ST 502 Project 1

Eric Warren, Chandler Ellsworth, Kevin Krupa

2024-02-26

1 Goals of Project

In this report, it is our goal to compare the performance of confidence intervals of binomial proportions by making inference for various values of sample size (n) and success probability (p) based on a Monte Carlo simulation study. Six different methods will be used, and performance will be assessed by proportion of intervals that contain the true value of p, proportion of intervals that miss above or below p, and the average length of the interval. We hope to show through our analysis that approximate intervals are better than exact intervals and select the most favorable interval methods.

2 Methods of Project

To conduct inference on p, six different confidence interval methods were used: Wald, Adjusted Wald, Score, Clopper-Pearson, Raw Percentile Parametric Bootstrap (Raw), and Bootstrap t Parametric Bootstrap (Boot t). The Wald, Adjusted Wald, and Score intervals are based on asymptotic normality assumptions, the differences being that Wald uses the sample proportion, Adjusted Wald adds two successes and two failures to the sample proportion, and Score attempts to correct for small sample sizes and extreme proportions. Clopper-Pearson is an exact confidence interval that, in theory, should give desired confidence levels for any p. Lastly, Raw and Boot t approximate confidence intervals based on empirical quantiles from bootstrap resampling distributions, where Raw approximates the sample proportion's distribution, and Boot t mimics a "t-type" statistic.

3 Creation of Data

The data was created by first identifying the ninety different combinations of p and n. A loop was then used to draw 1500 independent samples (y_i) using the `rbinom()` function for each combination, with the results stored in a data frame with three columns: n, p, and y_i. Each data frame within the loop was combined with the previous, resulting in a consolidated data frame of the 1500 independent samples for each combination of n and p. Lastly, $\hat{p_i} = \frac{Y_i}{n}$, the sample proportion, was then calculated for each sample. Below is a sample of the created data.

n	p	y_i	p_hat	n	p	y_i	p_hat
15	0.01	0	0	200	0.99	199	0.995
15	0.01	0	0	200	0.99	198	0.990
15	0.01	0	0	200	0.99	197	0.985
15	0.01	0	0	200	0.99	196	0.980
15	0.01	0	0	200	0.99	196	0.980
15	0.01	0	0	200	0.99	199	0.995

Note the data is in *long format* which means each combination of n, p, y_i, $\hat{p}_i$ are shown as one row. For example, the combination of $n = 15$ and $p = 0.01$ would need to be referenced by the 1500 rows that show the corresponding y_i / $\hat{p}_i$ values.

4 Calculating Quantities

Using the created data, 95% confidence intervals (setting $\alpha = 0.05$) were calculated from each of the independent samples for each method using the following interval formulas:

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

Adjusted Wald: $\hat{p} = \frac{y_i+2}{n+4}$ using the same interval formula as the Wald Interval

Score: $\frac{\hat{p}+\frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})+z^2_{\alpha/2}}{\frac{4n}{n}}}}{1+\frac{z^2_{\alpha/2}}{n}}$

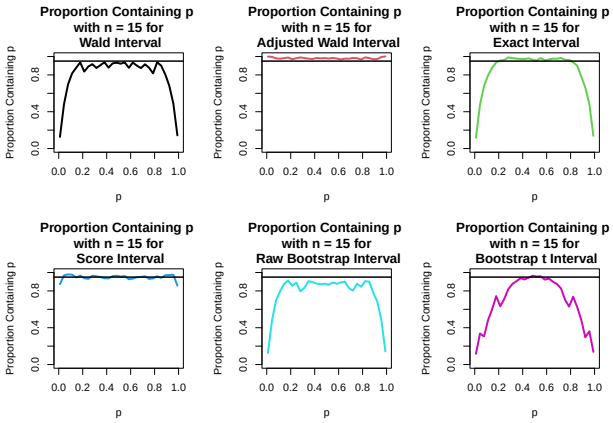
Clopper-Pearson (Exact): $((1+\frac{n-y_i+1}{y_i F_{2y_i,2(n-y_i+1),1-\frac{\alpha}{2}}})^{-1}, (1+\frac{n-y_i}{(y_i+1)F_{2(y_i+1),2(n-y_i),\frac{\alpha}{2}}})^{-1})$ where $F_{a,b,c}$ denotes the $1-c$ quantile from the F-distribution with degrees of freedom a and b

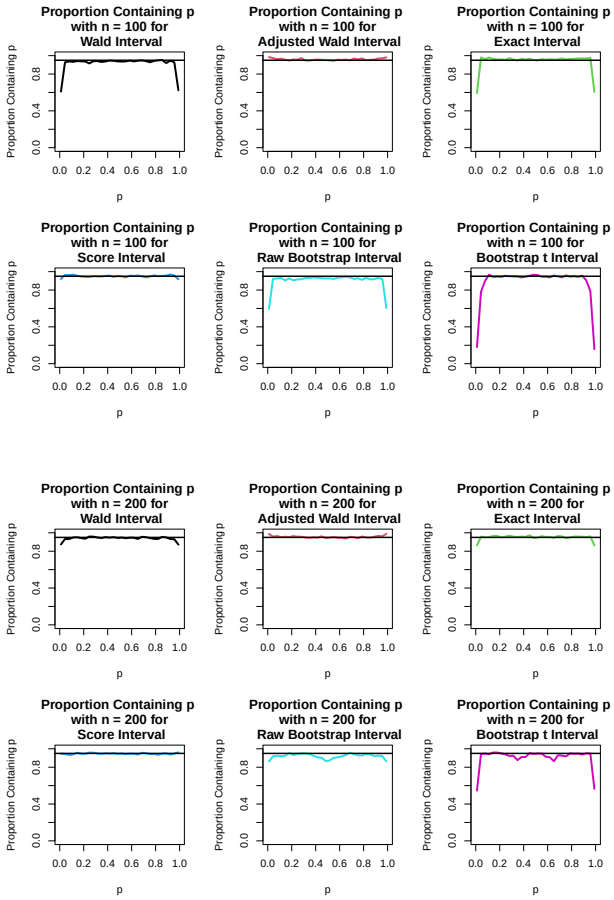
Raw: $(\theta^*_{lower}, \theta^*_{upper})$ where θ^*_{lower} and θ^*_{upper} are the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles of the bootstrap distribution of $\hat{p}$

Boot t: $(\hat{p} - \delta_{upper} * \hat{SE}(\hat{p}), \hat{p} - \delta_{lower} * \hat{SE}(\hat{p}))$ where δ_{lower} and δ_{upper} are the quantiles from the bootstrap distribution of $T = \frac{\hat{p}-p}{\hat{SE}(\hat{p})}$

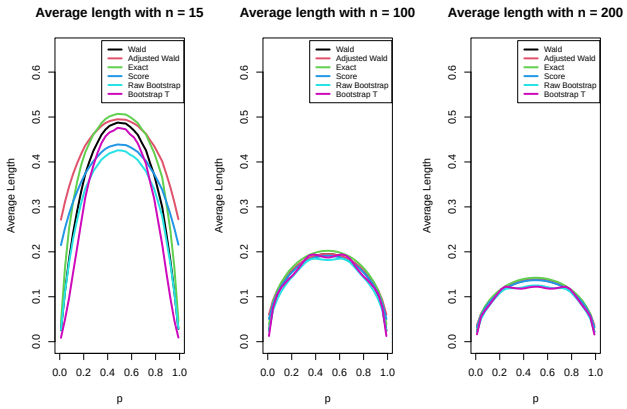
Each confidence interval across the different methods was evaluated to see if the true p was contained in the interval, or if it missed above or below, by classifying a 1 or 0 in three new indicator columns, "count," "count_under," and "count_over." This enabled proportions to be calculated for each of the new columns by taking the average of the count from the 1500 independent samples for each n and p combination. A similar procedure was done to calculate the average interval length, but instead of taking an average of the count, the average difference was taken between the upper and lower bound of the confidence interval.

5 Results

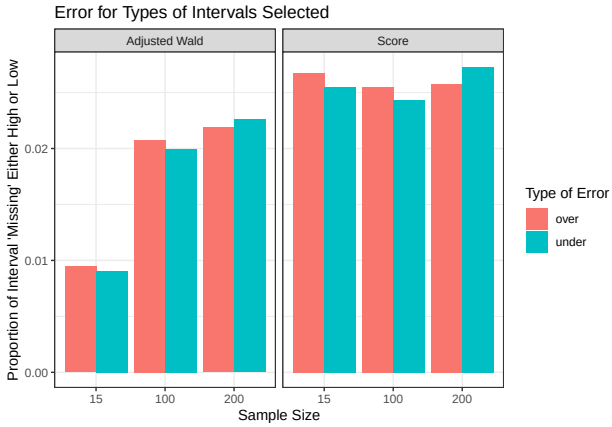




From the calculations and plots, there were clear trends of increased performance as the sample increased across all methods, but certain methods performed better than others for a set sample size. First, for a small sample size of $n = 15$, the Score interval method performed the best in relation to the proportion of confidence intervals that captured the true p value relative to our desired proportion of 0.95. The Adjusted Wald interval also had stable proportions across the support of p, with better results for extreme values close to 0 or 1, but generally had proportions greater than desired. The Exact interval did well for p between 0.3 to 0.7, but had significantly lower proportions than desired for other p values. The other interval methods did a poor job of creating confidence intervals that contained the true p based on our desired level of confidence. For a medium sample size of $n = 100$, the Adjusted Wald and Score interval methods seemed to perform the best in relation to proportion contained relative to desired, with the other four methods closely behind aside from p values close to 0 or 1. Lastly, for a large sample size of $n = 200$, the Adjusted Wald and Score intervals again performed well. The Wald and Exact methods had comparable results to the Adjusted Wald, except had lower proportions than desired for p values close to 0 to 1, while Adjusted Wald had more conservative results with proportions greater than desired for the same p range. For larger sample sizes, it became more apparent the Raw method outperformed the Boot t method for p values closer to 0 and 1 in relation to having proportions closer to desired of 0.95.



For small sample sizes of $n = 15$, the Bootstrap methods and the Score method limited our interval ranges for the same level of confidence, which is preferred. For medium sample sizes of $n = 100$, all methods were similar, while for even larger sample sizes of $n = 200$, the Bootstrap methods had slightly lower average interval lengths for p in the range of 0.3 to 0.7. Factoring the average interval length results into the results from how well the confidence intervals captured the true p, the Score method appeared even more favorable for smaller sample sizes.



Since the Adjusted Wald and Score intervals appeared to perform the best in relation to proportion of intervals containing the true p, while also having comparable average interval lengths relative to the other methods, we lastly checked the proportion of intervals that either overestimated or underestimated the true p for these two methods. Across all sample sizes for the two methods, the overestimated and underestimated proportions were about the same, meaning the interval estimation procedures were empirically unbiased and were reasonable for estimating intervals that contained the true p. The results also showed the general preference towards the Score method since, if the proportion of intervals that contained the true p were close to the desired proportion of 0.95, that would leave 0.05 to be separated between overestimated or underestimated. The preference for the remaining 0.5 would be to have it equally split amongst overestimated and underestimated at 0.025, which the Score interval more closely aligned to over the Adjusted Wald, especially for smaller sample sizes. This also showed the Adjusted Wald had slightly more conservative estimations, since the remaining proportion that did not contain was less than 0.05 for all sample sizes, likely indicated by a greater average interval length.

6 Conclusions

After reviewing the results, our preferred methods for creating confidence intervals of binomial proportions are the Adjusted Wald and Score Methods. These are our preferred methods since they had the best coverage and adequate interval lengths for our desired confidence level for all sample sizes, across the support of p. Further, these "approximate" methods performed better than the exact interval method, Clopper-Pearson, and had a much friendlier interpretation and calculation. Adjusted Wald and Score are especially preferred for lower samples sizes, as the other four methods performed poorly, especially when p was close to 0 or 1. We slightly favor the Score method over the Adjusted Wald for small sample sizes as the Score method performed better for values of p close to 0 and 1 and had a more narrow average interval length, meaning a more precise interval of p could be given with the same level of confidence. For large sample sizes, we are unbiased towards using the Adjusted Wald or Score method, but if the true p was assumed closer to 0 or 1, one may prefer to use the Adjusted Wald over Score since the Adjusted Wald had more conservative coverage for these values with a similar average interval length.