ST 502 Project 1

Eric Warren, Chandler Ellsworth, Kevin Krupa

2024-02-21

Contents

1 Goals of Project

In this report, it is our goal to review the performance of confidence intervals through six different methods. These methods will be further explained below; however, we will be assessing their performance for different combinations of success probabilities and sample sizes. In each combination, we will assess the following: proportion of intervals that contain the true value of the parameter, proportion of intervals that are below/above the true value, as well as the average length of the intervals. After completing the goals above, it is our intention to select the "best" interval method. This will be determined by an amalgamation of the listed goals, not the best method for each goal. We will further determine if there are methods, other than our chosen "best", that can also be considered useful.

2 Methods of Project

We are exploring six different types of interval estimation techniques that we can use to create our own 95% confidence intervals based on simulation of data. The intervals we are evaluating are the Wald Interval ($\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$), Adjusted Wald Interval (says $\hat{p} = \frac{y_i+2}{n+4}$ and using the same interval formula as the **Wald Interval**), Clopper-Pearson (Exact) Interval (lower bound is $(1 + \frac{n-y_i+1}{y_i F_{2y_i,2(n-y_i+1),1-\frac{\alpha}{2}}})^{-1}$ and upper bound is

$(1+\frac{n-y_i}{(y_i+1)F_{2(y_i+1),2(n-y_i),\frac{\alpha}{2}}})^{-1}$ where $F_{a,b,c}$ denotes the $1-c$ quantile from the F-distribution with degrees of freedom a and b), Score Interval $(\frac{\hat{p}+\frac{z^2_{\alpha/2}}{2n}\pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})+\frac{z^2_{\alpha/2}}{4n}}{n}}}{1+\frac{z^2_{\alpha/2}}{n}})$, using a Raw percentile interval using a parametric bootstrap, and using a Bootstrap t interval using a parametric bootstrap. As what will be described more in depth in the Creation of Data section of our report, we will have 1500 independent samples for each combination of p (our true proportion from our population) and n (the sample size we take to "estimate" our proportion, pretending we do not know what the true value is). We will then explore the six different interval techniques and compare the proportion of our simulated data intervals that capture our "true" proportion (p).

3 Creation of Data

To create the data needed to complete the goals of this report, we will first start by correctly identifying the sequences of success probabilities (the true value of p) and sample sizes (n). After successfully establishing these sequences, we can then establish a loop that will parse through all combinations of the two sequences. Furthermore, for each combination, we will establish a data frame with three total variables (sample size, success probability, y_i). The y_i column will be established using an R function (`rbinom()`) that randomly samples from a binomial distribution that takes the sample size and success probability variables as parameters. After successfully creating the data for that specific parameter combination, we will then do the same for all combinations and row bind all into one single data set that contains all observations from each combination of sample size and success probability. Lastly, to help with calculations later, we will create a column for the $\hat{p}_i$ estimates which can be calculated from the formula $\hat{p}_i = \frac{Y_i}{n}$. After generating this data, we will show briefly what it looks like to get a better understanding of our structure.

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

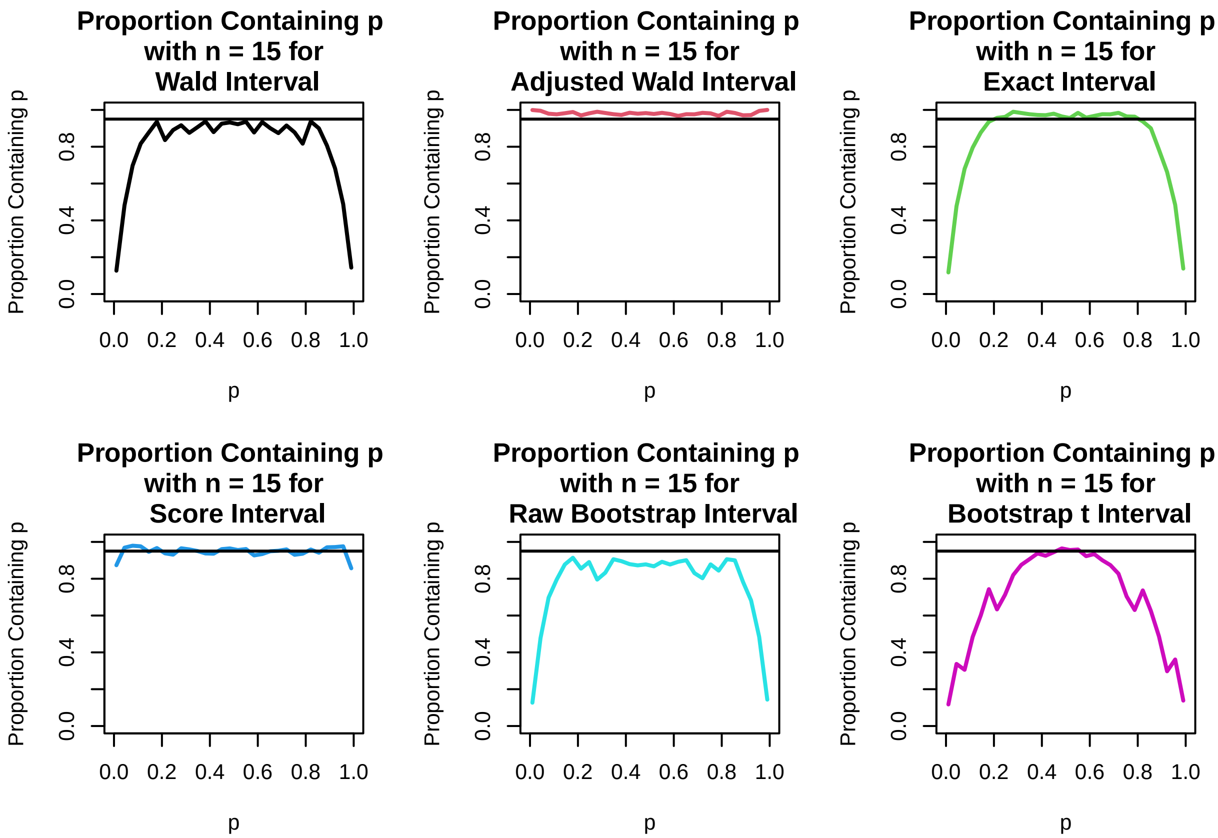
Please note for this data that it is in *long format* which means that each combination of n, p, y_i, $\hat{p}_i$ is shown as one row. So for example, if we want to find all the randomly generated data for the combination of $n = 15$ and $p = 0.01$, we would want to find the 1500 rows that show its corresponding y_i / $\hat{p}_i$ values.

4 Calculating Qunatities

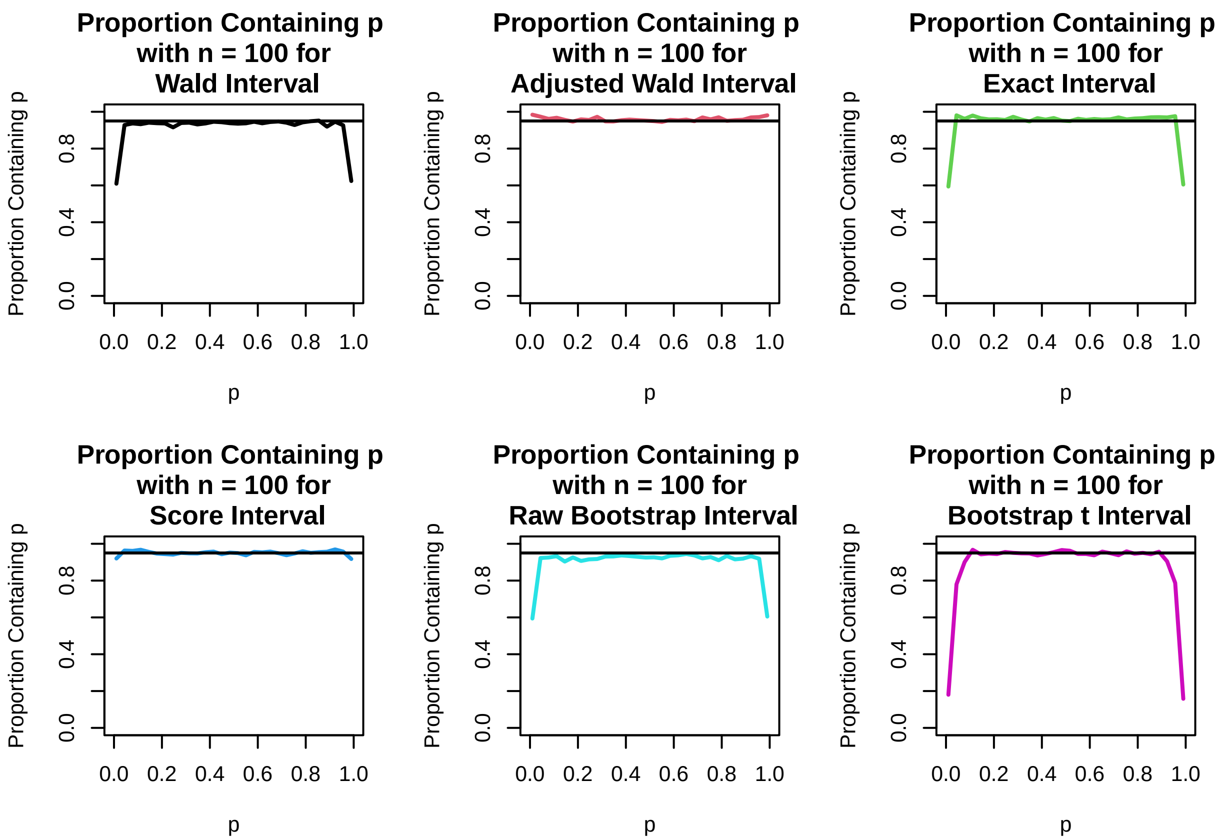
After creating our data, we are now going to use each interval method (mentioned in the *Methods* section) and use their respective 95% confidence interval formulas; this is setting $\alpha = 0.05$ when calculating the intervals for our $\hat{p}$ we generated for each independent sample. Afterwards, we will check to see if the interval created contains the true p proportion parameter in it. If it does, which means the lower bound of the interval created from our $\hat{p}$ statistic is less than the true p proportion parameter and the upper bound of the interval created from our $\hat{p}$ statistic is greater than the true p proportion parameter then we would create a new indicator variable called `count` which would be 1 if this statement is true and 0 if false (meaning our calculated lower bound was greater than or equal to p **or** our upper bound was less than or equal to p. Finally to get our proportion of intervals for each combination of n (our sample size) and p (our actual true proportion parameter) we would find the percentage of intervals that contain our true value of p by finding the average of the `count` indicator variable. These quantities calculated from our simulation will be helpful in making recommendations of which intervals are the best to use.

5 Results

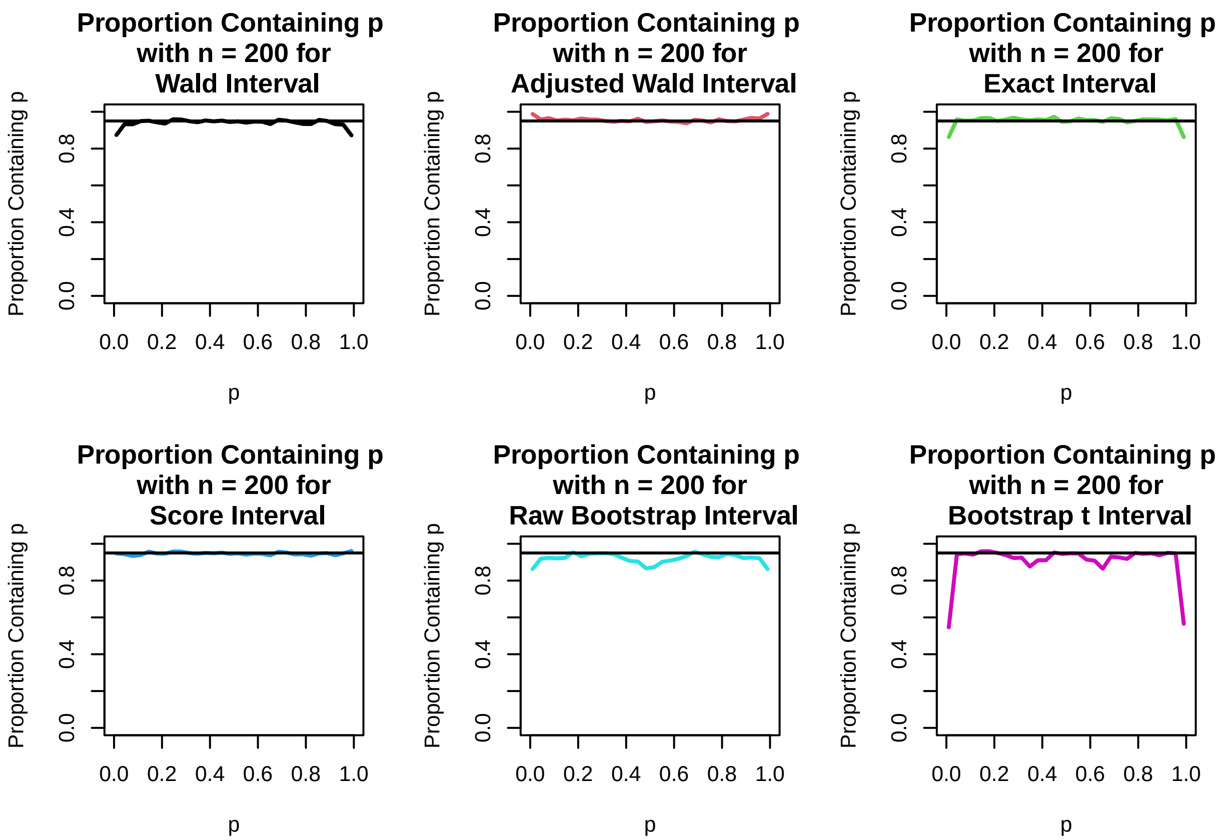
5.1 Proportion of Intervals Capturing Parameter for Each Sample Size



As 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.

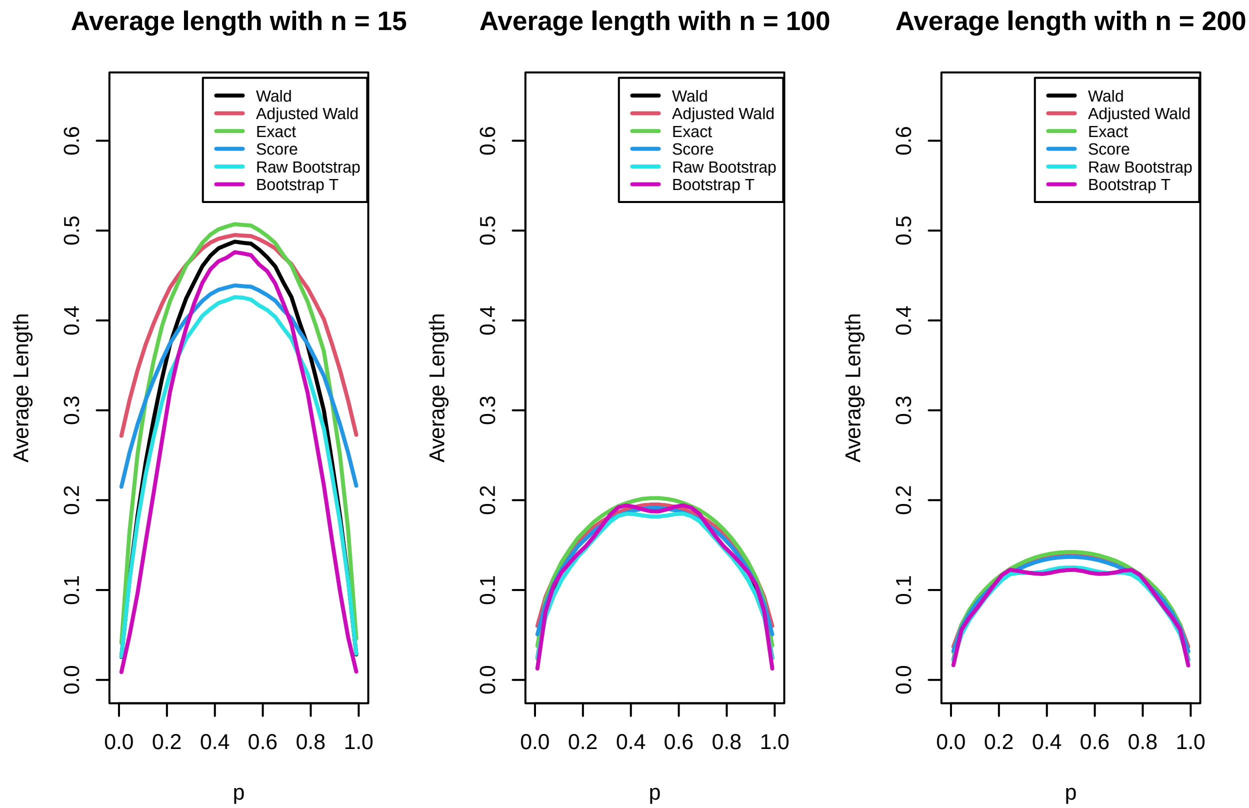


As we can see for our medium sample size of $n = 100$, the Adjusted Wald again seems to be performing the best of capturing our necessary proportion of confidence intervals that should capture our true proportion parameter p with the Score interval closely behind (minus the extremes being slightly below 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.



As 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%).

5.2 Average Confidence Interval Length Results for Each Sample Size



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.

5.3 Key Takeaways

Based on our computed graphs, there are several results that we can identify. When $n = 15$, the Adjusted-Wald and Score methods appear to have the best overall coverage in terms of achieving our desired confidence level. We can make the argument that the Adjusted-Wald method is better, because it achieves the desired level at the extreme values of p. With that being said, the average lengths for the Score method do appear to be smaller than the Adjusted-Wald interval lengths. This could explain why the Score method does not achieve the desired confidence level at the extreme levels of p. Moving to when $n = 100$, we can still consider the Adjusted-Wald and Score Methods to be the best in terms of overall coverage. The same argument for the Adjusted-Wald method can be made when discussing coverage at extreme levels of p, but this time the average length distributions for each method appear to be very similar. Finally, when $n = 200$, we can still see that Adjusted-Wald and Score methods have the best coverage; however, they are now very similar at the extremes, so a “best” distinction would be hard to make. Also, the distributions of average interval lengths once again are approximately the same for all six methods, so no distinction can be made here either.

6 Conclusions

After reviewing the results, we have come to the following conclusions for the "best" interval method. In terms of overall results, the best performing methods are the Adjusted-Wald and Score Methods. Regarding smaller sample sizes, the Score method appears to perform the best overall, as it does well in reaching our ideal confidence level and limits the length of our error bounds. Now, with that being said, the Adjusted-Wald interval obtained the desired confidence level, even at the extremes where the Score method will miss just slightly. We can see for lower sample sizes when we are looking at the lower true proportion p the other four methods tend to perform horribly when capturing the parameter in our confidence levels. On top of that, the Wald Interval tends to perform poorly when evaluating low sample sizes, but will perform much better as we increase our sample size. However, the average interval length for the Score method is narrower than the Adjusted-Wald, meaning we will get a more precise interval. Therefore, for smaller sample sizes, we recommend using the Score method when producing a Confidence Interval, because we only miss our confidence level slightly for very extreme values of p. This decision was referenced in the **Results** section. Now as the sample size increases (such as $n = 100$ or $n = 200$) we will recommend using the Adjusted-Wald interval. The Adjusted-Wald interval will always achieve our desired confidence level, whereas the Score Interval will hover at or just below it. Also, the average interval lengths for the six methods are approximately the same for all values of p, so no distinction on "best" method can truly be made, based on strictly looking at interval length. Therefore, we will put much more weight towards the fact that the Adjusted-Wald method will achieves our desired confidence level, even in the event of an extreme p (where p is close to 0 or 1). In conclusion, when producing confidence intervals, we will recommend using the Score method for smaller sample sizes and the Adjusted-Wald method for larger sample sizes.