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 n /p combination. 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

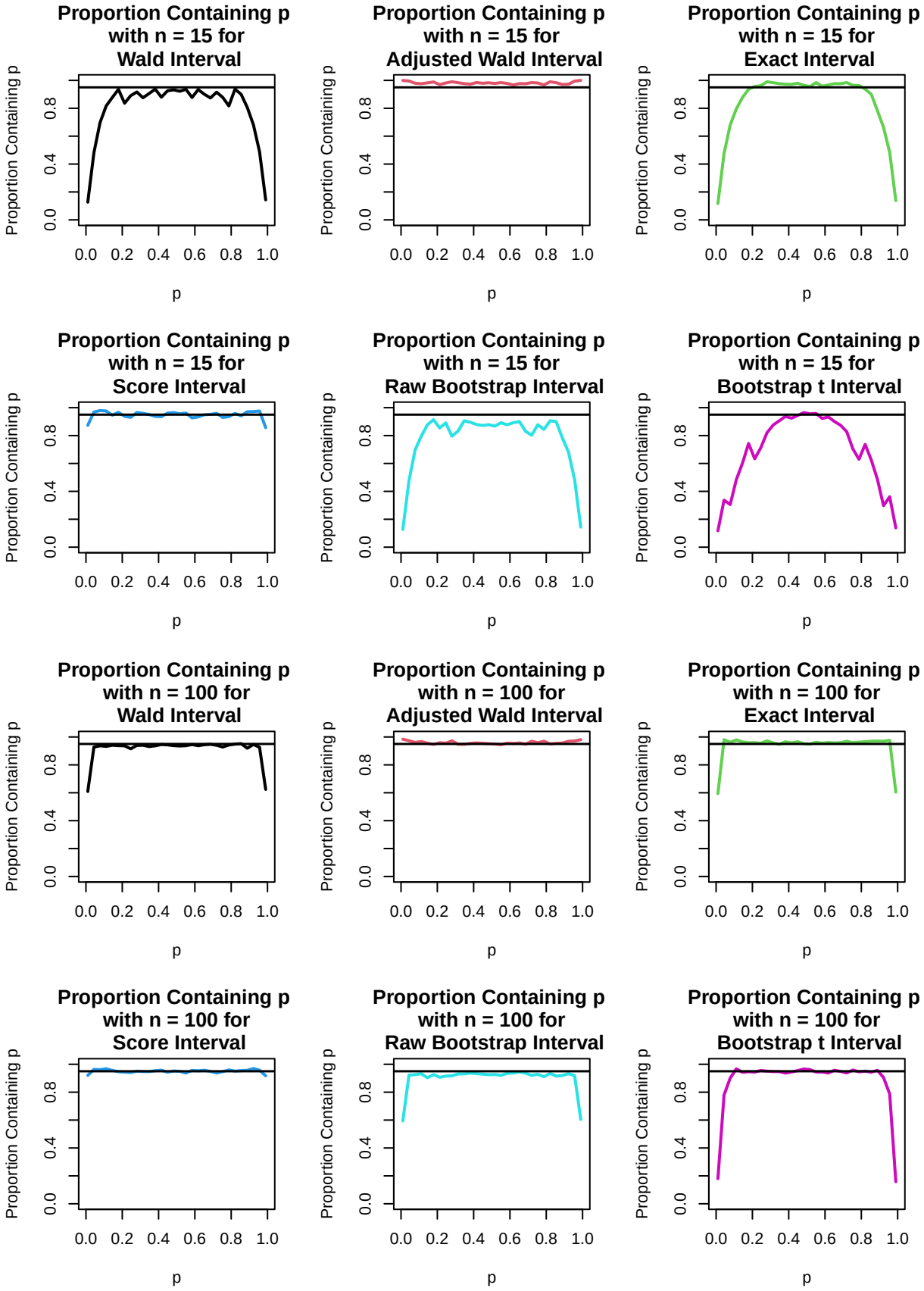
Note the data is in *long format* which means each combination of n, p, y_i, $\hat{p}_i$ are shown as one row. For example, the combination of $n = 15$ and $p = 0.01$ would need to be referenced by the 1500 rows that show the corresponding y_i / $\hat{p}_i$ values.

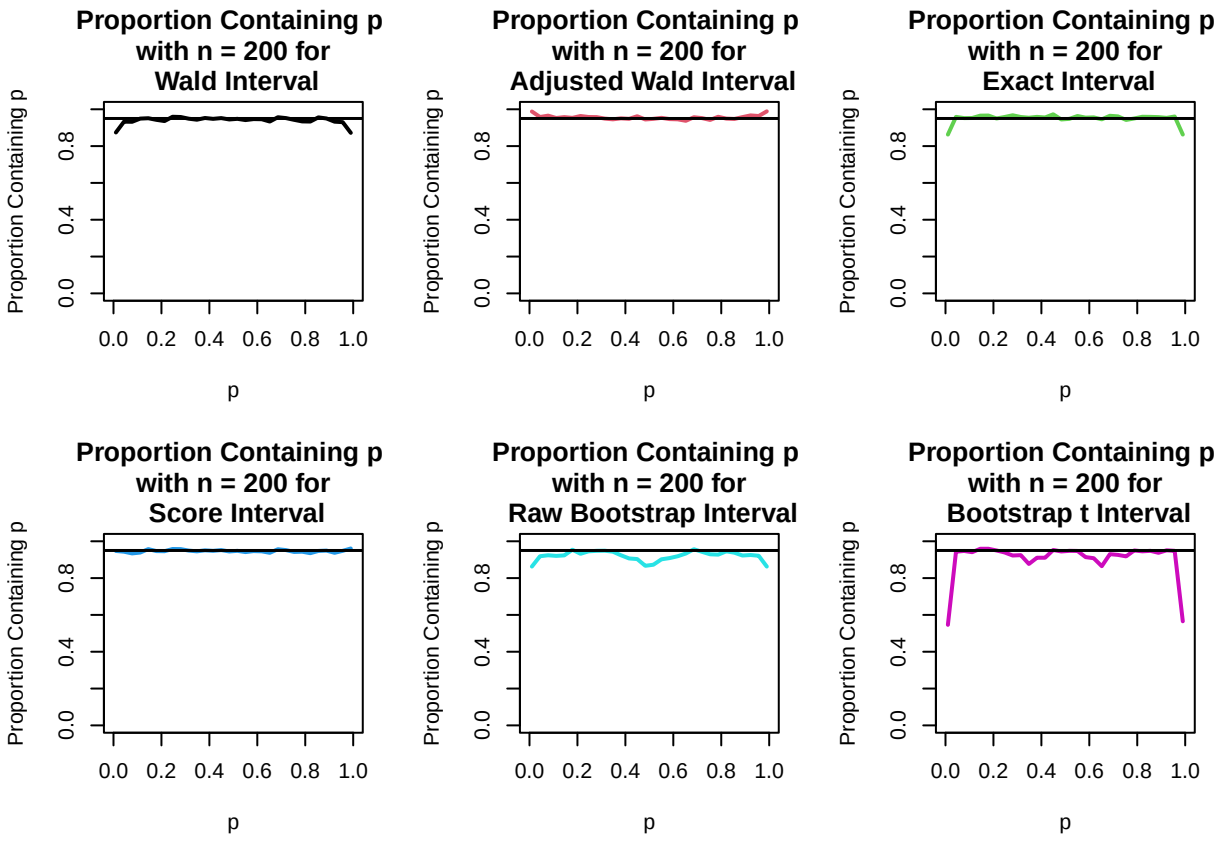
4 Calculating Quantities

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 $T = \frac{\hat{p}-p}{\hat{SE}(\hat{p})}$

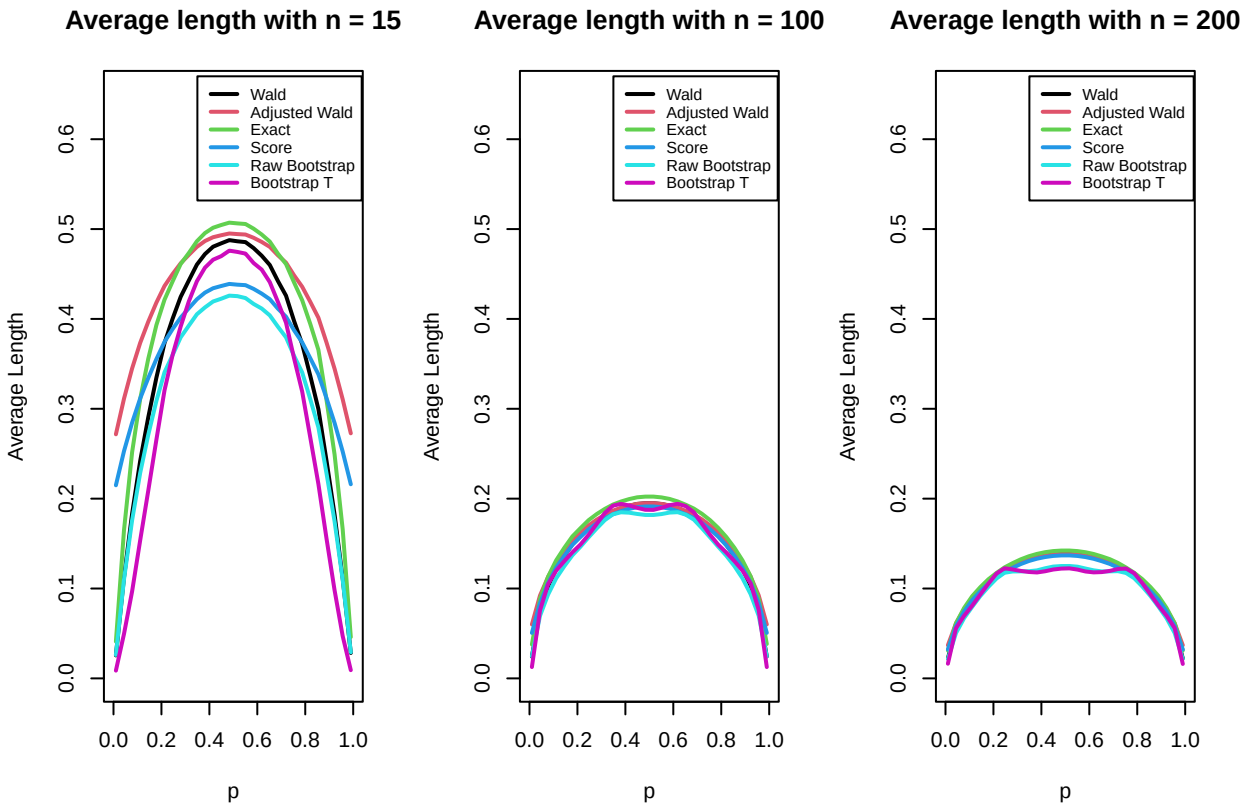
Each confidence interval across the different methods was evaluated to see if the true p was contained in the interval, or if it missed above or below, by classifying a 1 or 0 in three new indicator columns, "count," "count_under," and "count_over." This enabled proportions to be calculated for each 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

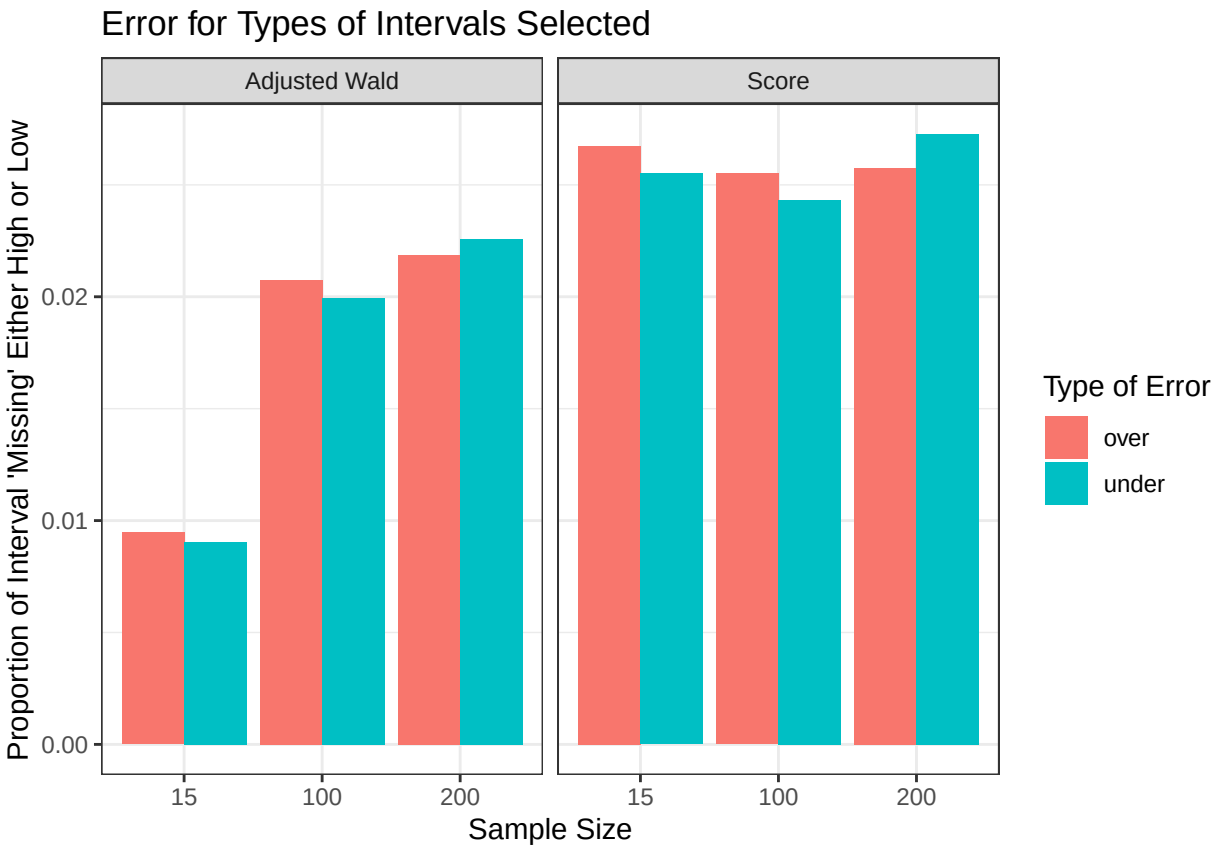

Proportion Containing p with n = 15 for Wald Interval
Proportion Containing p with n = 15 for Adjusted Wald Interval
Proportion Containing p with n = 15 for Exact Interval
Proportion Containing p with n = 15 for Score Interval
Proportion Containing p with n = 15 for Raw Bootstrap Interval
Proportion Containing p with n = 15 for Bootstrap t Interval
Proportion Containing p with n = 100 for Wald Interval
Proportion Containing p with n = 100 for Adjusted Wald Interval
Proportion Containing p with n = 100 for Exact Interval
Proportion Containing p with n = 100 for Score Interval
Proportion Containing p with n = 100 for Raw Bootstrap Interval
Proportion Containing p with n = 100 for Bootstrap t Interval
Proportion Containing p
p
0.0
0.4
0.8
0.0 0.2 0.4 0.6 0.8 1.0



When evaluating the execution level for the different intervals for different sample sizes there are some trends that are present. First, we can see for our lower sample size of $n = 15$, the Adjusted Wald seems to be performing the best of capturing our necessary proportion of confidence intervals that should capture our true proportion parameter p. The Score interval also seems to do a good job as well (minus the extremes) and the Exact interval also seems to do fairly well (with the "extreme" range is a little wider than the Score interval). The other interval methods do a poor job of correctly classifying our desired proportion of correct confidence intervals. Then, we can see for our medium sample size of $n = 100$, the Adjusted Wald again seems to be performing the best of capturing our necessary proportion of confidence intervals that should capture our true proportion parameter p with the Score interval closely behind (minus the extremes being slightly below par). The other four intervals seem to do fairly well (with the "extreme" range is a little wider than the Score interval) but from these plots alone it is hard to say that one of those four are much "better" than the others. Lastly, we can see for our large sample size of $n = 200$, the Adjusted Wald and Score intervals both seem to be performing the better than its "competitor" interval methods of capturing our necessary proportion of confidence intervals that should capture our true proportion parameter p with the Score interval closely behind. The Exact and Wald intervals are close behind with just the extreme p (true proportion values) having some slight faults. The other two bootstrap intervals seem to do fairly alright (with the "extreme" range is a little wider than the other intervals) and seem to be a little more inconsistent as getting to our ideal confidence level (which we set at 95%).



For small samples (in our case $n = 15$), the bootstrap methods and the Score interval both seem to limit our confidence ranges, which we prefer intervals with smaller length (or less error in the bounds). For our medium size sample of $n = 100$, they are all pretty similar and for our large sample of $n = 200$, we can see very similar lengths with the bootstrap methods actually limiting our error slightly better than the other four methods. Now comparing this to how well our confidence intervals captured our "true" value of p, we might lean on using the Score Interval for smaller sample sizes, while an Adjusted Wald Interval seems adequate for a larger sample.



Lastly, we want to check the Adjusted Wald and Score Intervals are both roughly missing the same proportion of intervals by overestimating compared to underestimating. Note that we want this to be even to show that the confidence intervals we are making are not biased. As we can see for our different sample sizes the over and under estimation values are about the same, meaning we are using unbiased interval estimators and further signaling that these are solid intervals to use for estimating the true proportion population value of p.

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 sample 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 can be given with the same level of confidence. For large sample sizes, we recommend using the Adjusted-Wald interval over the Score interval, as the Adjusted-Wald interval was more conservative for values of p close to 0 or 1 by having coverage higher than our desired level of confidence, while having similar average interval lengths to the other methods examined.