Reliability of Odds Ratios Confidence Intervals and Welch's correction among different scenarios

Bryce Anderson, Miles Standish, Lormel Bationo, Jordon Silcox

March 2025

1 Abstract

This paper seeks to determine the reliability of three different odds ratios: Woolfe's OR confidence interval, Gart's OR confidence interval, and the Agresti OR confidence interval, using a variety of normal and extreme parameters. Along with finding reliable confidence intervals, this paper also seeks to determine the performance of the two-sample t-test Welch's adjustment when applied to the three different OR confidence intervals. Through our study, we found that the Gart's OR confidence interval performed the best when dealing with small sample sizes and odds ratios. We also found that Welch's adjustment helped provide more consistent confidence intervals when working with small odds ratios. To achieve these results we performed 10,000 Monte Carlo simulations on 2352 different combinations of our chosen parameter values.

2 Background Information

To better understand the performance of different odds ratio confidence intervals, it is helpful to look at some of the history of each method. The following will introduce and briefly explain/motivate all of the methods referred to in this study.

2.1 Woolf OR Confidence Interval

The Woolf Confidence Interval, or Logit Confidence Interval, was first introduced in Bernard Woolf's paper "On Estimating the Relation Between Blood Group and Disease." In it, he tested for the heterogeneity of the Odds Ratio. The need for this arose from the challenge of accurately predicting the relationship between blood groups and disease, as it was expected that many more studies on the topic would be conducted. The motivation for increased accuracy stemmed from a method by Aird et al., which allowed for varying results even when the attack rate within a blood group remained constant (Woolf 1955). Although adjustments have been made to this confidence interval—such as the Agresti and Gart adjusted logit interval—methods like these have been widely used in several fields. Biostatistics, statistics, genetics, and economics are notable examples. The original logit confidence interval is:

$$\log\theta \pm z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}} \qquad \text{Equation 1}$$

2.2 Gart OR confidence interval

The Gart confidence interval, also known as the Haldane-Ans combe correction, is a method developed by Gart and Nam to construct confidence intervals for the odds ratio in stratified comparisons of binomial proportions. This method addresses the challenge of achieving accuracy in confidence intervals, particularly when working with stratified data (categorized into distinct groups). Their work aims to reduce potential skewness and propose a better alternative for odds ratio confidence interval calculation. Their work is a modification of a previous study by Farrington and Manning(1990) that came up with a difference test that corrected skewness.

The skewness-corrected test statistics Z_{GN} is the appropriate solution to the quadratic equation : $(-\tilde{\psi})\cdot Z_{GNR}^2+(-1)Z_{GNR}+(Z_{GNR})+Z_{FMR}(\phi)+\tilde{\psi})=0$

The confidence interval formula for odds ratio suggested by Wolf(1955)based on the delta method and expressed in equation 1 struggles with an overestimated coverage rate that often exceeds the theoretical level of 95% when $\alpha = 0.05$. A

popular adjustment to Woolf's research is suggested by Gart (1966) with the following equation:

$$log\tilde{\theta} \pm Z_{\alpha/2}\sqrt{\frac{1}{\tilde{n}_{11}} + \frac{1}{\tilde{n}_{22}} + \frac{1}{\tilde{n}_{21}} + \frac{1}{\tilde{n}_{22}}} \qquad \text{Equation 2}$$

where $\tilde{n}_{ij} = n_{ij} + 0.5$ and $\tilde{\theta}$ is estimated using $\tilde{n}_{ij}$ instead of n_{ij}. After that the upper and lower bound are exponentiated to obtain the confidence interval for the odds ratio.

2.3 Agresti OR confidence interval

When presented with data that has a small sample size or n_{ij} cell counts that are close to zero an adjustment may be needed to obtain a reliable coverage rate. Different interval adjustments have been used to deal with the inefficiency of the Woolf's confidence interval in such scenarios like the Gart adjusted OR confidence interval or an adaption to the Gart's adjustment is the Agresti OR confidence interval (Agresti 1999), also commonly called the Independent-smoothed logit confidence interval. As opposed to the Gart adjustment that changes each cell using a constant value, $c = 0.5$, the independent-smoothed logit confidence interval takes account of elements in the contingency table to make its adjustment. The independent-smoothed logit confidence interval is presented as:

$$\log\breve{\theta} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{1}{\breve{n}_{11}} + \frac{1}{\breve{n}_{12}} + \frac{1}{\breve{n}_{21}} + \frac{1}{\breve{n}_{22}}} \qquad \text{Equation 3}$$

where $\breve{n}_{ij} = n_{ij} + c_{ij}$ and $c_{ij} = \frac{2n_{i+}n_{+j}}{N^2}$. As shown in Eq. 3 the independent-

smoothed logit confidence varies from Gart's confidence adjustment. As the c_{ij} is an adjustment to each cell using the row sum and column sum, assuming there is independence between rows and columns, also it is divided by N^2 which is the total sample size. As commented before this adjustment works well when sample sizes are small which can happen in medical research like epinephrine in cardiac arrest with children (Fargerland et. al 2015) .

2.4 Welch's Adjustment

For all of our OR confidence intervals there is another adjustment that can be done which is Welch's adjustment. The Welch's t-test (Welch 1949) also called the Welch's adjustment comes from the two sample t-test that is used when the assumption of equal variance is violated or the data set has unequal sample sizes. The assumption of equal variance can be dropped as the t-test accounts for this in the degrees of freedom which also accounts for the varying sample sizes. Implementing Welch's adjustment into the OR confidence changes the test statistic from a Z-statistic to a T-statistic that is used for the calculation

of the confidence interval. The modified OR confidence is:

$$\log \hat{\theta} \pm t_{\frac{\alpha}{2},v} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \qquad \text{Equation 4}$$

$$v = \frac{2Nf_1^2}{f_3 - f_1^2}$$

$$f_j = (\frac{N}{n_{11}})^j + (\frac{N}{n_{12}})^j + (\frac{N}{n_{21}})^j + (\frac{N}{n_{22}})^j$$

The Welch's adjustment changes the normal degree of freedom (df) of $n-1$ to the v parameter shown. Like the Agresti OR confidence interval this adjustment uses information from the contingency table to find the proper df for the T-statistic.

3 Simulation Study

3.1 Overview

The monte carlo simulation with be using a total of 2352 different combinations. The results are stored in a table from which we can do analysis on.

Parameter	**Values**	**Number of Values**
n_{1d}	5, 10, 50, 100, 500, 1000, 2000	7
n_{2d}	5, 10, 50, 100, 500, 1000, 2000	7
p (i.e., p_1)	0.05, 0.10, 0.30, 0.50, 0.70, 0.90	6
OR	.001, .01, .3, .5, .7, .95, .99, .999	8

Table 1: Overview of Parameter Combinations

3.2 Preliminary Results

From these histograms we can see how liberal and conservative each tests seems to from our simulations. One test that looks interesting is the ZG as it has very good coverage rates for most cases, however it completely fails a small percentage of the time.
Now we can visualize the data we have and interpret the plots to analyze which tests seem to perform the best, and note any interesting results. Firstly, we will represent the data via a 3D scatter plot.

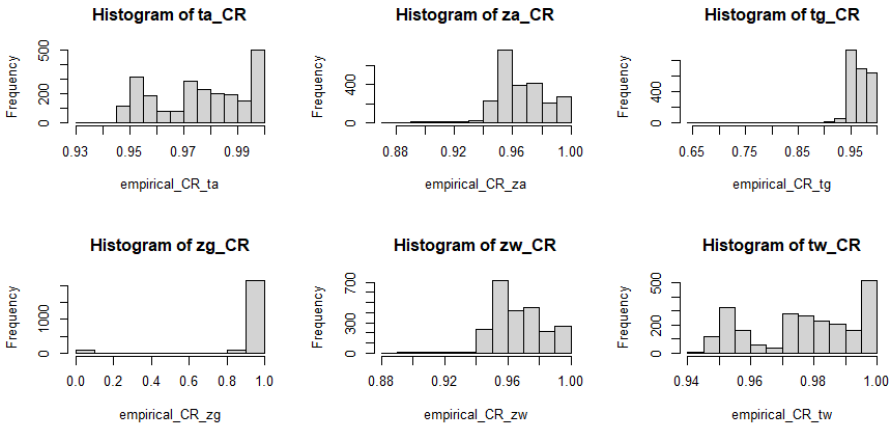

Figure 1: Histograms of the Tests

3.2.1 Agretsi Test

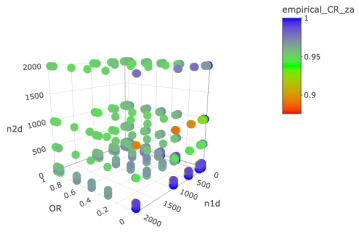

Figure 2: Agresti OR confidence interval (ZA_{OR})

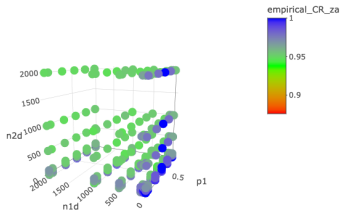

Figure 3: Agresti OR confidence interval (ZA_{P1})

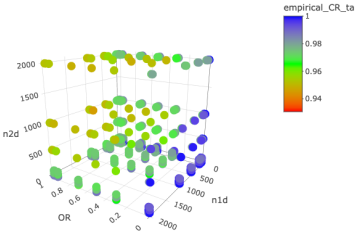

Figure 4: Agresti OR with Welch's adjustment (TA_{OR})

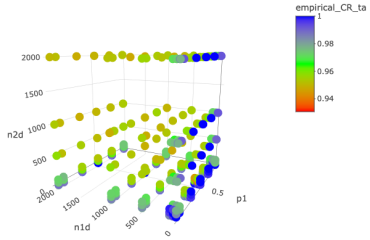

Figure 5: Agresti OR with Welch's adjustment (TA_{P1})

ZA_{OR}- From figure 2 the normal Agresti OR adjustment given odds ratio on the y-axis, n1d sample sizes on the x-axis and n2d sample sizes on the z-axis. It shows that as the OR and sample size gets smaller the adjustment improves the empirical coverage rate. Alternatively as the OR and sample size increases the coverage rate reduces. The adjustment appears to not perform as well when the OR is 0.001 and sample sizes are 500 or above, the coverage rate diminishes considerably.

ZA_{P1}- From figure 3 using sample size and the P1 probability, the Agresti adjustment performed best when the probabilities or sample sizes are small. As both of these parameters increase the empirical coverage rate decreases. Overall the empirical coverage rate is fairly consistent as the sample size and probability increases.

TA_{OR}- From figure 4 with Welch's adjustment using the new adjusted cell count improves the coverage rate when the OR is small and the sample size increases. As mentioned, the Welch's adjustment increased the coverage rate where the Agresti lacked when the OR was held very small at 0.001 while the sample size was 500 or above the coverage rate went from 90 or below to close to 100.

ZA_{P1}- From figure 5 using sample size and the P1 probability, the Agresti adjustment performed best when the probabilities or sample sizes are small. As both of these parameters increase the empirical coverage rate decreases. Overall the empirical coverage rate is fairly consistent as the sample size and probability increases.

TA_{P1}- Plugging in Welch's adjustment into the Agresti OR adjustment creates a wider range of coverage rates as the parameters change. It appears to perform worse as both sample sizes increase, but if the sample sizes are drastically different like n1d=2000 and n2d= 5 the Welch's adjustment still performs adequately.

3.2.2 Gart Test

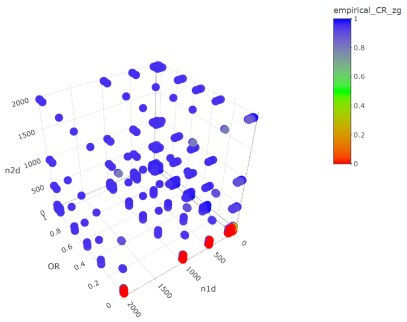

Figure 6: Gart Z OR Confidence Interval

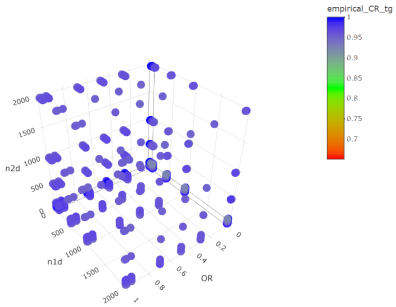

Figure 7: Gart T OR Confidence Interval

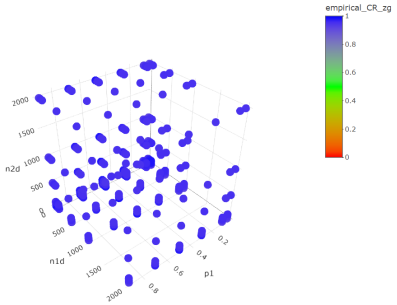

Figure 8: Gart Z p1 Confidence Interval

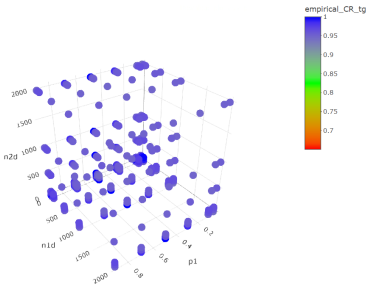

Figure 9: Gart T p1 Confidence Interval

Interpretations:

ZG_{OR} - In figure 6 we have the Gart Z OR confidence interval given odds ratio on the x-axis, n1d sample sizes on the y-axis and n2d sample sizes on the z-axis.The color of these spots represents the coverage rate of each combination of values with blue being very high and red very low. We obtain the best empirical coverage rate with smaller sizes of both samples n1d and n2d. We also notice a steady drop in our coverage rate when n1d and n2d are at their peak at 2000 with a coverage rate of 0.001. Alternatively, as the OR and sample size increases Our empirical coverage rate remains very small with smaller n2d even if n1d is high.

TG_{OR}- In figure 7 we have the Gart T OR confidence interval given odds ratio. We have the odds ratio on the x-axis,the n1d sample sizes on the y-axis and the n2d sample sizes on the z-axis. Here for high values of n1d and n2d

we obtain an acceptable odd ratio of 0.95 as seen with the purple color. Our highest coverage rate and Odd ratio is attained at n1d and n2d being exactly 1. As can be seen by the color, the overall performance of the odds ratio is better compared to the similar Z test with the odd ratio with the coverage rate around 0.95 in the worst case scenario making it a better alternative to the Z confidence interval.

ZG_{p1} and TG_{p1} - The last figures 8 and 9 show us respectively the Gart Z and T confidence interval given probabilities. We have the probabilities on the x-axis, the n1d sample sizes on the y-axis and the n2d sample sizes on the z-axis. For the confidence interval with probabilities we can see easily see how the Zg performs better compared to the Tg confidence interval. For both test, we obtain our highest probabilities of 0.9 with smaller samples sizes of 5 for both n1d and n2d. The T test performs as well as the Zg when n1d and n2d are very small but overall give lower coverage rate. This is easily noticeable with the purpleish overall color in the plot but the values still guve us an acceptable coverage rate in terms of accuracy.

3.2.3 Woolf Test

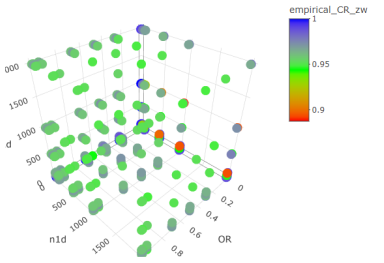

Figure 10: Woolf Z OR Confidence Interval

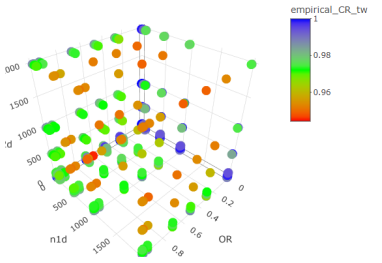

Figure 11: Woolf T OR Confidence Interval

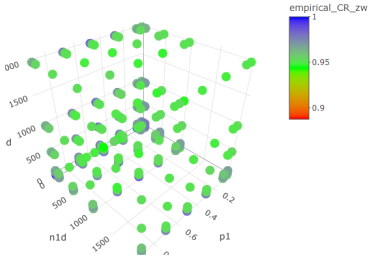

Figure 12: Woolf Z p1 Confidence Interval

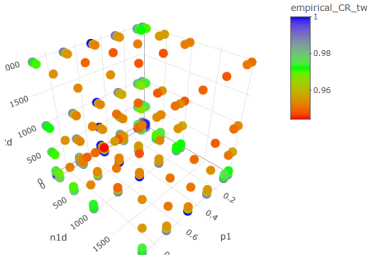

Figure 13: Woolf T p1 Confidence Interval

ZW_{OR} and ZW_{OR} - Figures 10 and 11, respectively, represent the Z and T Woolf Confidence Intervals for OR compared to n1d and n2d. The color of these spots represents the coverage rate of each combination of values. As seen in both graphs above, high coverage rate combinations commonly have low cell counts in at least one of the row totals (n1d or n2d). Some outliers occur when the OR is low, particularly with higher values of n1d and n2d for exapmle 2000 and 2000. Though these values are still slightly lower than those combinations where at least one of n1dn or n2d is small. However, outliers for theoretical probability differ, as a specific value of 0.5 appears to encourage higher coverage rates while still requiring at least one small value of n1d or n2d.

ZW_{p1} and TW_{p1} - Figures 12 and 13, respectively, represent the Z and T Woolf Confidence Intervals for p1 compared to n1d and n2d. These values follow trends similar to previous results, showing comparable patterns with OR and p1. However, the range of coverage rates in the heat map spans from 0.95 to 1, instead of 0.9 to 1 as seen with the Z-distribution. This indicates a higher average coverage rate for the T-distribution-based confidence interval compared to that of the Z-distribution.

3.3 Looking into the utilization of the Welch's adjustment

We are going to be looking at how the coverage rate changes between the tests, with and without the Welch's Adjustment. We will fix n2d and p1, then plot the coverage of the test vs. OR for multiple values of n1d. We have chosen to represent the data this way, because this can clearly show how the Welch's adjustment will account for extreme differences in the cell counts.

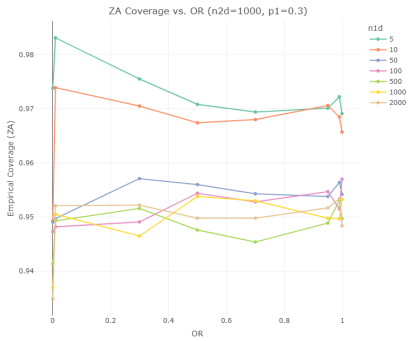

Figure 14: Agresti C.I

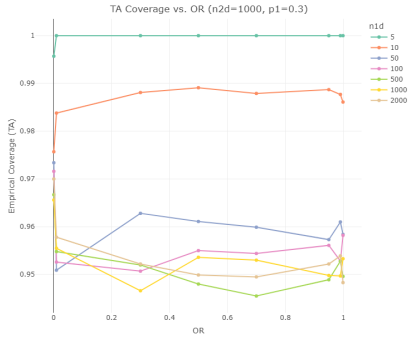

Figure 15: Welch's Adjustment of Agresti C.I

We clearly see with the Agresti's C.I that the Welch's Adjustment is effective for small sample sizes and levels out the graph. The effectiveness of Welch's adjustment compared to the original is due to the test statistic in the confidence interval. Where in the Welch's adjustement the test statistic takes into account cell counts through the degrees of freedom. While the original uses the Z test

statistic which only takes into account the alpha value. Explaining why as cell counts get smaller the coverage rates for Welch's adjustment increases. This also appears to be true for Gart and Woolf's coverage rates.

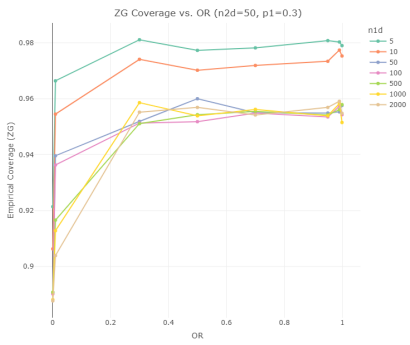

Figure 16: Gart C.I

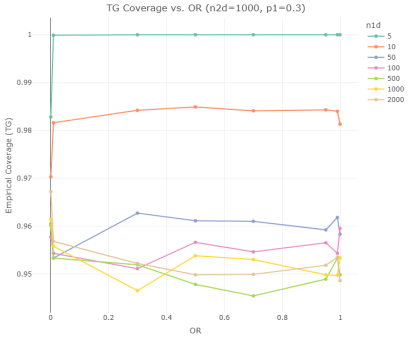

Figure 17: Welch's Adjustment of Gart C.I

Here we can see the same happens.

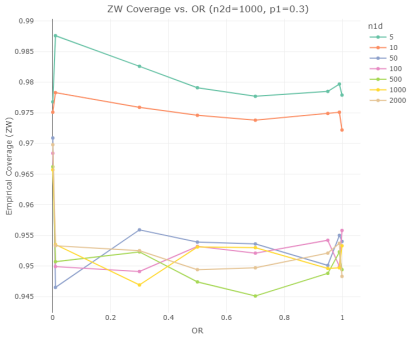

Figure 18: Woolf C.I Coverage

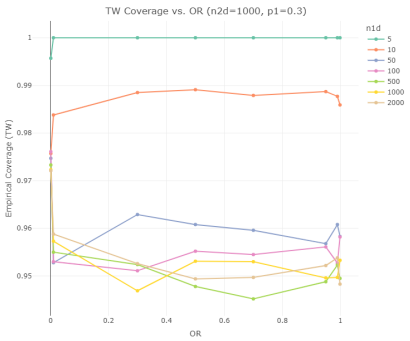

Figure 19: Welch's Adjustment of Woolf's C.I

3.4 Which Test is Best?

The ZG test shows strong potential for analyzing datasets and performs as a highly conservative method with excellent coverage in most scenarios. One of its key advantages is how easy it is to tell where the test works well and where it performs poorly. Now let us analyze this test more in depth for different scenarios. To more easily interpret how these different parameters affect the test, we will fix n2d and n1d, then plot the coverage of the test vs. p1 for multiple values of OR.

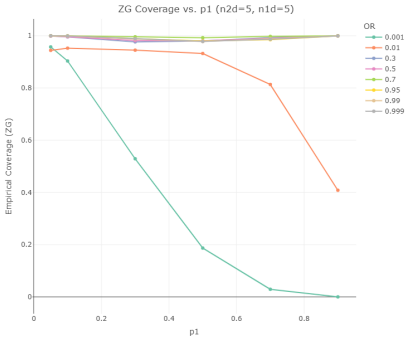

Figure 20: Coverage Rate of ZG, n1d = 5, n2d = 5

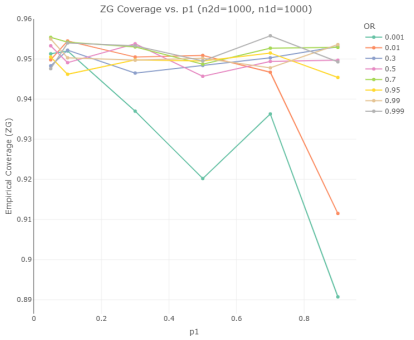

Figure 21: Coverage Rate of ZG, n1d = 1000, n2d = 1000

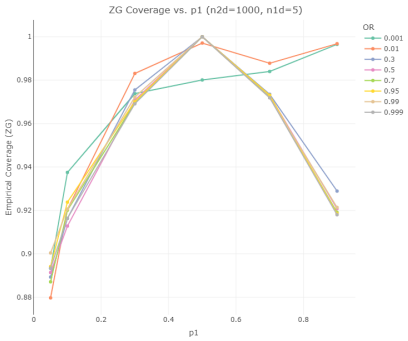

Figure 22: Coverage Rate of ZG, n1d = 5, n2d = 1000

First, let us highlight when this test should NOT be used. We can see a general trend that when the OR is extremely small, (i.e 0.001 or 0.01) as we increase p1, the lower coverage rate we get. Looking at figure 20, the test reaches a minimum when n1d and n2d are low, p1 = 0.9 and odds ratio = 0.001. If you were to encounter this situation, this is where you would want to use the Welch's Adjustment. Looking at the other graphs, it is clear that this test is not robust when there is a large difference in the OR and p1 values. However, outside of these extreme cases, we can see just how good of coverage rate this test has. Additionally, when the test does fail, they are perfect scenarios to use the Welch's Adjustment. It's reliability across most scenarios, combined with the clarity of when it fails, makes it a valuable method when used appropriately.

4 Additional Applications (Extra Credit)

4.1 Real Data Analysis

The dataset we will be analyzing is from a double-blind clinical trial investigating a new treatment for rheumatoid arthritis. Researchers wanted to see whether this new treatment led to better clinical outcomes compared to a placebo. Patients were randomly assigned to receive either the new treatment or a placebo. Each patient's improvement was categorized into three ordered levels: None, Some, and Marked. A total of 84 subjects, each with recorded demographic and response data. To create a simpler, binary outcome for our analysis, we combined the levels "Some" and "Marked" into a single category called "Better" while leaving "None" as is. This way, instead of three categories for the outcome, we have just two: "None" vs. "Better".

	Better	
Treatment	None	Better
Placebo	29	14
Treated	13	28

Table 2: 2x2 Contingency Table of Arthritis Data

	method	lower	upper
1	z_Agresti	1.741928	10.59110
2	t_Agresti	1.735691	10.62915
3	z_Gart	1.741882	10.59038
4	t_Gart	1.735650	10.62840
5	z_Woolf	1.784657	11.15359
6	t_Woolf	1.777648	11.19756

Table 3: 95% confidence intervals for the odds ratio using six different methods

All of the confidence intervals are roughly the same. Also, the interpretation of the confidence intervals is the same across all six intervals is the same, because none of the intervals include 1.0, i.e there is statistically significant evidence at $\alpha = 0.95$ that the odds of "None" (not improving) is higher for the Placebo group than for the Treated group. Therefore, we could conclude that there is enough evidence to suggest that the treatment is effective in improving arthritis.

When comparing the Welch's adjustment tests to their counterpart, we see that in all cases, the test with the Welch's Adjustment actually has a wider interval. The difference is very slight, only at the 0.01 level. This can be attributed to the relatively large cell counts for each cell. This can also explain how the Welch's Adjustment didn't have an effect on the coverage rate of each

test. That is, the degrees of freedom is large enough that the results will be similar among the Z-tests.

4.2 Adjusting the Simulation to 90%

We want to assess how adjusting the the confidence level effects the coverage rates of each test. Thus, we have re ran the simulation at $\alpha = 0.10$.

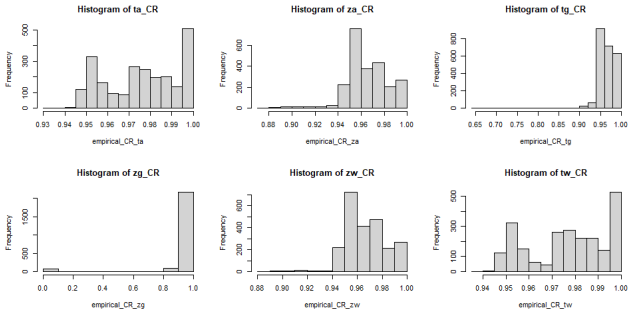

Figure 23: Histogram of Coverage rates at $\alpha = 0.10$

Looking at the histograms of coverage rates, there is hardly any difference vs. $\alpha = 0.05$. Therefore, we can conclude that the confidence intervals are still reliable at a type I error rate of 0.10.
To look at how much it did affect the tests, we will look into the Gart Z test.

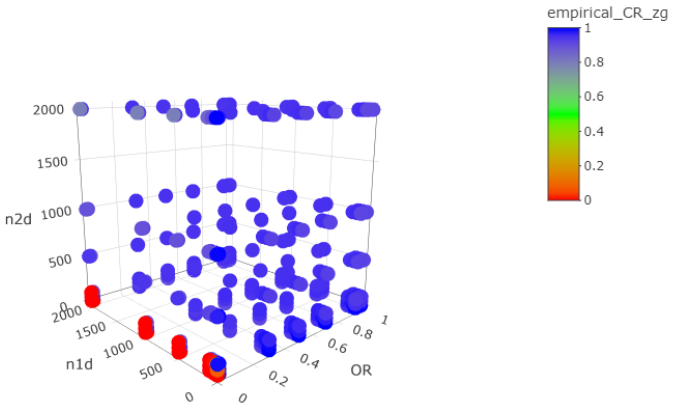

Figure 24: Gart-Z 3d Plot from $\alpha = 0.10$

From the plot we are still seeing the same results. The test is reliable in the

same places and performs poorly in the same places when $\alpha = 0.05$ highlighted in section 3.4. Overall, the adjustment in α did not seem to have a significant effect on the confidence intervals.

4.3 Comprehensive Literature Search

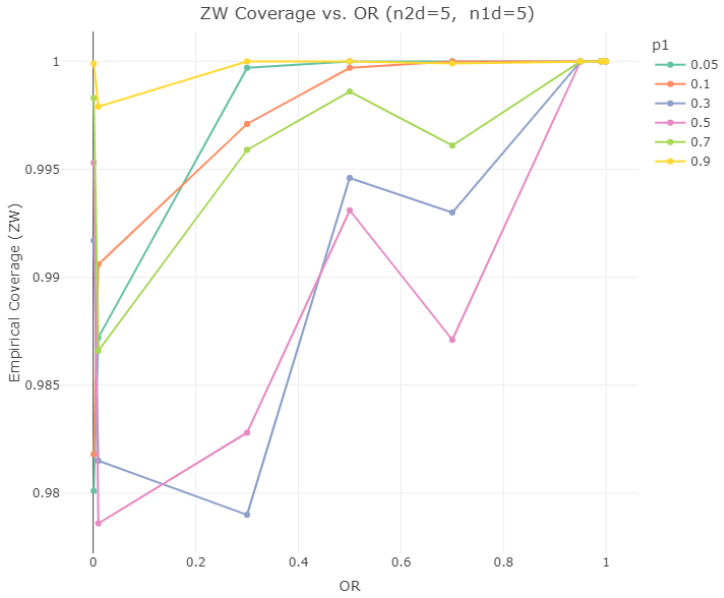

Figure 25: Coverage rate of Woolf Z Test

In the paper "Computing an Exact Confidence Interval for the Common Odds Ratio in Several 2×2 Contingency Tables," Cyrus R. Mehta, Nitin R. Patel, and Robert Gray introduce a computationally efficient algorithm for determining exact confidence intervals for the common odds ratio across multiple 2×2 contingency tables. They found that the Woolf Confidence Interval was more accurate when the true OR value was 1, especially with small sample sizes up to 64. When comparing this to our findings in Figure 25, the results are consistent with those reported in the previous article. Our graph shows that as the OR increases to 1, the coverage rate also increases, with the level of p1 reducing the volatility of the coverage rate.

5 Discussion

The use of the Welch's adjustment provides us with an advantage because it is designed to handle situations where the assumption of equal variances between groups is violated. Unlike the Z- test, which assumes equal variances, Welch's t-test does not require this assumption which makes it particularly useful when

comparing groups with different variances or sample sizes. In the case of our simulations study, to assess the reliability of the popular odds ratio (OR) this gives us an advantage because the homogeneity of variances assumption is almost impossible to observe in a real case scenario. Our analysis was consistent with the

When it comes to the simulation study, we used 3 main approaches: Agresti, Gart and Woolf's and confidence intervals and ran each of them with the adjusted Welch. Our preliminary suggested that the ZG test was particularly good in terms of computing coverage rate but had a high failure rate.

The Agresti test performs best when sample sizes are small, as it improves empirical coverage rates. However, its performance diminishes when the OR is extremely small (e.g., 0.001) and sample sizes exceed 500 Despite its good performance in certain scenarios this approach struggles with consistency. The Wolf confidence interval only showed reliable performances when the odds ratio was close to 1. Out of all the different test used the Gart test performed the best. The Z Odds Ratio confidence interval achieves its best empirical coverage rate when both n1d and n2d are small, but its performance declines sharply particularly with extreme values like 2000. Compared to the Z, The Gart T confidence interval demonstrates better overall performance, maintaining acceptable coverage rates (around 0.95) even in more challenging scenarios, which makes it the best of all the other alternatives.

One aspect of this study that was particularly difficult was dealing with the simulations which are computationally very demanding which limited our simulations to 10000 per combination. With better research settting, we could have significantlky increase the number of simulations and increase our chances of having more accurate results.

6 Citations

Agresti, Alan. “On logit confidence intervals for the odds ratio with small samples.” Biometrics, vol. 55, no. 2, June 1999, pp. 597–602, https://doi.org/10.1111/j.0006-341x.1999.00597.x.

Fagerland, Morten W, et al. “Recommended confidence intervals for two independent binomial proportions.” Statistical Methods in Medical Research, vol. 24, no. 2, 13 Oct. 2011, pp. 224–254, https://doi.org/10.1177/0962280211415469.

Gart, John J. “Alternative analyses of contingency tables.” Journal of the Royal Statistical Society Series B: Statistical Methodology, vol. 28, no. 1, 1 Jan. 1966, pp. 164–179, https://doi.org/10.1111/j.2517-6161.1966.tb00630.x.

Lawson, Raef. “Small sample confidence intervals for the odds ratio.” Communications in Statistics - Simulation and Computation, vol. 33, no. 4, Oct. 2004, pp. 1095–1113, https://doi.org/10.1081/sac-200040691.

Mehta, Cyrus R., et al. “Computing an exact confidence interval for the common odds ratio in several 2 × 2 contingency tables.” Journal of the American Statistical Association, vol. 80, no. 392, Dec. 1985, p. 969, https://doi.org/10.2307/2288562.

Subbiah, M., and M.R. Srinivasan. “Classification of 2×2 sparse data sets with zero cells.” Statistics & Probability Letters, vol. 78, no. 18, Dec. 2008, pp. 3212–3215, https://doi.org/10.1016/j.spl.2008.06.023.
Welch, B. L. “The generalization of ‘student’s’ problem when several different population variances are involved.” Biometrika, vol. 34, no. 1/2, Jan. 1947, p. 28, https://doi.org/10.2307/2332510.

Woolf, Barnet. “On estimating the relation between blood group and disease.” Annals of Human Genetics, vol. 19, no. 4, May 1955, pp. 251–253, https://doi.org/10.1111/j.1469-1809.1955.tb01348.x.