a comprehensive simulation to evaluate the reliability of six confidence interval methods for estimating odds ratios in 2×2 contingency tables: Woolf (original and Welch-adjusted), Gart (original and Welch-adjusted), and Agresti (original and Welch-adjusted). The study utilized a variation of four key parameters: p_1, OR (θ), n_{1d}(n_{1+}), and n_{2d} (n_{2+}) for a total of 1760 unique combinations for analysis. These parameters were varied as follows:

p_1 = [0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95],
θ = [1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100],
n_{1d} = [5, 10, 25, 50],
n_{2d} = [5, 10, 25, 50].

Our analysis was divided into two stages:

- First, we generated histograms of empirical coverage rates for all six methods across various parameter combinations to identify the top-performing methods. Based on these preliminary results, we observed that the Agresti method with Welch's adjustment (ta), the original Woolf method (zw), and the Woolf method with Welch's adjustment (tw) demonstrated superior performance, as their coverage rates were consistently close to or above the nominal 95% level.
- In the second stage, after identifying the top-performing methods (ta, zw, and tw) through the histogram analysis, we conducted a more detailed investigation using 3D surface plots to validate and further explore their performance under various scenarios. Specifically, we examined these methods in extreme and challenging conditions, such as small-small (n_{1+} = n_{2+} = 5), large-large (n_{1+} = n_{2+} = 50), moderate-moderate (n_{1+} = n_{2+} = 25), and unbalanced (small-large: n_{1+} = 5, n_{2+} = 50, large-small: n_{1+} = 50, n_{2+} = 5) sample size configurations. For each scenario, we generated 3D surface plots where:
    - The x-axis represents the success probability for Group 1 (p_1).
    - The y-axis represents the odds ratio (θ).
    - The z-axis represents the empirical coverage rate.

2. *Results and Interpretation*

By conducting 10,000 Monte Carlo simulations, histograms were constructed to visualize the empirical coverage rates achieved through each of six methods: Agresti with (ta) and without (za) Welch's adjustment, Gart with (tg) and without (zg) Welch's adjustment, and Woolf with (tw) and without (zw) Welch's adjustment (Figure 1). From these results, the tw case has the most conservative results with the most empirical coverage rates above 0.95. The ta and zw methods appear similar from their histograms, however, the ta method is better than both tw and zw because ta has more cases closer to 0.95. The za and tg methods also had similar coverage rate frequencies, but were more liberal cases with much higher frequencies below 0.95. The zg method performed the worst with a high frequency of coverage rates anywhere between 0.90 and 1.00. Because tw, ta, and zw were clearly the most successful cases, only these three methods were used for further analysis.

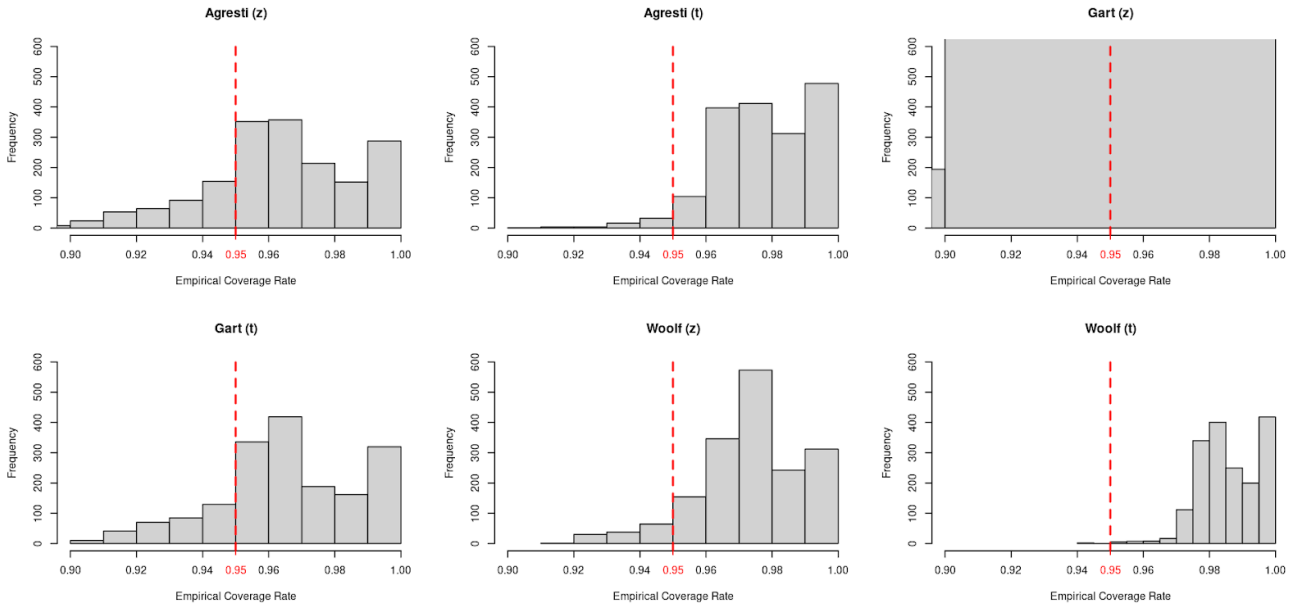

Figure 1. Histograms of empirical coverage rates for each of the six methods are shown with a vertical dashed line at 0.95, representing the nominal coverage rate set in the simulation studies. Graphs are labeled with respective OR methods used as well as "t" for Welch-adjusted methods or "z" for original methods.

Following preliminary studies, method analysis using 10,000 Monte Carlo simulations was done in which the same three methods, tw, ta, and zw, presented the most successful. To better visualize the empirical coverage rate success with respect to different sample size cases, 3D plots of these three methods were made. Cases analyzed include moderate-moderate ($n_{1+} = 25$, $n_{2+} = 25$), large-large ($n_{1+} = 50$, $n_{2+} = 50$), large-small ($n_{1+} = 50$, $n_{2+} = 5$), small-small ($n_{1+} = 5$, $n_{2+} = 5$), and small-large ($n_{1+} = 5$, $n_{2+} = 50$). These cases were chosen for analysis in order to determine which confidence interval would be the most successful with both equal

sample sizes as well as extremely unequal sample sizes. The small-large case for all three confidence interval methods performed similarly with narrow clustering between 0.97 and 1.00, and therefore are not presented.

Two of the most successful methods from the Monte Carlo simulations, ta and tw, performed very similarly at the moderate-moderate, large-large, and large-small sample size cases. The empirical coverage rates for tw (Figure 2) and ta (Figure 3) were all clustered around or above z = 0.95, representing the nominal coverage rate of 95%. The tw method was the most conservative with more narrow clustering, particularly in the moderate-moderate sample size case. The ta method presents the best results with more cases around 0.95.

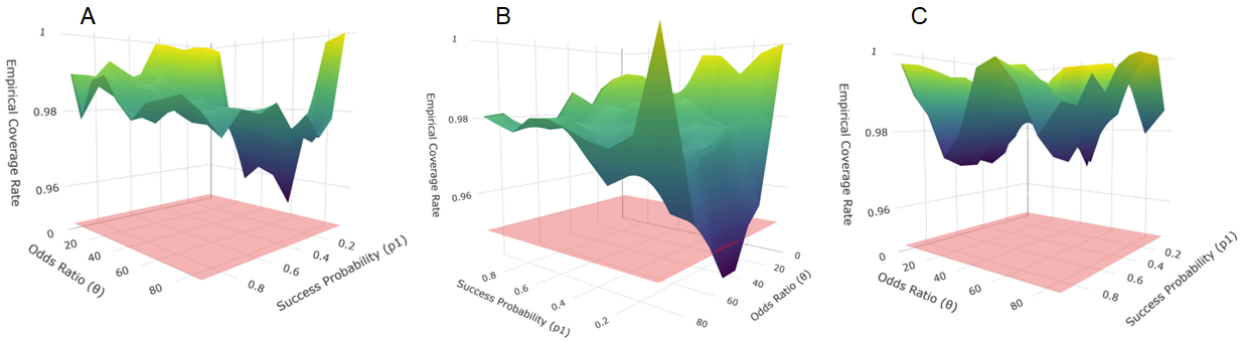

Figure 2. 3D plots of the Woolf Welch-adjusted (tw) method with the red plane highlighting z = 0.95, indicating the nominal coverage rate of 95%. The moderate-moderate sample size case with n_{1+} = 25 and n_{2+} = 25 (A), the large-large sample size case with n_{1+} = 50 and n_{2+} = 50 (B), and the large-small sample size case with n_{1+} = 50 and n_{2+} = 5 (C) are shown.

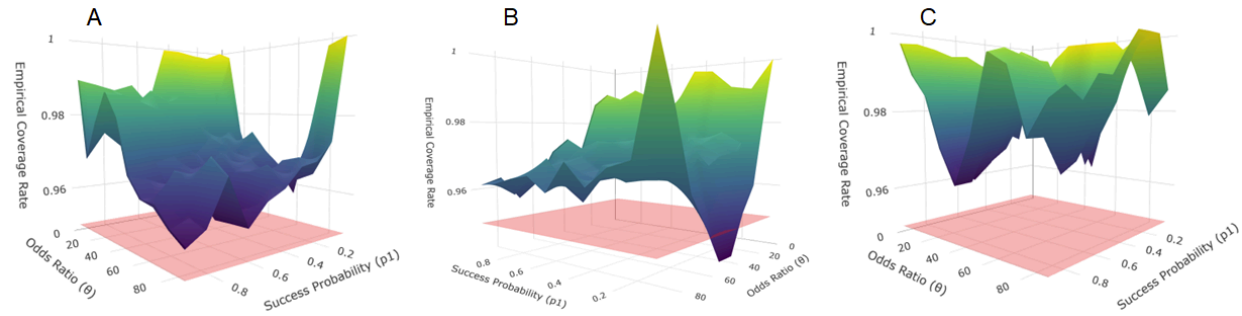

Figure 3. 3D plots of the Agresti Welch-adjusted (ta) method with the red plane highlighting z = 0.95, indicating the nominal coverage rate of 95%. The moderate-moderate sample size case with n_{1+} = 25 and n_{2+} = 25 (A), the large-large sample size case with n_{1+} = 50 and n_{2+} = 50 (B), and the large-small sample size case with n_{1+} = 50 and n_{2+} = 5 (C) are shown.

The third method analyzed via 3D plots was the zw method, which proved least successful in most cases (Figure 4). This method had much wider clustering both above and below 0.95, indicating a more liberal test than the two other methods analyzed.

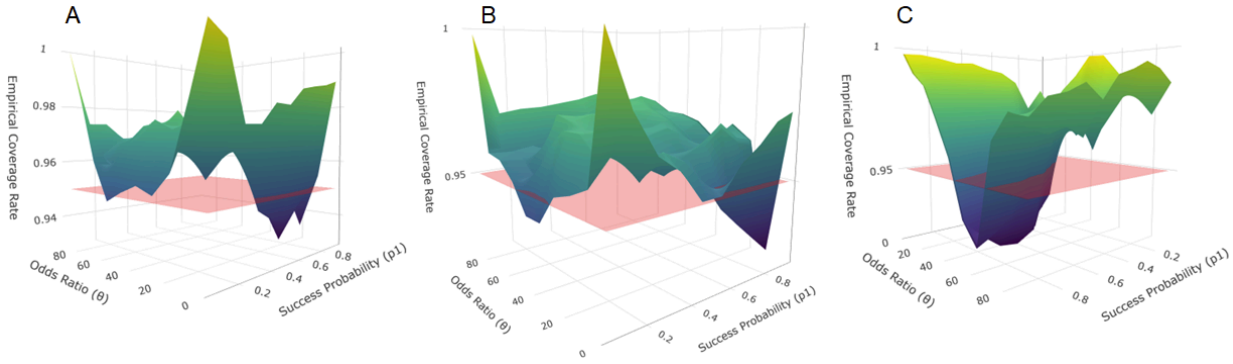

Figure 4. 3D plots of the Woolf original (zw) method with the red plane highlighting z = 0.95, indicating the nominal coverage rate of 95%. The moderate-moderate sample size case with n_{1+} = 25 and n_{2+} = 25 (A), the large-large sample size case with n_{1+} = 50 and n_{2+} = 50 (B), and the large-small sample size case with n_{1+} = 50 and n_{2+} = 5 (C) are shown.

Although this method was least successful in most sample size cases, it did show greater success in the small-small sample size case. Both Woolf methods performed well in this sample size case, with the Agresti method showing the least success (Figure 5), whose lowest z-value (empirical coverage rate) is approximately 0.91. In the case of both sample sizes being extremely small, the ta method was the most liberal and is the only method with empirical coverage rates well below 0.95. The tw and zw methods both remained clustered above 0.95, with the tw method maintaining the most conservative results.

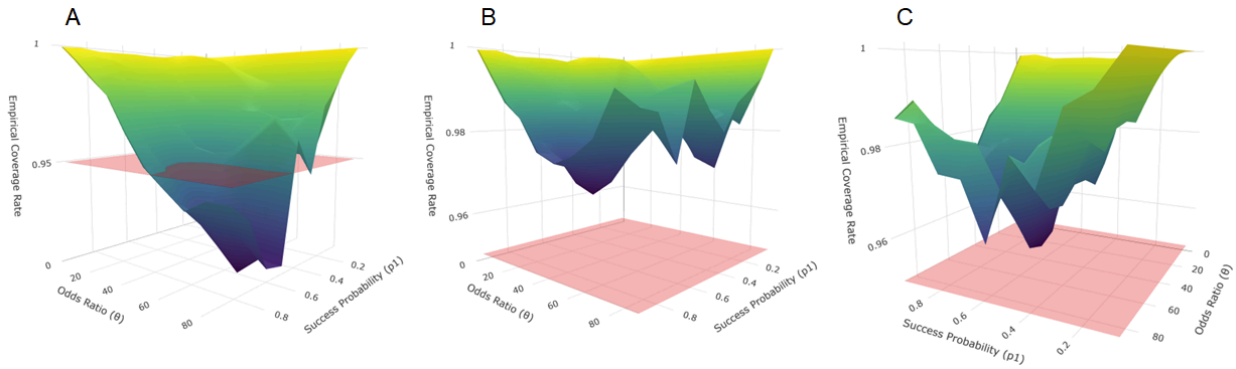

Figure 5. A 3D plot of the three methods, the ta method (A), tw method (B) and zw method (C) are shown for the small-small sample size case with n_{1+} = 5 and n_{2+} = 5. The red plane highlighting z = 0.95 indicates the nominal coverage rate of 95%.

Application to Real-World Data

The dataset selected for this analysis explores the relationship between smoking status and the incidence of lung cancer, providing a real-world application for evaluating odds ratio confidence intervals. The data is organized into a 2×2 contingency table, where individuals are categorized based on their smoking status (smokers vs. non-smokers) and lung cancer diagnosis (yes or no) (Table 1) and their theoretical probabilities are presented in Table 2.

Table 1: Contingency Table of Smoking Status and Lung Cancer Diagnosis

	Lung Cancer		Total
	Yes	No	
Smoking	155	19	174
Non-smoking	115	20	135
Total	270	39	309

Table 2: Theoretical Probabilities of Lung Cancer Diagnosis by Smoking Status

	Yes	No
Smoking	0.89	0.11
Non-smoking	0.85	0.15

Both the OR calculation and Monte Carlo simulations were carried out for this data set.

Calculate the Odds Ratio (OR):

$$\theta = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{(155 \cdot 20)}{(19 \cdot 115)} = 1.42$$

Compute empirical coverage rates:

By running R using the parameters $n_{1d}(n_{1+})$ = 174, $n_{2d}(n_{2+})$ = 135, p_1 = 0.89, and OR(θ) = 1.42, the six confidence interval methods were analyzed (Table 3).

Table 3: Empirical Coverage Rates for Lung Cancer Diagnosis Based on R

Method	Original (z)	Welch's adjustment (t)
Agresti (1999)	0.9511	0.9532
Woolf (1955)	0.9500	0.9532
Gart (1966)	0.9531	0.9565

Based on Table 3, the simulation results reveal that all six confidence interval methods achieve empirical coverage rates close to the nominal level of 95% for the large-large sample size scenario derived from the real-world dataset. Among these, Woolf original (zw) stands out as the most precise method, achieving an empirical coverage rate of exactly 95%. The other five methods also performed exceptionally well with coverage rates ranging from 95.11% (za) to 95.65% (tg), indicating a conservative yet reliable approach with any of the confidence interval methods.

The application of the real-world dataset is consistent with the findings from our analysis using 10,000 Monte Carlo simulations, ensuring that the conclusions drawn are both theoretically robust and practically relevant. Notably, our theoretical simulations reveal that the Agresti with Welch's adjustment (ta) and Woolf with Welch's adjustment (tz) methods exhibit similar performance in nearly all cases, except for scenarios involving small-small sample sizes. The results further demonstrate their consistency, as both methods achieve an empirical coverage rate of 0.9532, highlighting their reliability. Interestingly, although the Woolf original method (zw) is the least successful overall, it still achieves an empirical coverage rate of 0.95, which is remarkably close to the nominal $1 - \alpha = 0.95$, indicating that it remains a highly effective method.

Conduct a simulation study to assess the cases where the confidence level is something other than 95%

1. ***Simulation study to assess the cases at a 99% confidence level (i.e. $\alpha = 0.01$)***

   At the 99% confidence level ($\alpha = 0.01$), the Agresti method with Welch's adjustment (ta) and Woolf method with Welch's adjustment (tw) remain the most successful methods. These methods maintain their reliability and consistency even at this higher confidence level, achieving empirical coverage rates close to the nominal 0.99 level. The original Woolf method (zw) continues to perform adequately but is slightly less successful compared to ta and tw. The results from Figure 6 demonstrate that the performance hierarchy of the methods remains largely consistent with the findings at the 95% confidence level. This consistency suggests that these methods can be trusted to maintain appropriate coverage rates even when more stringent confidence requirements are needed. However, it's worth noting that as the confidence level increases, all methods tend to become more conservative, with empirical coverage rates generally higher than the nominal level.

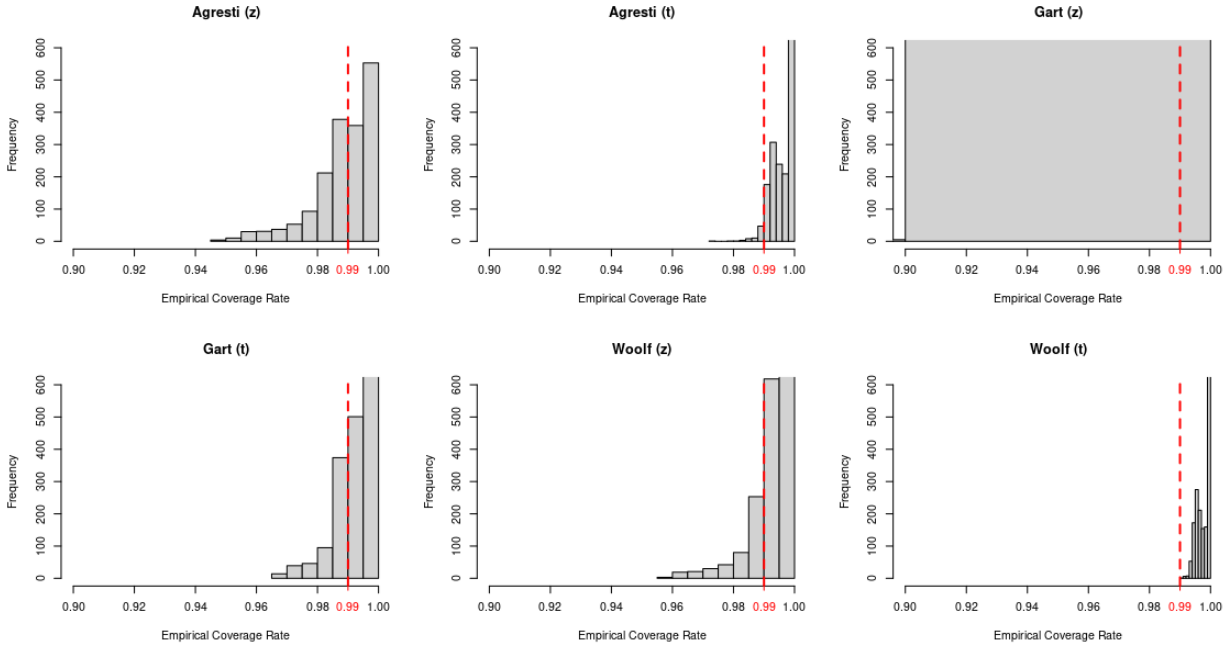

Figure 6. Histograms of empirical coverage rates for each of the six methods are shown with a vertical dashed line at 0.99, representing the nominal coverage rate set in the simulation studies. Graphs are labeled with respective OR methods used as well as "t" for Welch-adjusted methods or "z" for original methods.

2. *Simulation study to assess the cases at a 90% confidence level (i.e. α = 0.10)*

At the 90% confidence level (α = 0.10), the Agresti method with Welch's adjustment (ta), Woolf with Welch's adjustment (tw), and the original Woolf method (zw) continue to perform well, achieving empirical coverage rates close to the nominal 90% level. Among these, zw stands out as the most reliable method due to its favorable performance characteristics. Specifically, zw exhibits fewer data points below the nominal coverage rate (z < 0.9) compared to the other methods, indicating reduced undercoverage. Additionally, the empirical coverage rates for zw are tightly clustered around z = 0.9, demonstrating its precision and consistency at this confidence level.

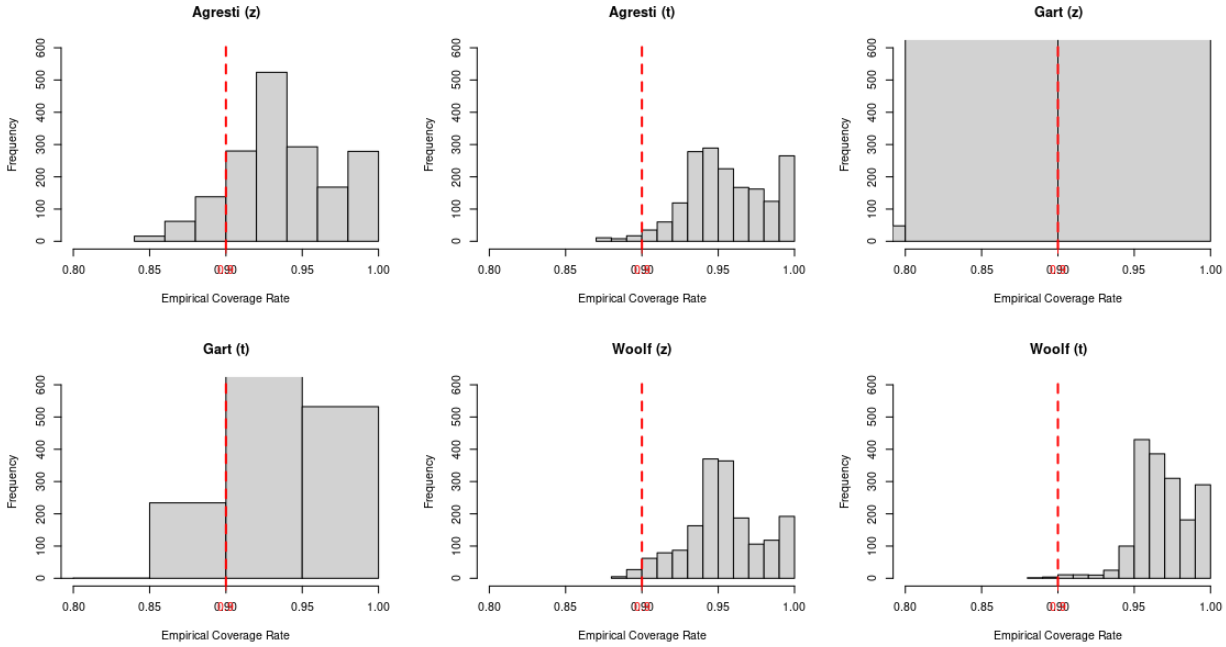

Figure 7. Histograms of empirical coverage rates for each of the six methods are shown with a vertical dashed line at 0.90, representing the nominal coverage rate set in the simulation studies. Graphs are labeled with respective OR methods used as well as "t" for Welch-adjusted methods or "z" for original methods

Conclusion

This project conducted a comprehensive simulation study to evaluate the reliability of six confidence interval methods for estimating odds ratios in 2×2 contingency tables: Woolf (original and Welch-adjusted), Gart (original and Welch-adjusted), and Agresti (original and Welch-adjusted). Through systematic exploration of various parameter combinations—including success probabilities (p_1), odds ratios (θ), and sample sizes (n_{1+}, n_{2+})—we identified scenarios where these methods perform reliably or unreliably.

Our findings reveal that Woolf with Welch's adjustment (tw) is the most robust method overall, achieving empirical coverage rates consistently close to the nominal level across diverse scenarios. Additionally, the original Woolf method (zw) performs exceptionally well in large-large sample size cases, particularly when applied to real-world lung cancer datasets and at reduced confidence levels (e.g., 90%). In these scenarios, zw achieves empirical coverage rates tightly clustered around the nominal level, demonstrating its precision and reliability. On the other hand, Agresti with Welch's adjustment (ta), while robust in most cases, struggles significantly in small-small sample size scenarios, where it exhibits undercoverage with empirical coverage rates falling well below the nominal level.

Overall, this study enhances our understanding of OR confidence intervals and provides practical guidance for their application in fields like medicine, epidemiology, and social sciences. By combining theoretical simulations with real-world data analysis, we have demonstrated the critical role of method selection in ensuring accurate and reliable statistical inference.

Suggested Further Discussion

In this study, we've conducted a simulation for calculating the Confidence interval by running Monte Carlo simulation and visualizing the outcomes. Through this process, we've concluded that the Agresti method with Welch's adjustment performed the best, however, it fails in the case of both population sizes being extremely small. Other successful methods include the Woolf confidence interval, both with and without Welch's adjustment. However, both Woolf methods showed decreased performance as the nominal coverage rate changes from 0.95. In future studies, we suggest exploring ways to improve the performance of these methods at different nominal coverage rates. This can be done through greater variation in both the success probability and the odds ratio values. We also suggest variation in these values to try to mitigate the failure of the adjusted Agresti method with extremely small sample sizes.

Reference

Agresti A. (1999). On logit confidence intervals for the odds ratio with small samples. *Biometrics 55*, 597—602.

Fagerland, M.W., Lydersen, S., and Laake, P. (2015). Recommended confidence intervals for two independent binomial proportions. *Statistical Methods in Medical Research 24*(2), 224—254.

Gart, J.J. (1966). Alternative analyses of contingency tables. *Journal of Royal Statistical Society, Series B: Statistical Methodology 28*, 164–179.

Lawson R. (2004). Small sample confidence intervals for the odds ratio. *Communications in Statistics: Simulation and Computation 33*(4), 1095—1113.

Woolf, B. (1955). On estimating the relation between blood group and disease. *Annals of Human Genetics 19*(4), 251—253.

Extra Credit: real-world dataset applied
Kaggle. (n.d.). *Lung Cancer Dataset* [Dataset]. Retrieved from https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer