Introduction/Goal

Confidence Intervals are used to describe the variation around a statistic and predict the value of your estimate. There are various methods used to calculate confidence intervals. Our goal in this project is to compare six different confidence interval methods for capturing the binomial parameter p (the probability of success) and report our findings.

In our study we performed a Monte Carlo simulation study in R to investigate the properties of the confidence intervals. We looked at the following:

- observed coverage rate; that is the proportion of confidence intervals that capture the true value p
- proportion of intervals that miss above and below
- average length of the confidence interval
- observed standard error (SE) in the average length of the interval

We compared six methods: Wald, Adjusted Wald, Clopper-Pearson, Score, Raw Percentile Parametric Bootstrap, and Bootstrap t interval (also parametric). These methods are detailed in the next section.

Confidence Intervals Methods to Compare

The first confidence interval that we explored was the Wald confidence interval. If we let X denote our binomial, and n be our sample size then $\hat{p} = X/n$ is our sample proportion. The 100(1- α)% Wald Confidence Interval is then

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

Where $z_{\alpha/2}$ is the 1- $\alpha/2$ quantile of the standard normal distribution. (Agresti, 1998) This has the advantage of being the simplest method, but according to sources referenced in Agresti & Coull this interval performs quite poorly unless n is large.

The Adjusted Wald interval uses a quick modification: add two successes and two failures. This simple adjustment changes the interval from too liberal to slightly conservative. (Agresti, 1998) The formula for the Adjusted Wald interval is therefore

$$\hat{p}_a \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_a(1-\hat{p}_a)}{n+4}}$$

where $\hat{p}_a = (X + 2)/(n + 4)$.

The Clopper-Pearson confidence interval for p is the "exact" confidence interval. This interval is guaranteed to have coverage of at least (1- α) for every possible value of p. It is therefore necessarily conservative. For x=1 to n-1 the Clopper-Pearson confidence interval is

$$[1 + \frac{n-x+1}{xF_{2x,2(n-x+1),1-\frac{\alpha}{2}}}]^{-1} < \mathrm{p} < [1 + \frac{n-x}{(x+1)F_{2(x+1),2(n-x),1-\frac{\alpha}{2}}}]^{-1}$$

Where $F_{a,b,c}$ denotes the 1-c quantile of the F distribution with degrees of freedom a and b. (Agresti, 1998)

The score confidence interval, also called the Wilson confidence, has the form

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

Where $z_{\alpha/2}$ is the 1- $\alpha/2$ quantile of the standard normal distribution. The score performed better than the Wald and Clopper-Pearson in terms of having coverage probabilities closest to the nominal and can be used and "can be recommended for use with nearly all sample sizes and parameter values. (Agresti, 1998)

The last two confidence intervals are parametric bootstrap based. The Raw Percentile interval randomly takes B=100 samples of the binomial distribution to create the confidence interval for $\hat{p} = \frac{x}{n}$ using the empirical 1- $\alpha/2$ and the $\alpha/2$ quantiles of the bootstrap distribution.

The Bootstrap t Interval creates B=100 resamples of the fitted binomial distribution. For each of those resamples the quantity $\frac{\hat{p}^*_j - \hat{p}}{\hat{SE}(\hat{p}^*)}$ is calculated to approximate $\underline{\delta}$ *and* $\bar{\delta}$ to form the confidence interval

$$(\hat{p} - \underline{\delta}\,\hat{SE}\left(\hat{p}^*\right), \quad \hat{p} - \bar{\delta}\,\hat{SE}(\hat{p}^*))$$

Methodology and Results

For all of the confidence intervals, we generated N=1000 random samples from a binomial distribution for n= 15, 30, and 100. For each n we varied p from 0.01 to 0.99 for 15 total values of p. We calculated the proportion of intervals within each method that capture the true value of the probability of success, p, for each value of n and p. The results are plotted in figure 1a, b, c for various values of n.

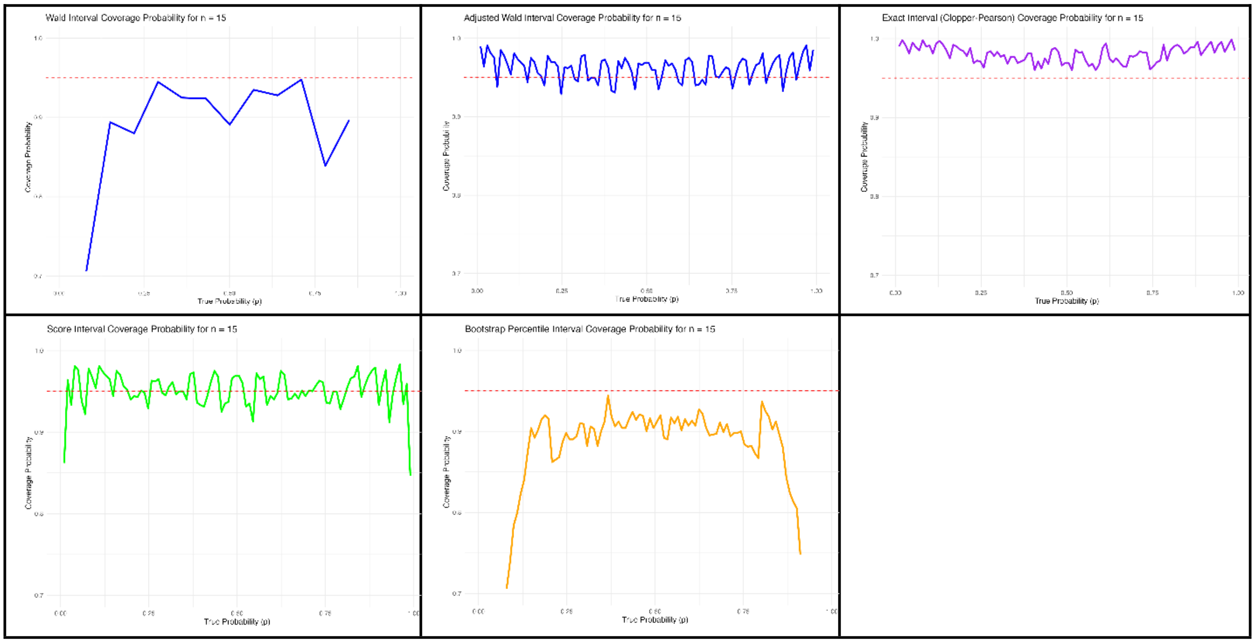

Figure 1a: Comparison of Coverage Probabilities for the binomial distribution with n=15 for the Wald, Adjusted Wald, Clopper-Pearson, Score, Raw Percentile Interval Bootstrap, and Bootstrap t Interval

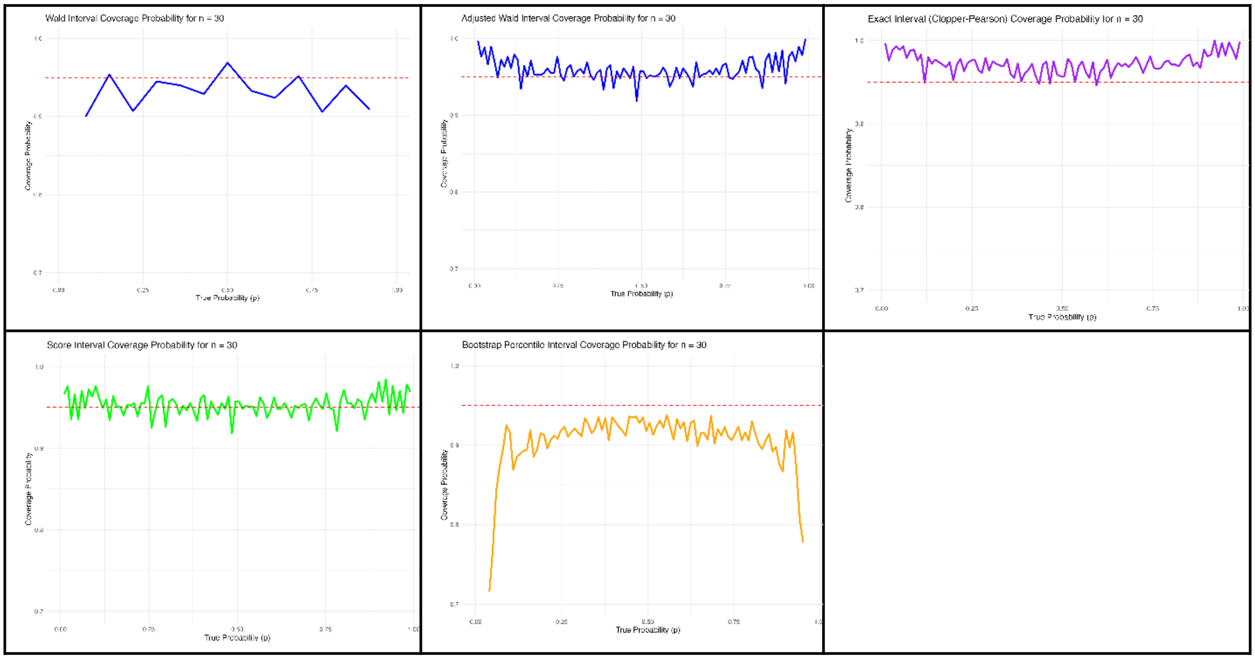

Figure 1b: Comparison of Coverage Probabilities for the binomial distribution with n=30 for the Wald, Adjusted Wald, Clopper-Pearson, Score, Raw Percentile Interval Bootstrap, and Bootstrap t Interval

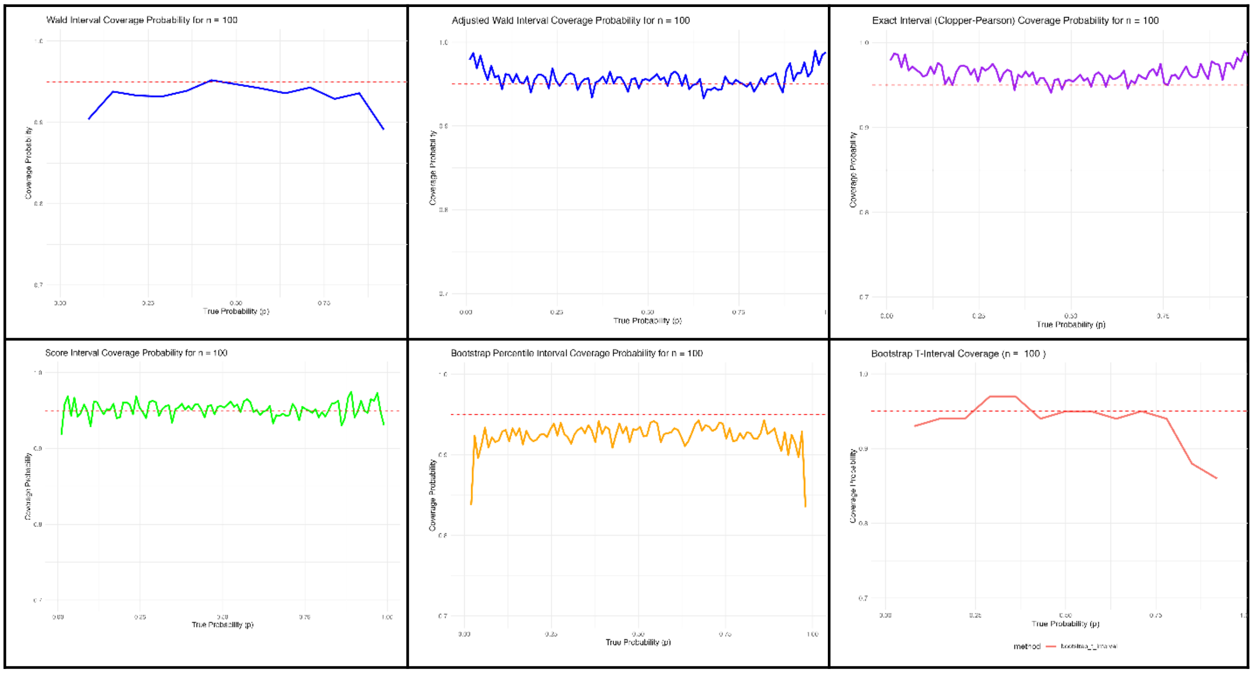

Figure 1c: Comparison of Coverage Probabilities for the binomial distribution with n=100 for the Wald, Adjusted Wald, Clopper-Pearson, Score, Raw Percentile Interval Bootstrap, and Bootstrap t Interval

We then calculated the width of the confidence intervals. Obviously we want a narrower confidence interval or the prediction value of the interval has less value. These results are plotted in figures 2 a, b, and c. We tabulated these results and included the standard error of these widths in table 1.

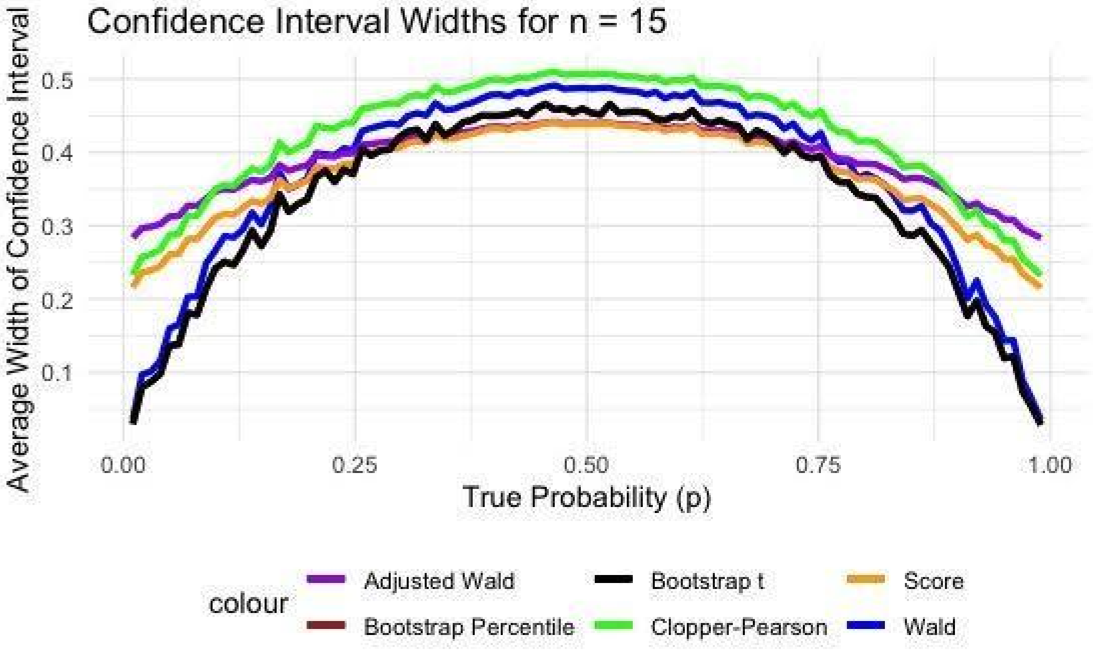

Figure 2a: Comparison of Confidence Interval widths for the binomial distribution with n=15 for the Wald, Adjusted Wald, Clopper-Pearson, Score, Raw Percentile Interval Bootstrap, and Bootstrap t Intervals

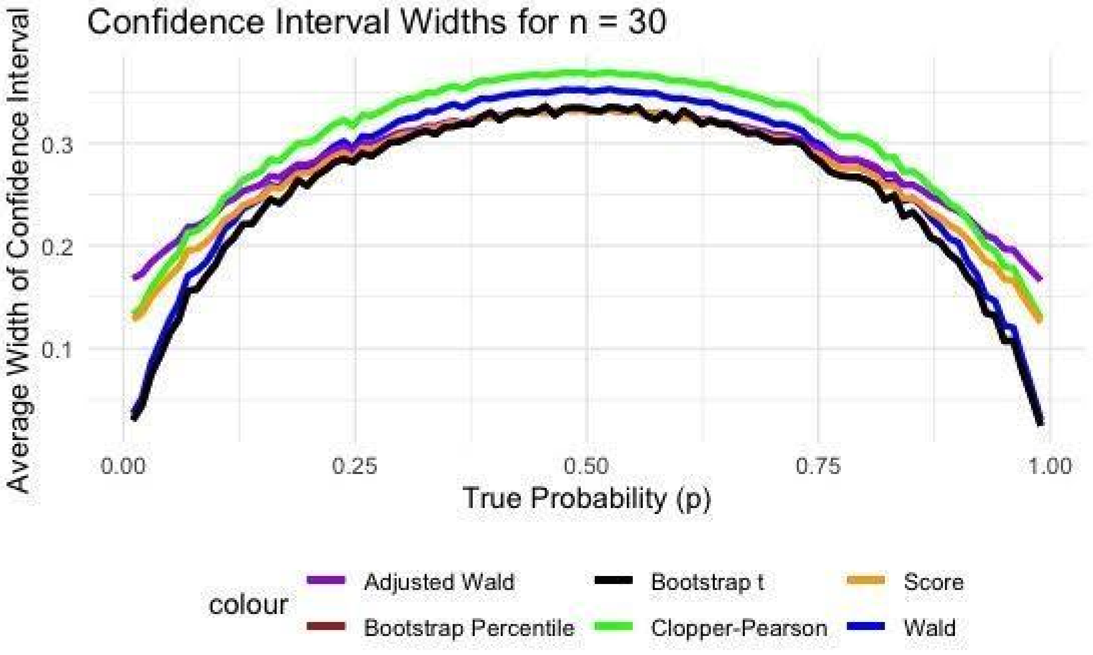

Figure 2b: Comparison of Confidence Interval widths for the binomial distribution with n=30 for the Wald, Adjusted Wald, Clopper-Pearson, Score, Raw Percentile Interval Bootstrap, and Bootstrap t Intervals

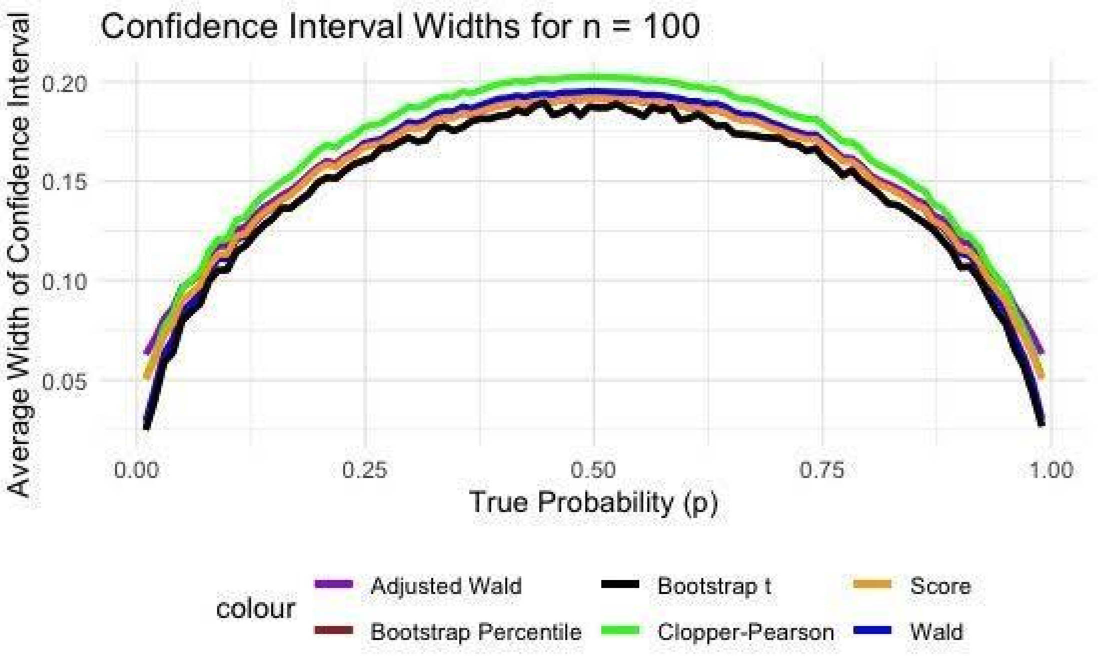

Figure 2c: Comparison of Confidence Interval widths for the binomial distribution with n=100 for the Wald, Adjusted Wald, Clopper-Pearson, Score, Raw Percentile Interval Bootstrap, and Bootstrap t Intervals

In choosing a confidence interval method, we want the actual coverage probability to be close to 95%. We want the width of the confidence interval to be as narrow as possible so it is giving us a meaningful estimate, and we want the standard error of that width to be as small as possible. In Table 1 we capture the long-running average of times the procedure is correct when used for various values of p. In (Agresti, 1998) this is discussed as possibly more relevant than looking at worst performance.

Table 1. Mean Coverage probabilities of Nominal 95% Confidence Intervals for the Binomial Parameter p

Method	n=15	se	n=30	se	n=100	se
Adjusted Wald	0.40054	0.00157	0.29271	0.00134	0.16381	0.000859
Wald	0.40088	0.00399	0.29388	0.00213	0.16408	0.00097
Exact	0.44164	0.00268	0.31666	0.00183	0.17234	0.000948
Score	0.38546	0.00217	0.28590	0.00161	0.16247	0.000909
Bootstrap Raw %	0.37200	0.00409	0.27706	0.00227	0.15672	0.00105
Bootstrap t	0.37213	0.00411	0.27709	0.00226	0.15692	0.00105

Conclusions

In choosing a confidence interval method, we want the actual coverage probability to be close to 95%. We want the width of the confidence interval to be as narrow as possible so it is giving us a meaningful estimate, and we want the standard error of that width to be as small as possible. It is also desirable to work for a wide range of parameters.

The score method, though complex, is perhaps the best performer of them all. It's long run average is very close to 95% and it has a relatively narrow average width and small standard error of the width. It performs well for all values of n and p.

Perhaps there is a place for Exact confidence intervals in statistical inference, but the definition of confidence interval I have committed to memory is "if the experiment were repeated many times, 95% of those intervals would contain the true proportion p". Under that definition, the Clopper-Pearson is too conservative and there are better performers out there.

The Wald Confidence interval method does not perform well for any of the values of n, but it has especially poor performance for low values of n. The adjusted Wald confidence interval method has a surprisingly big improvement over the Wald and is quite good even for small values of n. It works for all the values of n and p that we considered. It's width is relatively narrow as is the standard error of the width. It is also a solid performer.

The bootstrap methods may be useful for calculating confidence intervals if the data distribution is unknown, but here we are assuming a binomial and using parametric bootstrapping. The raw percentile does not achieve the 95% coverage probability. The bootstrap t distribution is much better but does not perform well for large n.

Bibliography:

1. Alan Agresti & Brent A. Coull (1998) Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions, The American Statistician, 52:2, 119-126, DOI: 10.1080/00031305.1998.10480550