Reliability of Popular Odds Ratio Confidence Intervals: A Comparison of Simulation Studies

Sydney Cambell

Zoe Absalonson

Parker Perry

Jalen Black

Western Washington University

Math 445: Computational Statistics

Dr. Kimihiro Noguchi

March, 2025

Abstract

Odds ratios are a vital statistical measure used to measure the association between an exposure and an outcome. They are very useful in medical and social sciences, specifically for case control studies and clinical trials in epidemiology, public health, economics, and more. In this paper, we explore the efficacy of three different odds ratio confidence intervals. Specifically Woolf's confidence interval, Gart's adjusted confidence interval, and Agresti's adjusted confidence interval (CI). This report aims to analyze the robustness of each CI by comparing the empirical coverage rate of each confidence interval to the nominal confidence level over a variety of parameters used to compute the CI's. To test the robustness of each interval, we tested the combinations of a variety of n1 sample sizes, n2 sample sizes, odds ratios, p1 probabilities, and p2 probabilities.

Background

An odds ratio (OR) is a statistical measure that measures the association between an exposure and an outcome. Typically, it's used to compare two groups, one with some exposure, and one without. It then measures the likeliness of an outcome between the different exposure groups. Its particularly useful in case-control studies, cohort studies, and clinical trials. If an odds ratio is greater than 1, that indicates that some exposure is associated with a higher chance of an outcome for one group when compared to another group. The OR can provide insights into how exposure to a certain factor (like a treatment, a risk factor, or a behavior) affects the likelihood of an outcome (such as disease occurrence or recovery).

An odds ratio provides a useful estimate of the relationship between an outcome and two groups, but determining the exact odds ratio of something can be difficult. Because of this, a 95% OR confidence interval is often used. This gives an approximate interval where a researcher can say the true odds ratio lies with some level of confidence. Narrower intervals indicate higher levels of confidence. These confidence intervals can be very useful for hypothesis testing; if 1 lies within the OR confidence interval, that implies there is no statistically significant evidence that an event is more likely with one group than another.

Multiple OR confidence intervals have been created. In this report, we focus on Woolf's confidence interval, Gart's adjustment, and Agresti's adjustment. Woolf's confidence is very popular and is expressed in terms of the estimated log OR. However, its actual coverage rates tend to be higher than the nominal confidence level when α=0.05 is chosen. For small sample sizes, it performs especially poorly, with actual coverage probabilities dipping below the nominal confidence level. Usually, the values don't dip too low however unless the OR is very large.

Garts adjustment, where 0.5 is added to each n_{ij}, is a common adjustment. Gart's method is typically used for rate ratios or odds ratios when the data involves small cell counts in contingency tables (especially with rare events). It applies a correction to the variance of the estimate, leading to more accurate confidence intervals. This is used more specifically in cases where small expected frequencies might lead to misleading results if using standard methods. Typically, Gart performs worse when the OR is far from 1 (>10 or <0.1). Garts adjustment to the OR confidence interval performs best when sample sizes of n1 and n2 are similar and the OR is small, such as less than 5.

The Agresti adjustment to the confidence interval is often used in situations involving proportions (e.g., binomial data), particularly when sample sizes are small or when the observed proportion is close to 0 or 1. This adjustment is preferred in these cases because it corrects for the bias that can occur when constructing confidence intervals for proportions, improving the accuracy and reliability of the interval estimates.

The reason to use the above OR confidence intervals instead of the Cornfield Exact is that they provide narrower intervals than the exact interval. If one can tolerate the relatively low risk of the coverage probability falling below the nominal confidence interval, the above OR confidence intervals will give more accurate estimates than the exact OR confidence interval. When sample sizes are small, all the above confidence intervals are somewhat conservative, with Woolf's being the most conservative and Gart's the least. When sample sizes increase, the confidence intervals behave similarly and coverage probabilities get closer to the nominal confidence level.

Due to the widespread use of logistic regression, the odds ratio is widely used in many fields such as medical, social, and economic fields. The odds ratio is commonly used in survey research, in epidemiology, and to express the results of some clinical trials, such as in case-control studies.

In medical research, ORs are often used to measure the strength of associations between exposures (like smoking or vaccination) and health outcomes (like cancer or heart disease). Confidence intervals around these ORs help clinicians assess whether a treatment or risk factor is likely to have a meaningful impact on patient outcomes. For example, in clinical trials, the OR might quantify the odds of recovery in treated vs. untreated patients, and the CI helps to evaluate the robustness of the findings.

In social science research, ORs are used to assess relationships between different binary outcomes, such as the likelihood of certain behaviors (e.g., voting, job acceptance) based on demographic factors (e.g., age, education). Confidence intervals are crucial for policymakers to understand the variability of these associations and make informed decisions

Economists use ORs to assess the likelihood of economic events (e.g., default on loans, adoption of new technology) based on factors like income or education. Confidence intervals help gauge the precision of these estimates, which is important for decision-making in policy and financial planning.

Overall, odds ratios are a powerful statistical tool, but the confidence intervals used to predict them are not perfect. Due to this, three approaches to calculating confidence intervals for odds ratios are analyzed below.

Methodology

Research Approach

Research Goals

This study examines the reliability of three commonly used confidence intervals for odds ratios – Woolf's, Gart's, and Agresti's – by assessing their empirical coverage rates. We evaluate how well these confidence intervals maintain their nominal coverage levels across different combinations of large, medium, and small sample sizes, varying ORs, and varying probabilities.

Justification

A quantitative research approach was chosen because this study involves statistical simulations to evaluate confidence interval performance. The simulations allow for systematic comparisons under controlled conditions, ensuring that the results are generalizable across a range of sample sizes.

Data Collection Methods

Data Source and Sampling Procedures

Data was generated through a Monte Carlo simulation, where 2x2 contingency tables were simulated under predefined conditions. The odds ratio (θ) was held constant at different values while varying sample sizes (n_{1+} and n_{2+}) and event probability for group 1 (p_1).

A computational approach was used instead of traditional sampling techniques. Each combination of parameters was simulated 10,000 times to estimate empirical coverage rates accurately. This large number of iterations ensures stability and robustness.

Statistical Analysis

Statistical Techniques

The study measured empirical coverage rates by computing the proportion of times each confidence interval contained the true odds ratio using 10,000 simulations. The following confidence intervals were evaluated:

- Standard normal (z) with Agresti's adjustment, Z-Agresti (za)
- Welch's t (t) with Agresti's adjustment, T-Agresti (ta)
- Standard normal (z) with Gart's adjustment, Z-Gart (zg)
- Welch's t (t) with Gart's adjustment, T-Gart (tg)
- Standard normal (z) with no adjustment (Woolf), Z-Woolf (zw)
- Welch's t (t) with no adjustment (Woolf), T-Woolf (tw)

Data Collection Tools

The statistical software used for this research was R, and the graphing package that was used for visualizing empirical coverage rates was ggplot2. Custom R scripts were provided for the purposes of this study in generating contingency tables, computing OR estimates, constructing confidence intervals, and evaluating empirical coverage rates.

Sample Size and Power

We tested a variety of sample size combinations to reflect real-world research conditions:

- Small: (3, 10, 20)
- Medium: (30, 35, 75)
- Large: (100, 600, 1000, 1005)

The large number of simulations (10,000 per condition) ensures high statistical power, making the results reliable.

Data Diagnostics

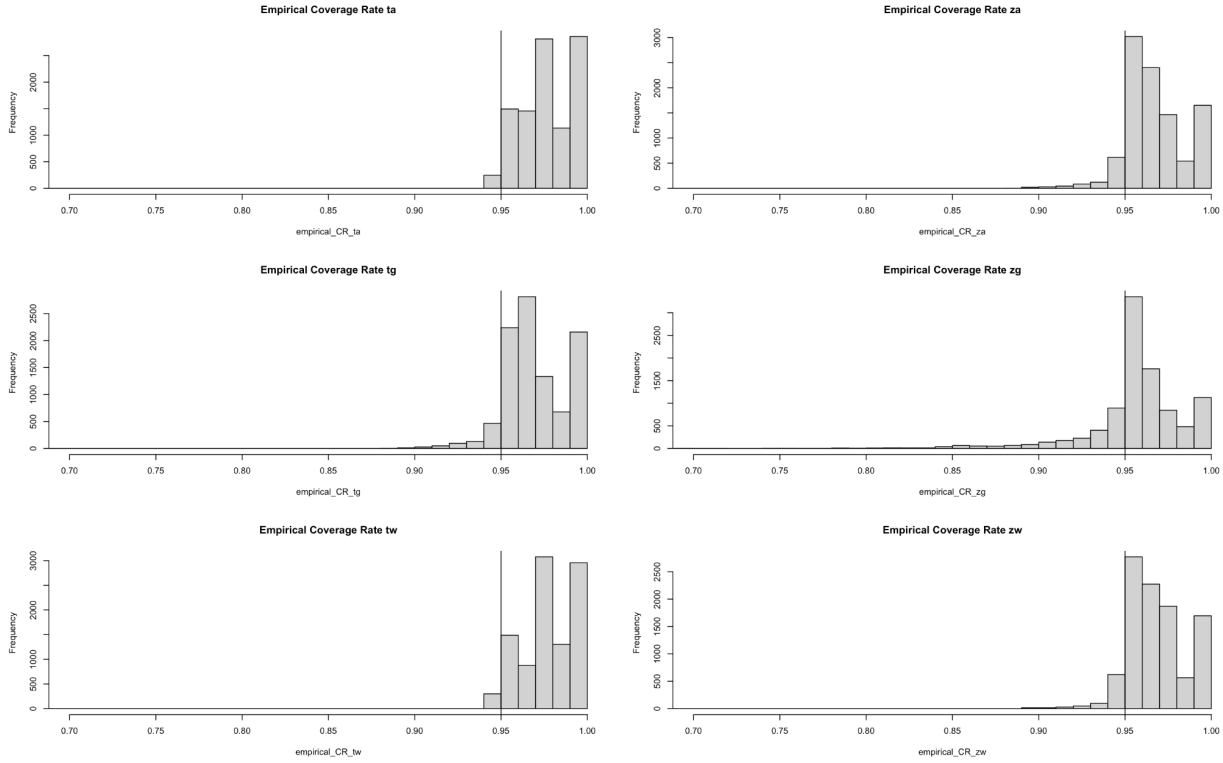
After computing the empirical coverage rates for each confidence interval, results were analyzed for consistency across different conditions. Graphs were used to visualize trends and confirm expected patterns. Anomalies or deviations from expected results were carefully examined.

Reliability

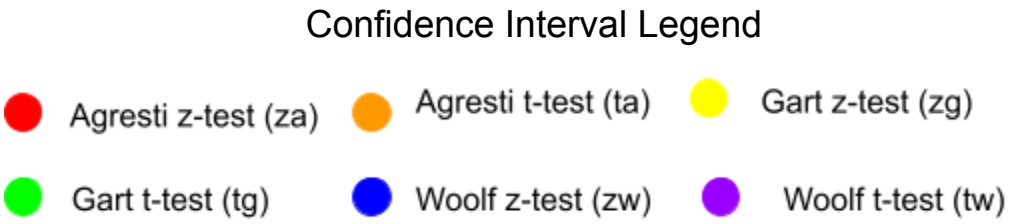
The use of 10,000 simulations per parameter combination ensures that the findings are replicable and not due to random variation. The entire analysis was conducted in R, with clearly defined procedures that can be rerun to obtain the same results. Results were compared against findings in existing research studies to confirm consistency with prior research.

Simulation Study Results

Below, histograms of each confidence interval display the approximate distribution of empirical coverage rates. In general, z intervals tended to dip below the nominal confidence level more frequently than t intervals, but also had less left skew, and are more centered around 0.95.



The data was split into six categories of combinations for analysis. The categories, with relevant analysis are displayed below.



Category 0

Fix θ, n_1 set to vary between large sizes (100, 600, 1000, 1005) and n_2 between medium sizes (30, 35, 75). Vary p_1 values.

When n_1 and n_2 were very far apart, such as 1000 and 75, the confidence intervals performed significantly differently for low p values, but as p approaches 1, the confidence intervals all converged to roughly the nominal confidence level as shown in Figure 0.1.

Gart's confidence interval with a standard normal distribution performed especially poorly for extreme p_1 values close to 0. For large odds ratio (θ) values and sample sizes with large differences, this poor performance was exacerbated. Figure 0.2 shows this trend.

Figure 0.1

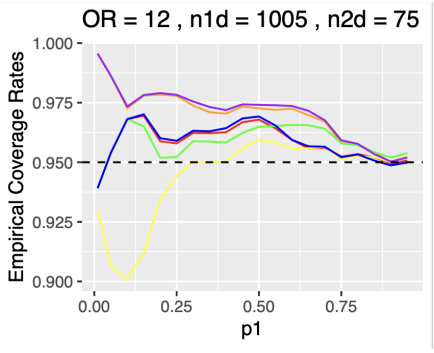
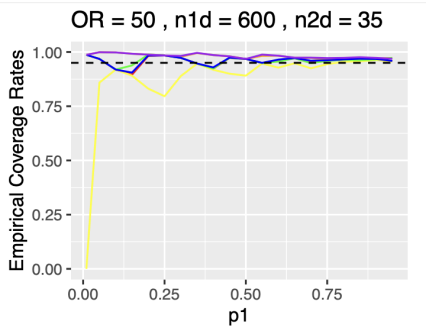


Figure 0.2



All confidence intervals performed best when the OR was relatively close to 1 and the sample sizes were similar. Figures 0.3 and 0.4 illustrate this. Although they may dip somewhat below the nominal confidence level, it happens infrequently and they don't dip below by very much in these cases.

Figure 0.3

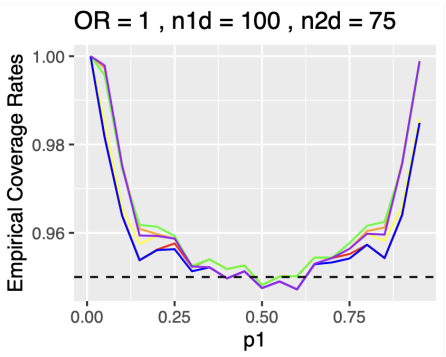
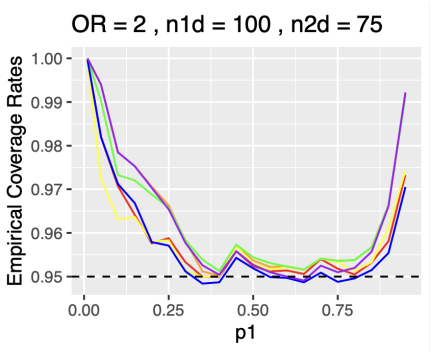


Figure 0.4



Category 1

Fix θ, n_1, n_2, set to vary between small size (3, 10, 20) ranges. Vary p_1 values.

When n_1 and n_2 values were both under 30 and odds ratios were high (θ = 50) , interesting graphs were generated. As seen in Figures 1.1, and 1.2 below, three of the coverage rates dipped below the standard nominal coverage rate of 95%. The yellow and red lines, representing the coverage rate for the Gart-adjusted standard normal test and the Agresti-adjusted standard normal test respectively, both had more liberal coverage rates than the others. The green line, indicating the Gart-adjusted t-test, had a significant dip in coverage rates when p-values were between 0.25 to 0.6 however it managed to pull itself back into the comfortable 0.95 coverage rate range as probability grew closer to one.

Figure 1.1

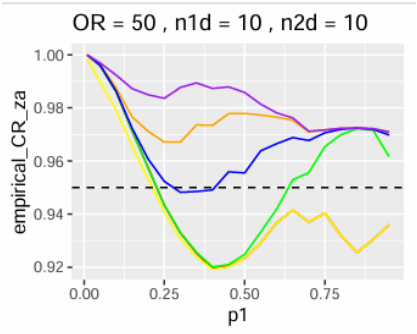


Figure 1.2

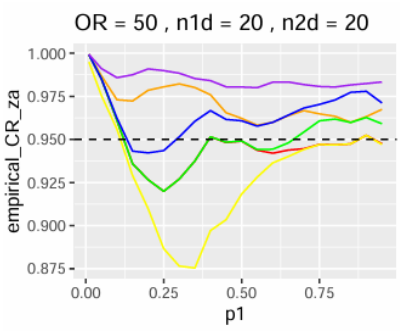


Figure 1.3 had two, the Gart-adjusted standard normal test and the Agresti-adjusted standard normal test, go below the standard coverage rate of 0.95. As probability increased these two intervals went as low as having a coverage rate of under 0.92. However, the majority of the tests were liberal, staying above the 0.95 coverage rate even with an extremely small sample size value.

Figure 1.4 had no confidence interval coverage rates dip below the standard rate of 0.95. Even acting slightly liberal by keeping the coverage rates closer to 0.96 than 0.95, most

likely due to how smaller sample sizes widen the empirical coverage range because its relationship to margin of error $z_{\alpha/2} * \sigma/\sqrt{n}$.

Figure 1.3

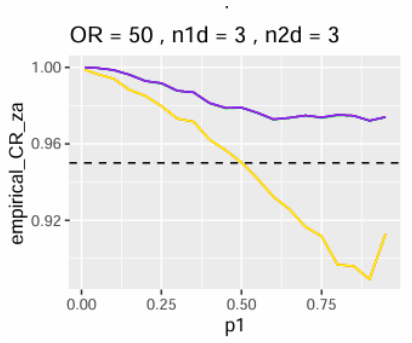
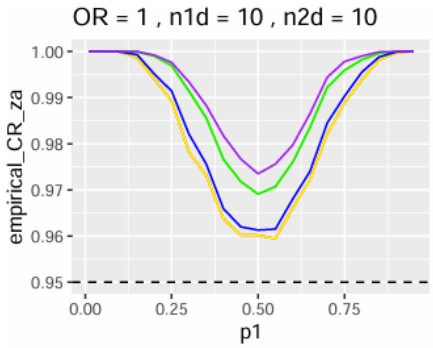


Figure 1.4



Category 2

Fix θ, n_1, n_2, set to vary between small size (3, 10, 20) and medium size (30, 35, 75). Vary p_1 values.

45 graphs were generated in total from each of the θ, n_1, n_2 combinations. When θ = 12, the graphs followed an interesting pattern. When probability increased to 1 all size empirical coverage rates converged to the nominal coverage rate of 0.95 (Figure 2.1). This convergence could be an indication of stabilization of the tests since this convergence occurred regardless of the difference in the sample size. Figure 2.3 has a similar occurrence but it is extreme since the sample sizes are very far apart and small sample sizes result in more liberal coverage rates.

Figure 2.1

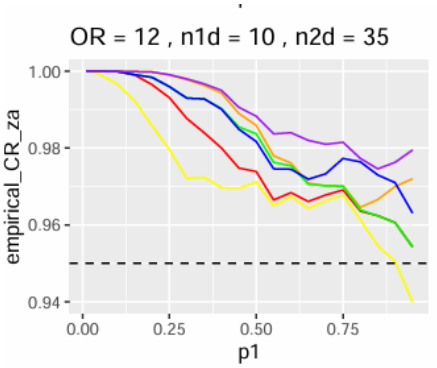
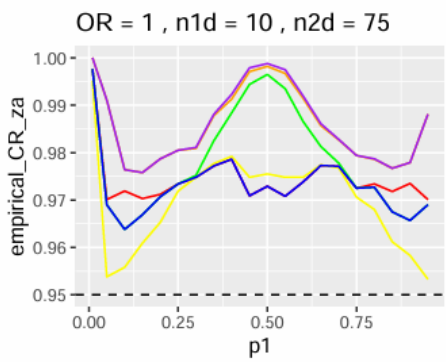


Figure 2.2



In Figure 2.2, while none of the coverage rates went below the nominal coverage rate there is a unique bump to a coverage rate of 1 for when probability is equal to 0.5. This is a common phenomenon interestingly enough known as the symmetry of the binomial distribution (Buchan,

lain). Since our simulations are of the binomial distribution we can see that when probability is equal to 0.5 it maximizes the variance. Though variance is still dependent on our sample size, maintaining the relationship $n/4$, derived from the equation for variance of the binomial distribution. Having a small sample size with this relationship seems to result in coverage rates of 1, since smaller sample sizes tend to increase the margin of error.

When odds ratios were less than 10 and sample size was higher than 20 the empirical rates behaved more predictable even when the two sample sizes were far apart (Figure 2.4).

Figure 2.3

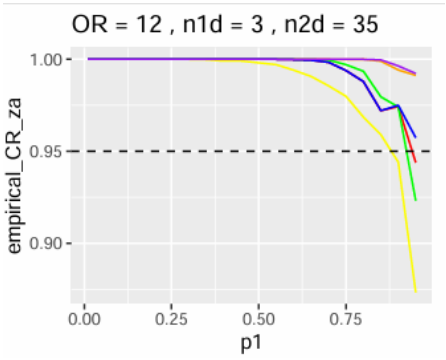
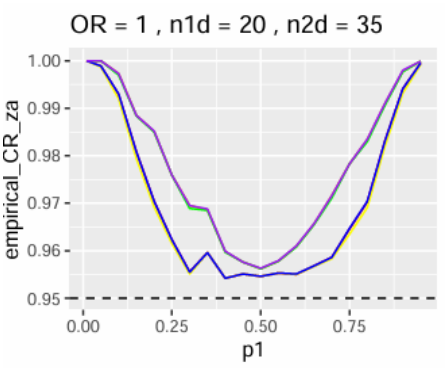


Figure 2.4



Category 3

Fix θ, n_1, n_2, set to vary between large size (100, 600, 1000, 1005) and small size (3, 10, 20). Vary p_1 values.

When analyzing cases where the first sample size (n_1) is small (3, 10, 20) and the second sample size (n_2) is large (100, 600, 1000, 1005), we observe interesting trends in empirical coverage rates for the confidence intervals.

For extremely different sample sizes, such as $n_1 = 3$ and $n_2 = 1000$, the results become less informative due to the extreme imbalance. More realistic cases, such as $n_1 = 20$ and $n_2 = 100$, provide a more meaningful analysis.

A consistent pattern emerged when holding OR (θ) at 1 and n_2 at 100 while increasing n_1 from 10 to 20. This resulted in a noticeable transformation in the shape of the coverage rate graphs as seen in Figures 3.1 and 3.2. When $n_1 = 10$ (Fig. 3.1), the coverage rate curves were concave down, performing best for mid-range p_1 values and worse (yet still above 0.95) for extreme p_1 values. However, when $n_1 = 20$ (Fig. 3.2), the pattern flipped, with coverage rates against p_1 values appearing to be concave upwards, performing worst at mid-range p_1 values and improving towards the extremes.

Figure 3.1

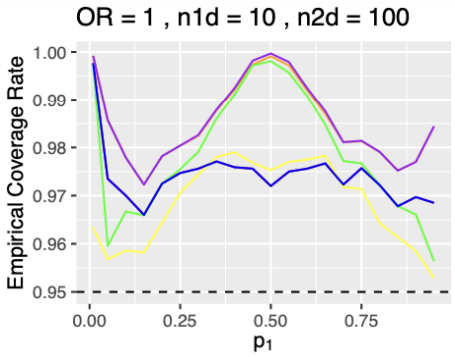
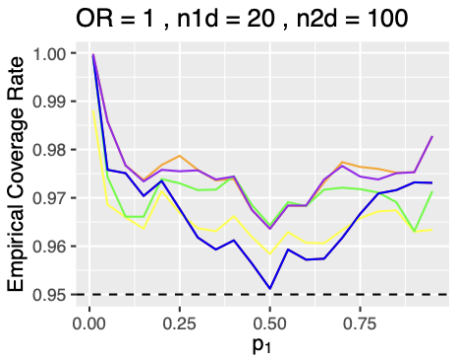


Figure 3.2



Another interesting trend was observed when increasing OR (θ) from 5 to 12 to 50 while holding n_1 at 20 and n_2 at 100. All three of these graphs, represented by Figures 3.3, 3.4, and 3.5, exhibited a similar shape: empirical coverage rates were closest to 1 for low p_1 values, nearest to the nominal coverage rate, 0.95 (yet still above), for mid-range p_1 values, between 1 and 0.95 for p_1 values around 0.75, and worst for p_1 values beyond that, dipping far below 0.95. T-Woolf (tw), T-Agresti (ta), and T-Gart (tg) performed slightly better than the other confidence intervals for OR values of 5 and 12 (see Figures 3.3 and 3.4), but T-Gart performed significantly worse for OR $= 50$ (see Figure 3.5). Meanwhile, T-Woolf and T-Agresti perform, in this case significantly, better than the other confidence intervals. Z-Gart (zg) consistently performed the worst across all graphs.

Figure 3.3

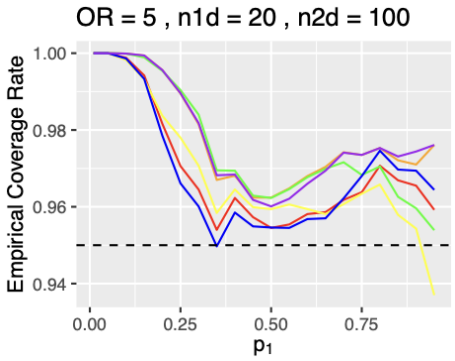


Figure 3.4

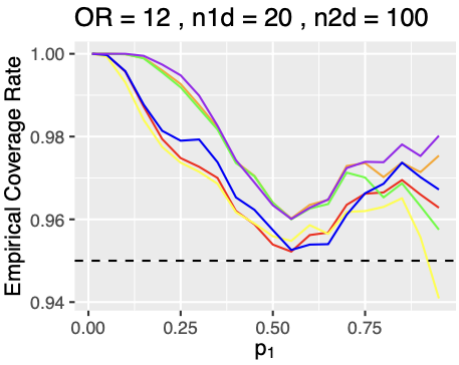
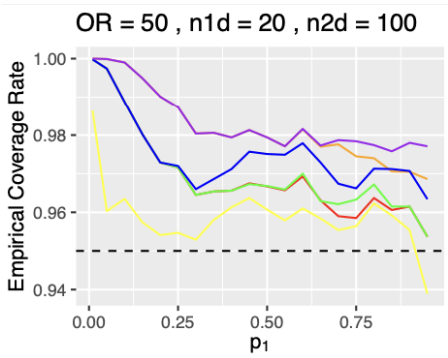


Figure 3.5



Across all these comparisons, Z-Gart consistently performed the worst among all confidence intervals, further reinforcing its instability in cases with highly imbalanced sample sizes.

Category 4

Fix θ, n_1, n_2, set to vary between large size range (100, 600, 1000, 1005). Vary p_1 values.

To further examine the behavior of confidence intervals, we analyzed cases where both sample sizes were large (n_1 and $n_2 = 100, 600, 1000, 1005$) to observe how empirical coverage rates change under these conditions with a loose goal of determining whether larger sample sizes stabilized coverage rates and how different odds ratios (OR) influenced interval performance.

When examining vastly different sample sizes (e.g. $n_1 = 1005$, $n_2 = 100$ or $n_1 = 1000$, $n_2 = 100$), as represented by Figures 4.1, 4.2, 4.3, and 4.4, it can be seen that all of the confidence intervals exhibit markedly different performance for small p_1 values, with Z-Gart (zg) having remarkably low performance for these p_1 values, yet all confidence intervals appear to gradually converge to the nominal coverage rate, 0.95, as p_1 approaches 1. This effect becomes even more pronounced for OR (θ) values of 12 or greater, as seen in Figures 4.3 and 4.4, where the deviations from 0.95 are more drastic for low p_1 values.

Figure 4.1

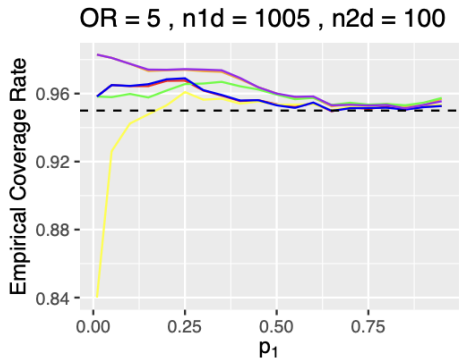


Figure 4.2

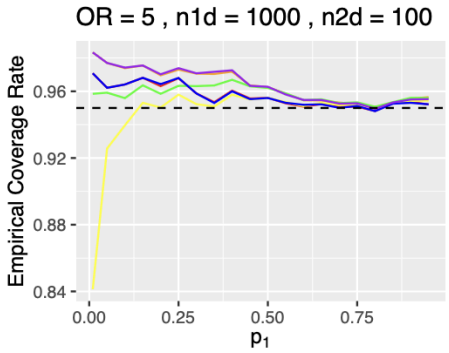


Figure 4.3

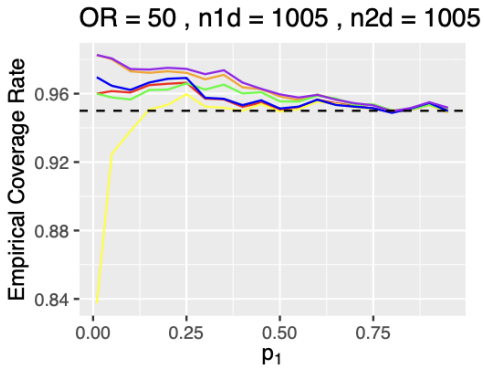
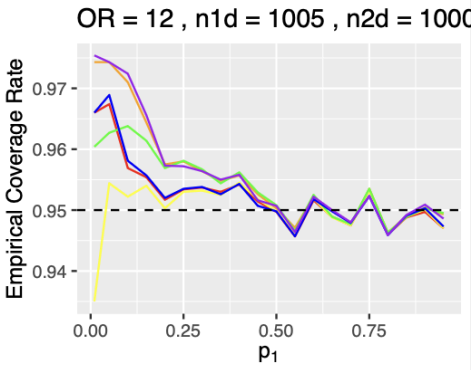


Figure 4.4



Among the tested intervals, the T-Woolf (tw) and T-Agresti (ta) performed slightly better in terms of empirical coverage rates compared to the others. In Figures 4.5 and 4.6, this advantage became more apparent. When the OR value was set at or exceeded 12 these two intervals maintained higher, more stable coverage rates relative to their counterparts, especially for low p_1 values.

Figure 4.5

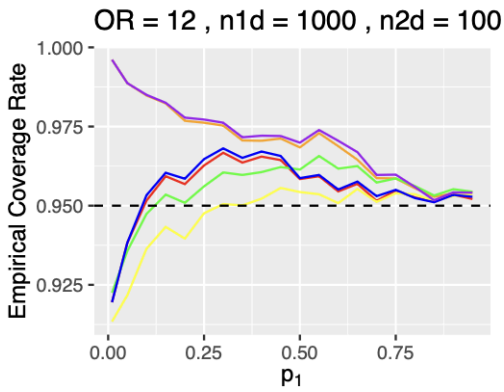
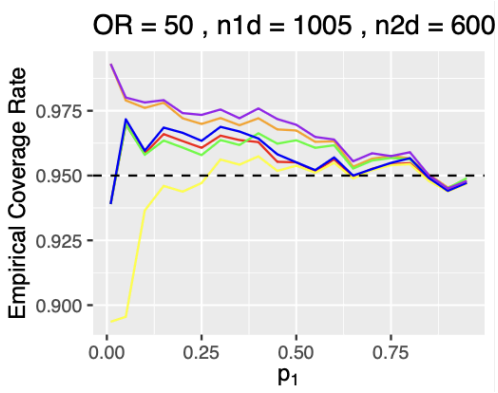


Figure 4.6



Overall, T-Woolf (tw) and T-Agresti (ta) had significantly better performances than the other confidence intervals in terms of empirical coverage rate. Adjustments in confidence interval selection are necessary here in optimizing empirical coverage behavior, especially when OR values are high and/or sample sizes are exceedingly unbalanced.

Category 5

Fix θ, n_1, n_2, set to vary between medium size ranges. Vary p_1 values.

For small θ values, results generated with equal n_1 and n_2 values ($n_1 = n_2$) in the medium range tend to behave normally with no dips below the coverage rate of 95% as shown in Figure 5.1. As we increase the θ to more extreme values, we see all tests start to perform worse.

Notably, for θ values 12+, Z-Gart is the first to show worse results. For a θ of 50, we see the Z-Agresti Z-Woolf, and T-Gart followed suit. The T-Agresti and T-Woolf tests seemed to perform better even in these scenarios.

Figure 5.1

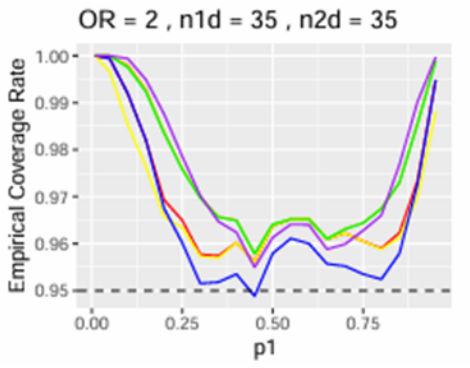
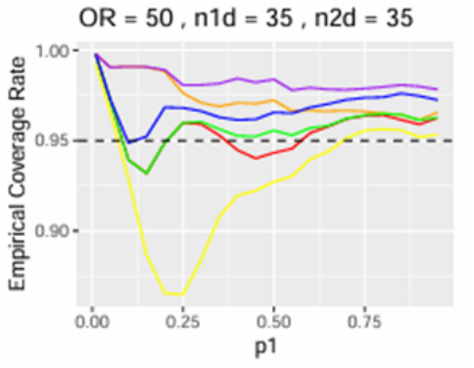


Figure 5.2



Additionally, an interesting finding shown in Figures 5.3 and 5.4 is when we differ n_1 and n_2. Low values of n_1 (30), and larger values for n_2 (75) yielded promising results. In all cases, every test except the Z-Gart test managed to stay above a coverage rate of 95%. However, when we did the opposite ($n_1 = 75, n_2 = 30$) we got slightly different results. For all θ under 50, all tests except the Z-Gart did decently, with some interesting dips around a p_1 value of 0.75. For a $\theta = 5$ (As seen in Figure 5.4), the Z-Gart test performed very poorly for all p_1 values less than 0.5. The Z-Agresti and Z-Woolf tests also dipped below the 95% coverage rate periodically.

Figure 5.3

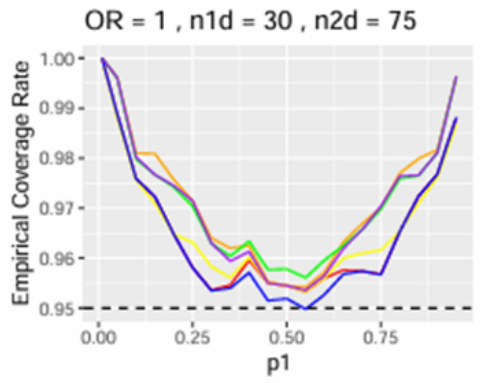
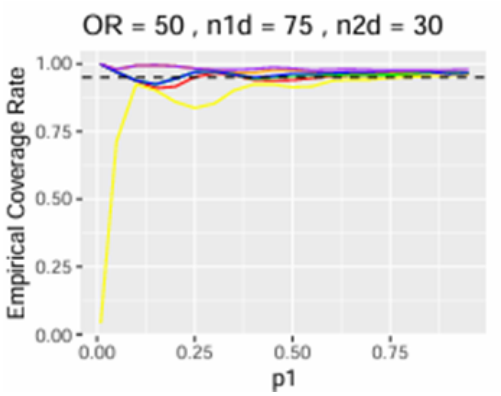


Figure 5.4



Category 6

Fix θ and vary $n_1 = n_2$ for all sample sizes.

Similar to previous results we see how low θ values paired with low $n_1 = n_2$ values are exceptionally liberal for all tests with deviations as θ is increased. For medium $n_1 = n_2$ refer to 5.

Large values of $n_1 = n_2$ were only slightly more interesting (refer to section 4).

For all low θ values, there are minor deviances bouncing just slightly above and below the coverage rate threshold of 95% (refer to Figures 6.1 and 6.2). There appear to be dips around p_1 values of 0.25, ~0.5 and ~0.75, which may warrant further research. Large θ values and large $n_1 = n_2$ indicate very liberal results (apart from the Z-Gart test) as seen in Figures 6.3 and 6.4. Overall, there seems to be little to no reason to make any judgments on varying sample sizes for $n_1 = n_2$.

Figure 6.1

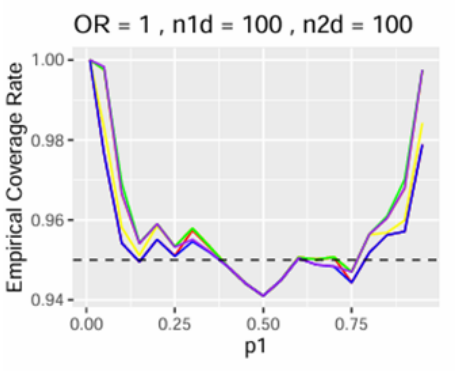


Figure 6.2

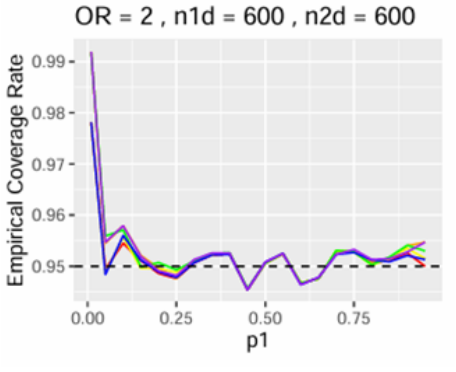


Figure 6.3

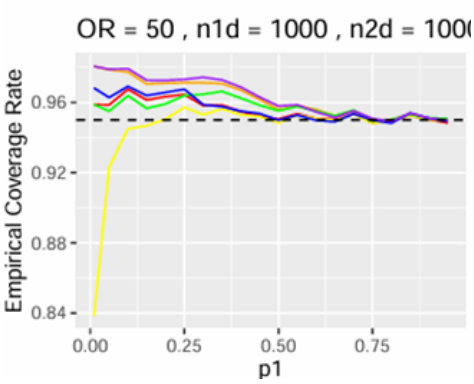
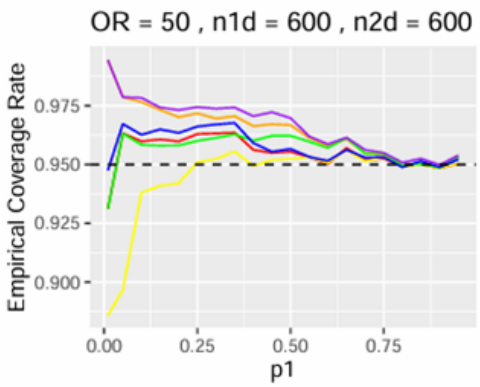


Figure 6.4



Discussion

Our results confirmed some of the assumptions we made about the Woolf, Gart and Agresti confidence intervals. One of the most prevalent trends, observed across all confidence levels, was that Gart's confidence interval with a standard normal distribution performed extremely poorly, generally being significantly more liberal than the other confidence intervals, with its empirical coverage rate frequently dropping below the nominal confidence level. When sample sizes were similar (or the same), specifically the T-Gart did well, but not as well, as the T-Woolf test. Both of the Agresti confidence intervals performed adequately for small sample sizes.

Multiple trends are visible from the above graphs. One being convergence of all tests towards a flat empirical coverage rate as p_1 values approach extremes. The tests converged as $p_1$values approached 1 when n_1 was significantly larger than n_2, and converged when p1 values approached 0 when n_2 was significantly larger than n_1. Another was downward trends showing high to low empirical coverage rates as p_1 values increase. Concave and convex parabolic

(symmetrical) results, which were even on both tails but with empirical coverage rates straying lower/higher respectively, were common for odds ratios closer to 1 and similar sample sizes. Lastly, we observed as seen in figure 2.3 extremely high empirical coverage rates which then dropped drastically.

In the case of peaks or troughs at $p_1 = 0.5$, we may be able to attribute this to the symmetric nature of the binomial distribution. At $p_1 = 0.5$, the variance is the greatest which results in the wider empirical coverage rates. For Woolf tests that yield high enough empirical coverage rates when theta was high, this may be due to a shrinking variance that occurs when we take the log transformation (log odds). The confidence intervals should be narrower and thus the empirical coverage rate should converge. With the Gart tests, adjustments are centered around better predicting small sample sizes, however as the p_1 value increases, the adjustment takes a significantly less effect and so we may see convergence.

Each test comes with its own set of limitations as we have seen throughout our models, we struggle in many cases to estimate the empirical coverage rates for all sample size ranges with high odds ratios. Although the Agresti tests were modified to better handle these situations, future research should be conducted to include ratios for events that have extreme likeliness to occur over others. We have also seen weak, liberal results in the event that one of our sample sizes is small, and the other is sufficiently large. The assumption of symmetry is a downfall of the Z tests as this relies on large sample sizes to be able to give good estimates. This is another reason why the Z-Gart test consistently was underperforming compared to the other (Microbe Notes. n.d.) . And why the Welch's t tests may have done well (Statology, n.d).

The assumptions made by these tests are reasonable, but show that there are flaws, and results generated need to be closely looked at. Further study should also be done into other possible distributions that would be more forgiving in large differentials in sample sizes. Similarly it would be interesting to investigate perhaps other transformations of the odds ratio that could prove helpful in these situations.

Additional Details and Applications

Lung cancer death rates hit an all time high towards the 1990's (WHO Mortality Database) coincidentally corresponding to when tobacco companies spent $4.6 billions on advertising for cigarette products (Institute of Medicine (US) Committee on Preventing Nicotine Addiction in Children and Youths). Multiple studies have shown that smoking has been linked to lung cancer, which is to say that smoking directly increases a person's odds of developing lung cancer "male smokers have 21 times the risk of dying from lung cancer as those who have never smoked" (Dattani, Saloni). Considering 13% of all men in the U.S. are smokers, methods like Woolf's odds ratio confidence interval and adjusted methods, such as Gart or Agresti intervals, can be extremely useful for helping quantify uncertainty in risk estimates, aiding public health decisions (Center for Disease and Prevention Control). Of course each one of these confidence intervals comes with its own set of limitations, Woolf's confidence interval tends to be "too wide rather than too narrow" indicating that these nominal coverage rates are often conservative (Lawson, p.1109). Gart's interval does well with smaller and medium sample sizes but has trouble when $\theta \geq 8$ which can result in low coverage probability. Agresti performwell with small and large sample sizes but struggles with unequal sample sizes, often resulting in conservative coverage rates (Fagerland et. al., p.23, 29)

The article "Review of alternative approaches to calculation of a confidence interval for the odds ratio of a 2 × 2 contingency table" by Graeme D. Ruxton, and Markus Neuhäuser explore different methods for calculating confidence intervals for odds ratio in 2x2 contingency tables. Specifically, it uses Fisher's exact test, Woolf's test, Blaker's test, and an R package that makes use of epitools. Since the only overlap between this paper and our study is Woolf's odds ratio confidence interval the same conclusion can be drawn for both studies. As seen in every one of our figures, Woolf's coverage rates stayed conservative throughout each combination of n_1 , n_2 and θ value. Equivalently, in Ruxton and Neuhäuser's simulation study Woolf's coverage rates also stayed conservative and remained an attractive alternative to their more complex confidence interval coverage rate calculations.

References

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

2. 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.

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

4. WHO Mortality Database. (2024). *Lung cancer death rates*. Our World in Data. https://ourworldindata.org/grapher/lung-cancer-deaths-per-100000-by-sex-1950-2002

5. Institute of Medicine (US) Committee on Preventing Nicotine Addiction in Children and Youths; Lynch BS, Bonnie RJ, editors. Growing up Tobacco Free: Preventing Nicotine Addiction in Children and Youths. Washington (DC): National Academies Press (US); 1994. 4, TOBACCO ADVERTISING AND PROMOTION. Available from: https://www.ncbi.nlm.nih.gov/books/NBK236761/

6. Saloni Dattani (2023) - “Risk ratios, odds ratios, risk differences: How do researchers calculate the risk from a risk factor?” Published online at OurWorldinData.org. Retrieved from: 'https://ourworldindata.org/risk-ratios-odds-ratios-risk-differences-how-do-researchers-calculate-the-risk-from-a-risk-factor' [Online Resource]

7. Institute of Medicine (US) Committee on Preventing Nicotine Addiction in Children and Youths. (1994, January 1). *Tobacco advertising and promotion*. Growing up Tobacco Free: Preventing Nicotine Addiction in Children and Youths. https://www.ncbi.nlm.nih.gov/books/NBK236761/

8. Buchan, I. E. (n.d.). *Binomial distribution*. StatsDirect. https://www.statsdirect.com/help/distributions/binomial.htm

9. Microbe Notes. (n.d.). *Z test vs t test: 8 major differences*. Microbe Notes. Retrieved March 21, 2025, from https://microbenotes.com/z-test/#z-test-vs-t-test-8-major-differences

10. Statology, (n.d) *Welch's t-test: Definition, formula, and example*. Statology. Retrieved March 21, 2025, from https://www.statology.org/welchs-t-test/