Estimating the necessary sample size for a binomial proportion confidence interval with low success probabilities

by

ZACHARY AHLERS

B.S., Regis University, 2013

A REPORT

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2017

Approved by:

Major Professor
Dr. Christopher I. Vahl

Abstract

Among the most used statistical concepts and techniques, seen even in the most cursory of introductory courses, are the confidence interval, binomial distribution, and sample size estimation. This paper investigates a particular case of generating a confidence interval from a binomial experiment in the case where zero successes are expected. Several current methods of generating a binomial proportion confidence interval are examined by means of large-scale simulations and compared in order to determine an ad-hoc method for generating a confidence interval with coverage as close as possible to nominal while minimizing width. This is then used to construct a formula which allows for the estimation of a sample size necessary to obtain a sufficiently narrow confidence interval (with some predetermined probability of success) using the ad-hoc method given a prior estimate of the probability of success for a single trial. With this formula, binomial experiments could potentially be planned more efficiently, allowing researchers to plan only for the amount of precision they deem necessary, rather than trying to work with methods of producing confidence intervals that result in inefficient or, at worst, meaningless bounds.

Table of Contents

List of Figures

Acknowledgements

I'd like to thank the entire Statistics department at Kansas State University for the opportunity to learn (and teach) a fascinating subject. In particular I'd like to thank my advisor Dr. Vahl for his enthusiasm and interest in this report, which served as a great motivator to me. I also thank Dr. Jager and Dr. Dubnicka for their role as my committee members, as well as for teaching some of my favorite classes. Lastly, I would like to thank my fiancé Lisa for her support and belief in my abilities.

Chapter 1 - Introduction

Background

One of the primary methods of statistical inference is the Confidence Interval (CI), a concept taught in every introductory stats class and a tool powerful enough to be used even in complex situations. Representing an estimate for a population parameter not only by a point estimate but with an additional measure of uncertainty is an idea fundamental to the philosophy of statistics. It is because of its widespread use that the proper construction and interpretation of these CI is of the utmost importance. It is with this in mind that we consider a specific problem with respect to the construction of a CI of the probability of success in a binomial experiment. A number of methods exist for the construction of this sort of confidence interval (discussed in more detail in section 2.1), and all perform well for many values of p and n (representing the probability of success and the sample size, respectively), albeit with their respective strengths and weaknesses. However, the case that poses the most trouble is the case where the number of successes that result from a binomial experiment is either zero or very close to zero (or by reflection, close to one). In this case the likelihood function becomes increasingly skewed, which does not easily lend itself to some approximations. Of particular note is the normal approximation to the binomial distribution, which should not be used when the proportion of successes is very close to zero or one except when the sample size is (often prohibitively) very large. Similarly, many other approximations are built upon the normal approximation and suffer a similar problem. This paper seeks to compare the available options in this specific case.

Also of interest is the planning stage of an experiment. No experiment is without constraints, and many experimenters consult a statistician in order to determine what sample size is necessary to produce useful results without overextending their budget. A researcher might wish to construct a confidence interval from the resulting binomial experiment and desires the confidence interval be sufficiently narrow. The sample size must be selected so that the width of the confidence interval will likely be less that some nominal width, but keep the sample size as small as possible. This estimate depends on the method used to construct the confidence interval, and should be examined closely.

Objective

This report seeks to first conduct a comparison of the various methods of constructing a Binomial Proportion Confidence Interval and discuss the criteria for doing so. This comparison will then be used to determine which methods produce the "best" confidence intervals under various parameter values in order to form an ad-hoc methodology of constructing a confidence interval based on the data observed. This ad-hoc rule of thumb can then be used to construct an estimate for the necessary sample size given a nominal width, an estimate of p, and a factor to control the likelihood that the width will be less than the nominal width.

Chapter 2 - An Overview of Methods and Properties

There are numerous methods for constructing a confidence interval for a binomial proportion, many of which are listed in section 2.1. This is not intended to be an exhaustive list, but should include the most common methods in use today. They will be compared using 3 criteria of varying importance, described in section 2.2. Below and throughout the remainder of the paper z will represent the $1-\frac{\alpha}{2}$ quantile of the normal distribution, n is the sample size, X is the result of the binomial experiment (number of successes), p is the binomial probability of success, and $\hat{p}$ is X/n.

Current Methods

Normal

This is arguably the most common method of approximating the distribution of $\hat{p}$, and is simple enough to be taught even in introductory statistics courses. This method is often justified by use of the central limit theorem as the sampling distribution of p approaches normality when n is large. As such, it becomes an increasingly poor approximation for small sample sizes and where the likelihood becomes skewed, as in the case where p is close to 0 or 1. The confidence interval is as follows:

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Wilson's (with and without Continuity Corrections)

Wilson's method is derived from Pearson's chi-square test, which corresponds to using the normal distribution but with an estimate of p that is a weighted average of $\hat{p}$ and 0.5, with the weighting depending on the confidence level and sample size. The bounds can be calculated as shown below:

$$\frac{1}{1+\frac{z^2}{n}}\left[\hat{p}+\frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z^2}{4n^2}}\right].$$

Adding a continuity correction aligns the minimum coverage instead of the average coverage:

$$\frac{1}{2(n+z^2)}\left[2n\hat{p}+z^2 \pm \left(1+z\sqrt{z^2-\frac{1}{n}+4n\hat{p}(1-\hat{p})+(4\hat{p}-2)}\right)\right].$$

(Wilson (1927) and Newcombe (1998)). If either bound is outside [0,1], it is replaced with the corresponding limit.

Agresti-Coull

This method adds z^2 hypothetical observations to the data (half successes, half failures), then uses the normal approximation as above. This is the rationale behind the "plus four" methods taught in introductory classes:

$$\tilde{n} = n + z^2, \qquad \tilde{p} = \frac{1}{\tilde{n}}\left(X + \frac{z^2}{2}\right), \qquad \tilde{p} \pm z\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}$$

(Agresti (1998)).

Jeffrey's Interval

This method gives a Bayesian credible interval, using the conjugate Jeffrey's prior for a beta distribution, $Beta(\frac{1}{2}, \frac{1}{2})$, which results in a posterior distribution of $Beta(X + \frac{1}{2}, n - X + \frac{1}{2})$. The interval is defined as the two corresponding quantiles of this posterior distribution, $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$. This method has the additional property of being more decidedly equal-tailed, whereas other methods can be centered too close to 0.5; although in the case where $X = 0$ we will use a one sided CI instead so that the lower bound of the CI is equal to 0. Several other techniques already do this, as they are not inherently equal-tailed (Cai (2005)).

Clopper-Pearson

This is the "exact" method, as it is based directly off the cumulative binomial probabilities. It selects the interval by defining the entire set of p such that

$$P(X \geq x) > \frac{\alpha}{2} \;\; and \;\; P(X \leq x) > \frac{\alpha}{2}$$

(Clopper (1934)).

Rule of 3

This method is a rule of thumb used in medicine as a simple way of estimating a 95% confidence interval when no successes are observed. This is precisely the situation we are most interested in, and though it is not immediately clear why it should work, it has been included

here to determine its usefulness as more than a rule of thumb. The interval always has a lower bound of 0, and an upper bound of $3/n$ (Simon (2010)).

The Criteria for Comparison

Width

The method that will give the narrowest interval while maintaining the same coverage is preferable as that indicates that it is capturing the areas of higher density and in some cases may be closer to an equal-tailed confidence interval rather than capturing the entirety of a single tail and none of the other.

Reasonable Values

The method should not give values that have no useful interpretation, for example when the lower bound is less than 0. This is not strictly necessary, as the confidence intervals can always be "clipped" to be within [0,1], but it is useful for researchers not trained in statistics who may not recognize an impossible value.

Coverage

Arguably the single most important factor is that the overall coverage of each method should be as close to the nominal coverage as possible. Any significant deviation from the nominal coverage indicated that the method is producing confidence intervals that are either too wide (when the coverage is too large) or too narrow (when the coverage is too small). In this paper, the nominal coverage will be fixed at 95%, though all these simulations could be carried out for a different coverage if desired.

It is worth noting that the "coverage" used in this paper is not truly the frequentist coverage, which is defined for a single value of p across all resulting values of X. Here X is fixed and the coverage is a measure of what proportion of p's that fall inside the confidence interval, weighted by the p's likelihood of resulting in x successes.

Chapter 3 - Simulation

The Case of Zero Successes

For the first two comparison criteria we can simply calculate the confidence intervals directly from hypothetical scenarios without the need for any simulation. We use 41 different values of *n*, which are roughly evenly spaced on a log scale, beginning at 3 and ending at 32,000. Figure 1 shows the widths given by each method over our examined range of *n* on a logarithmic scale. If we multiply the widths by *n*, we can get an easier-to-read graph in terms of *x* instead of $\hat{p}$. This can be considered as a transformation of interpretation such that the y-axis represents the number of successes "wide" the confidence interval is for each sample size. This produces a graph that is more easily interpretable.

The normal approximation is based on an estimate of the standard error that does not give a meaningful answer in this case and results in a "Confidence Interval" from 0 to 0 regardless of sample size. The rule of three is perfectly flat (as it basically says that the CI for p-hat is between 0 observed successes and an additional 3 hypothetical successes). But every other method accounts for an increasing relative width as the sample size increases. Which interval is widest (relatively) changes depending on the sample size, but after about 50 observations the methods remain in a consistent order with no further crossings.

Now we consider the lower bounds of each confidence interval in Figure 3. It is perhaps the most reasonable to expect the lower bound to be equal to 0, as all but two methods do. The continuity corrected variation of Wilson's method has a strictly positive lower bound, which does not pose a computational problem, but might be philosophically unappetizing. The Agresti-Coull method produces a lower bound outside the realm of possibility, making it one method which should be avoided by researchers unable to recognize an impossible value. Although if this is acknowledged and corrected for (by replacing the bound with 0), it needn't necessarily cause a problem. This also explains why the CI were so wide in **Error! Reference source not found.**, as a portion of the CI was describing a meaningless interval. Correcting for this results in an updated graph of the widths. Now Agresti-Coull is roughly parallel with and slightly narrower than the Wilson's interval with continuity correction.

The last thing to consider is perhaps the most essential. If a confidence interval claims to have $\alpha = 0.05$, it is extremely important that the amount of coverage the CI actually has is 95% (at least nominally). While calculating the confidence intervals for a variety of hypothetical

sample sizes could be done by straightforward calculation, the estimation of coverage will be done by simulation for simplicity.

Given that x is fixed, let's sample over a large set of possible p (for efficiency we will ignore any p that has a very small (<0.00001) probability of resulting in an $x = 0$, which would slow down the brute-force simulation). The amount of simulation error in the coverage estimates can be estimated ahead of time by considering that each simulation is independent of each other with some probability of being in the calculated CI (the coverage probability). The number of successes in our simulation therefore follows a binomial bistribution, of which we are trying to estimate the proportion of successes (being in the CI!). To prevent any potential recursion within this paper, we will simply use the typical standard error formula to estimate the simulation error: $s.e. = \sqrt{q(1-q)/n}$, where q is the coverage.

The coverage simulations all have 100,000 iterations for each level of n, which even in the worst case (50% coverage) will result in a normally aproximated standard error of only 0.158%, and if the coverage rate is closer to the nominal 95% will have an error of only 0.069%. In Figure 5, a band about each point is shown, which represents a three-SE distance from each point. Of course, the normal distribution has no coverage as the "interval" it creates doesn't capture any possible $p \neq 0$ and so will be ignored for the moment.

Simulation Algorithm

1. Select a value of n.
2. Select a CI method and construct a Confidence Interval for n trials and 0 successes.
3. Select a random p uniformly from all p such that $X \sim Bin(n, p)$ & $P(X = 0) > 0.00001$.
4. Draw an $X \sim Bin(n, p)$.
5. If $X = 0$, determine whether p falls within the CI.
6. Repeat Steps 3-5 until 100,000 cases have been observed.
7. Calculate the Coverage as the number of p's that fell in the CI divided by 100,000.
8. Repeat for all CI methods and values of n.

Now we can see how each of the methods compares in terms of actual coverage over different values of *n*. Some methods tend to increase in coverage with larger *n*, while others tend to decrease in coverage. But after about *n* = 50 they fall into a consistent order (as with the widths) with Agresti-Coull resulting in the highest coverage and Jeffrey's method giving the lowest. Interestingly the only method that appears to approach the nominal 95% coverage is the simplest of the group, i.e. the Rule of 3. It is worth taking a slight tangent to investigate how this works.

How Does the Rule of 3 Work?

We are concerned with the specific case where *p* is likely to be small as the number of successes is zero or near-zero. The Poisson approximation starts to apply well here (as *n* increases and the corresponding likely values of *p* decrease), so if we write out $P(X = 0)$ over *n* trials with a rate of *p* for each trial, we get:

$$P(X = 0) = \frac{(np)^0 * e^{-np}}{0!} = e^{-np}.$$

To consider the largest *p* for which the probability of getting 0 successes is alpha, $P(X = 0)$ should be equal to alpha, say 0.05, so that $e^{-np} = 0.05$ and

$$p = \frac{-\ln(0.05)}{n} = \frac{2.99573}{n} \approx \frac{3}{n}.$$

This gives the largest value that *p* can take such that the probability of getting 0 successes is alpha under the Poisson approximation of *p* going to 0 and *n* going to infinity proportionally.

Which Method to Use?

At this point we are in a position to consider which method does the "best" job when describing zero-success data. In this case we don't have to look very hard, as the near perfect coverage (which was the most important criteria) of the Rule of 3 makes it a clear frontrunner with no other method demonstrating surprising aptitude. Though when *n* is less than about 15, the coverage becomes significantly higher and Wilson's method is arguably preferable.

So, Are We Done Yet?

Not quite. The Rule of 3 seems to be obviously superior when *X* = 0 (at least for sufficiently large sample sizes), but we need to plan for the inevitable, i.e., what happens if we

plan for 0 successes, but get some anyway? Let's consider the methods for other cases and see what happens.

Cases where $X = 1$ or $X = 2$

Unsurprisingly, the Rule of 3 doesn't apply well any more as it gives the same confidence interval regardless of the value of $\hat{p}$. The normal approximation has improved (as it would effectively be impossible not to), but is still the worst of the bunch in terms of coverage, not to mention having a negative lower bound. For large samples, Jeffrey's and Wilson's intervals are the closest to the nominal 95% coverage, though the latter is slightly conservative (~96.2%) and the former is slightly permissive (~94.1%) as shown in Figure 6. A similar result is achieved using $X = 2$, seen in Figure 7, ignoring the Rule of 3 which gives progressively worse results as the observed value of X increases.

Constructing an Ad-Hoc Methodology for a Binomial CI

Which method is best? Both Wilson's and Jeffrey's look similarly close to nominal in the nonzero case. In general we tend to prefer conservative estimates, but it is also worth noting that Jeffrey's method is significantly narrower than Wilson's. This should be unsurprising as coverage and interval width are directly connected, but this level of difference indicates that Jeffrey's interval is a more efficient one, and indeed it is the only truly equal-tailed method of the group. This can be shown by comparing the widths of both intervals at the sample sizes where they both have very similar coverages, with Jeffrey's interval being consistently narrower in all tested situations. We are therefore much more likely to have captured the regions of higher density with Jeffrey's method better than with Wilson's. While this makes Jeffrey's method preferable, it does not provide an easy calculation of width as Wilson's does which a necessary step in the remainder of the paper. With that in mind, the rest of the paper will use Wilson's interval as a tool for sample size estimation, even though Jeffrey's method provides arguably superior results.

We now have everything we need for the basis of an ad-hoc method for generating Binomial Proportion Confidence Intervals. When our sample size is low (<20) Wilson's method gives the closest to the nominal coverage, regardless of the number of successes. When the sample is large, the Rule of 3 is used for zero-success samples and Wilson's method (or

Jeffrey's) is used in the nonzero case. This ad-hoc method should produce a useful confidence interval for any binomial experiment, but we'll also use it to plan an experiment.

Figures

Figure 1. The widths of the confidence interval produced by each method

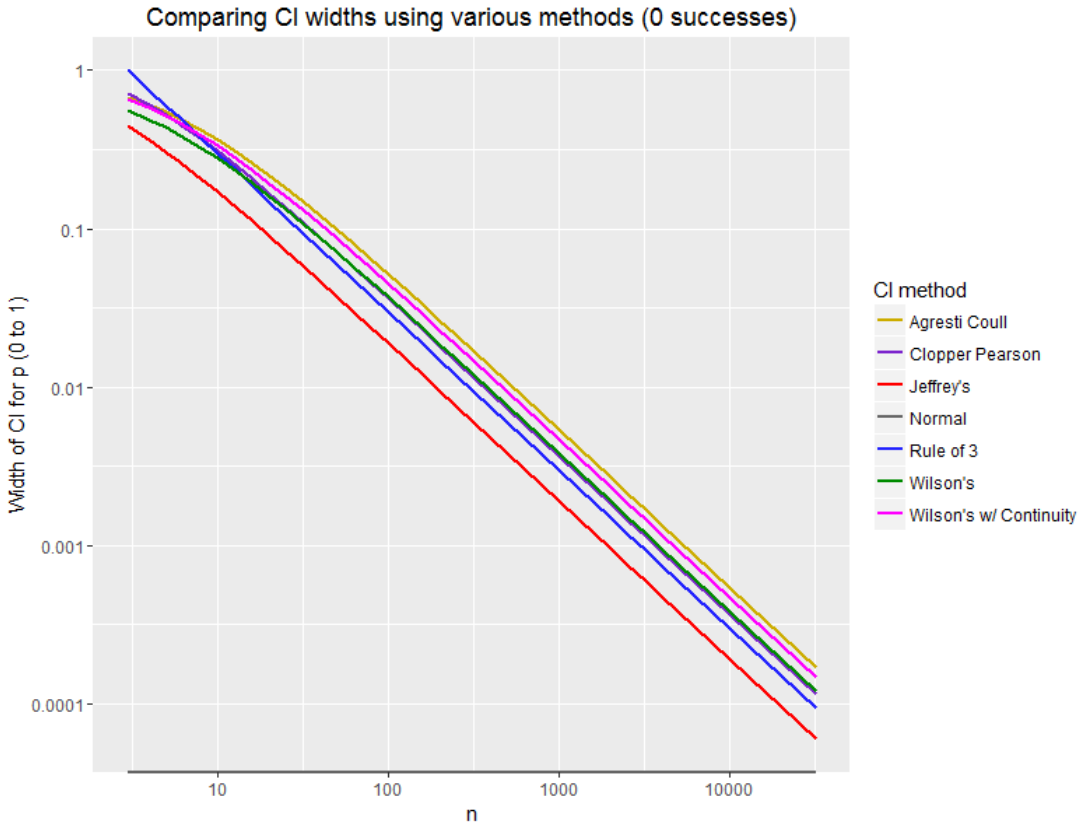


Figure 2. The widths of the confidence interval in terms of the number of successes.

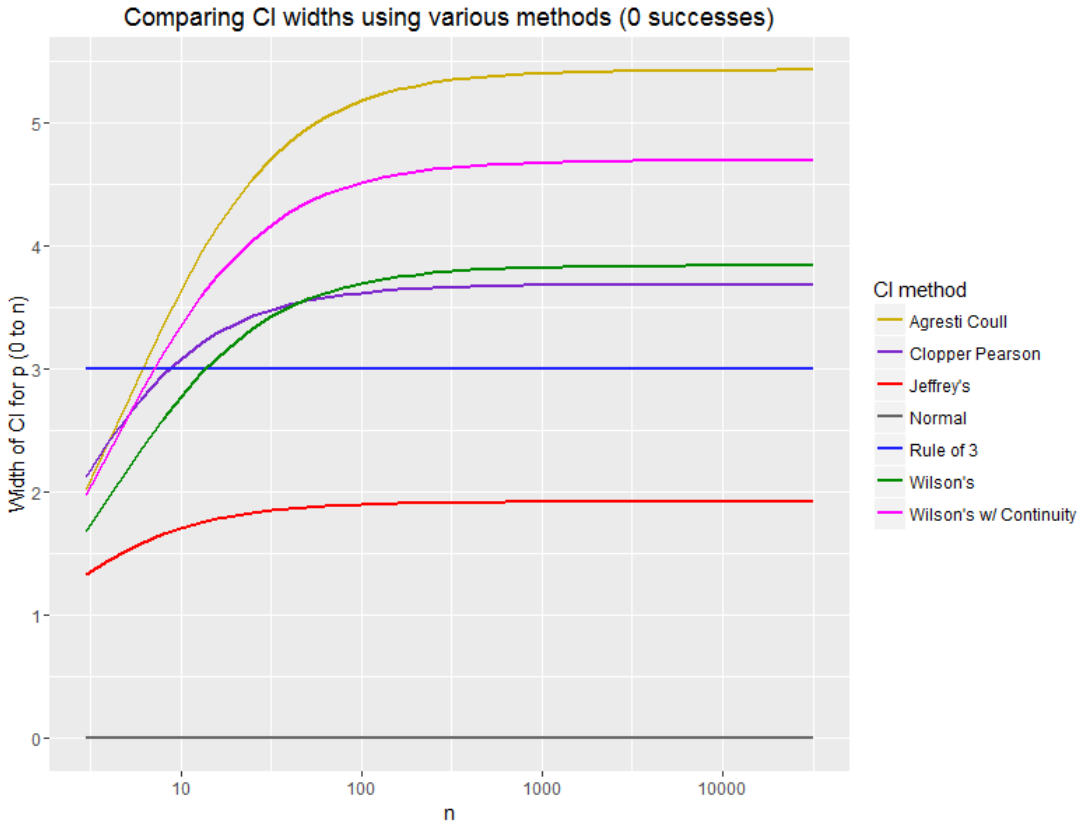


Figure 3. The lower bounds of each confidence interval.

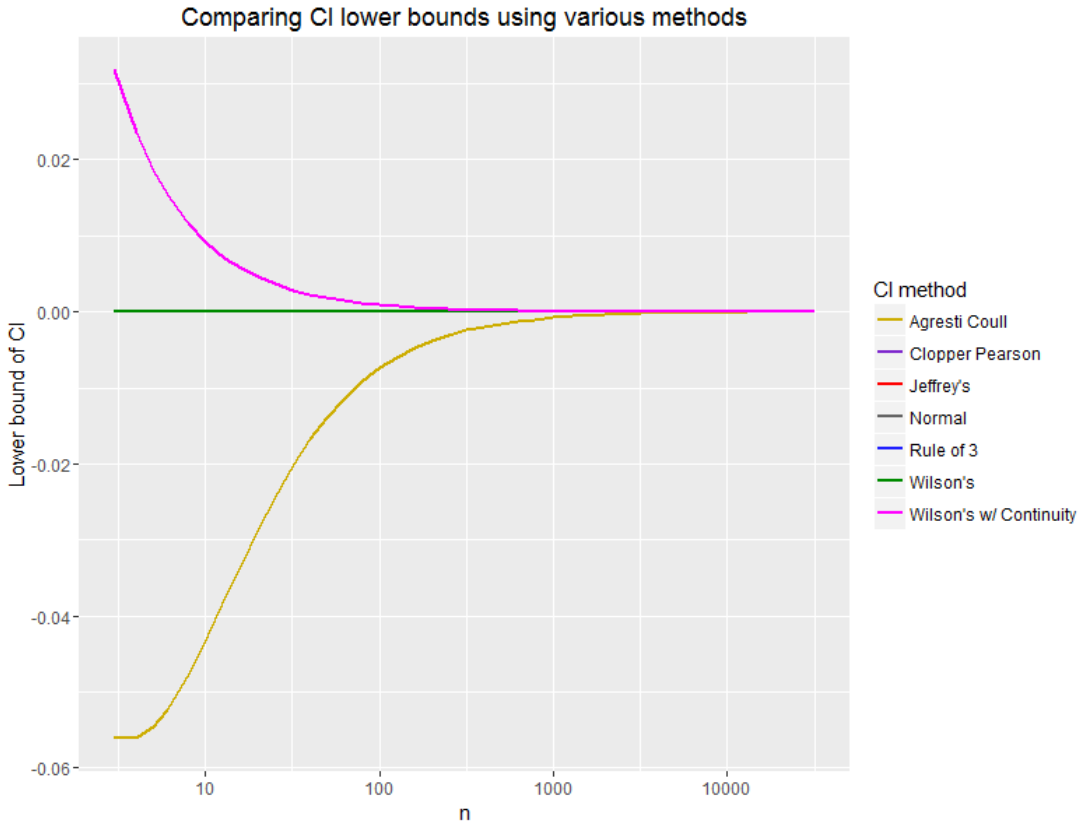


Figure 4. Widths of the confidence interval, correcting to ignore any portion of the interval below 0.

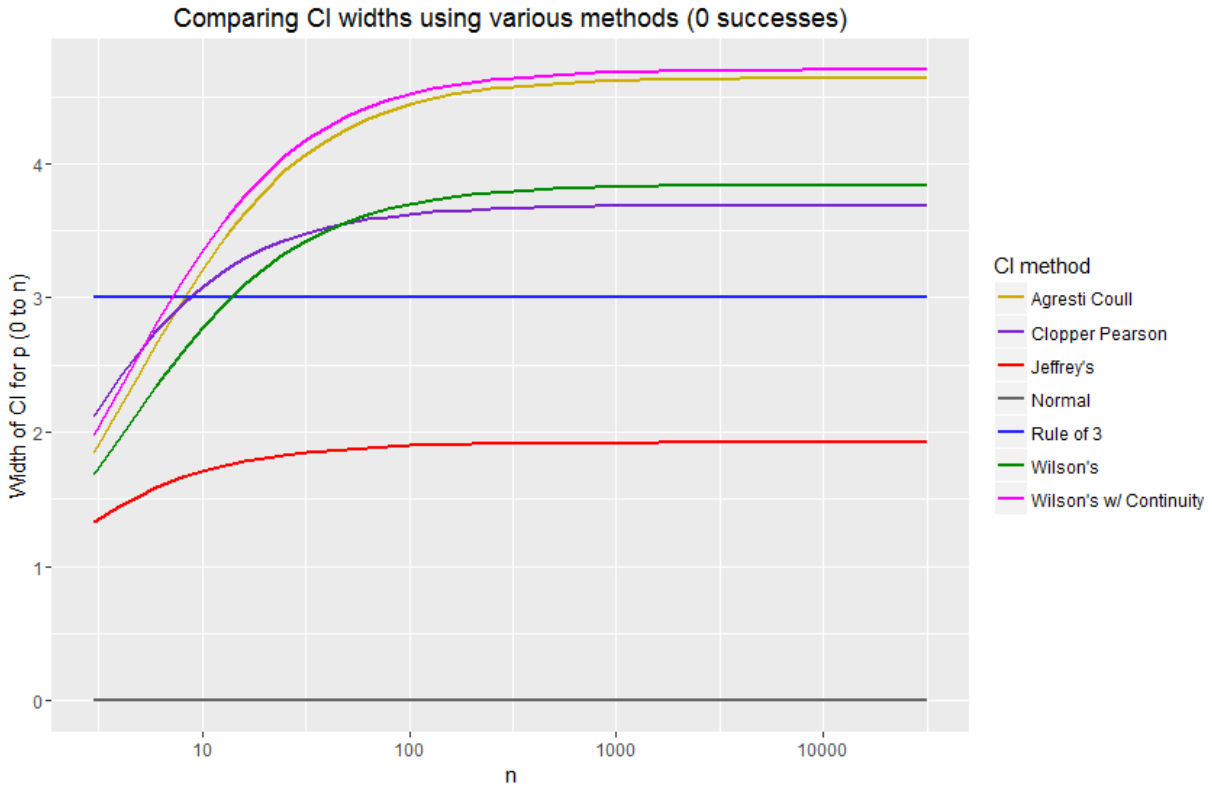


Figure 5. The Coverage for each method over different values of *n* with zero successes.

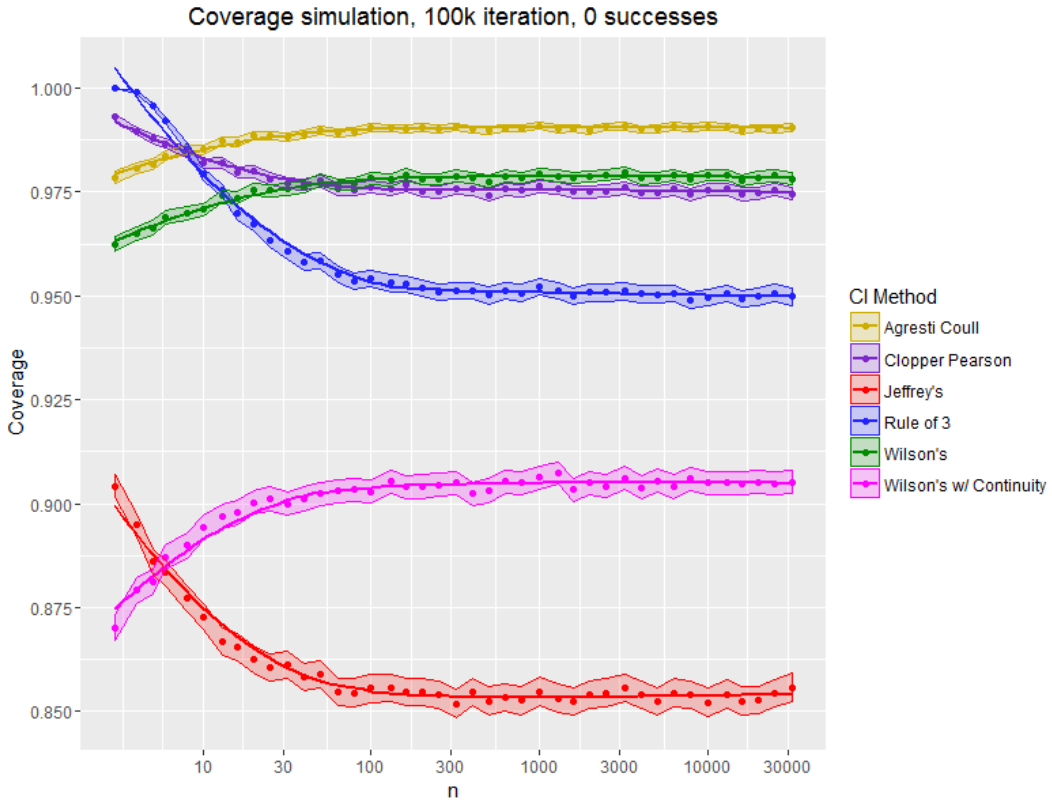


Figure 6. The coverage estimates for the $X = 1$ case.

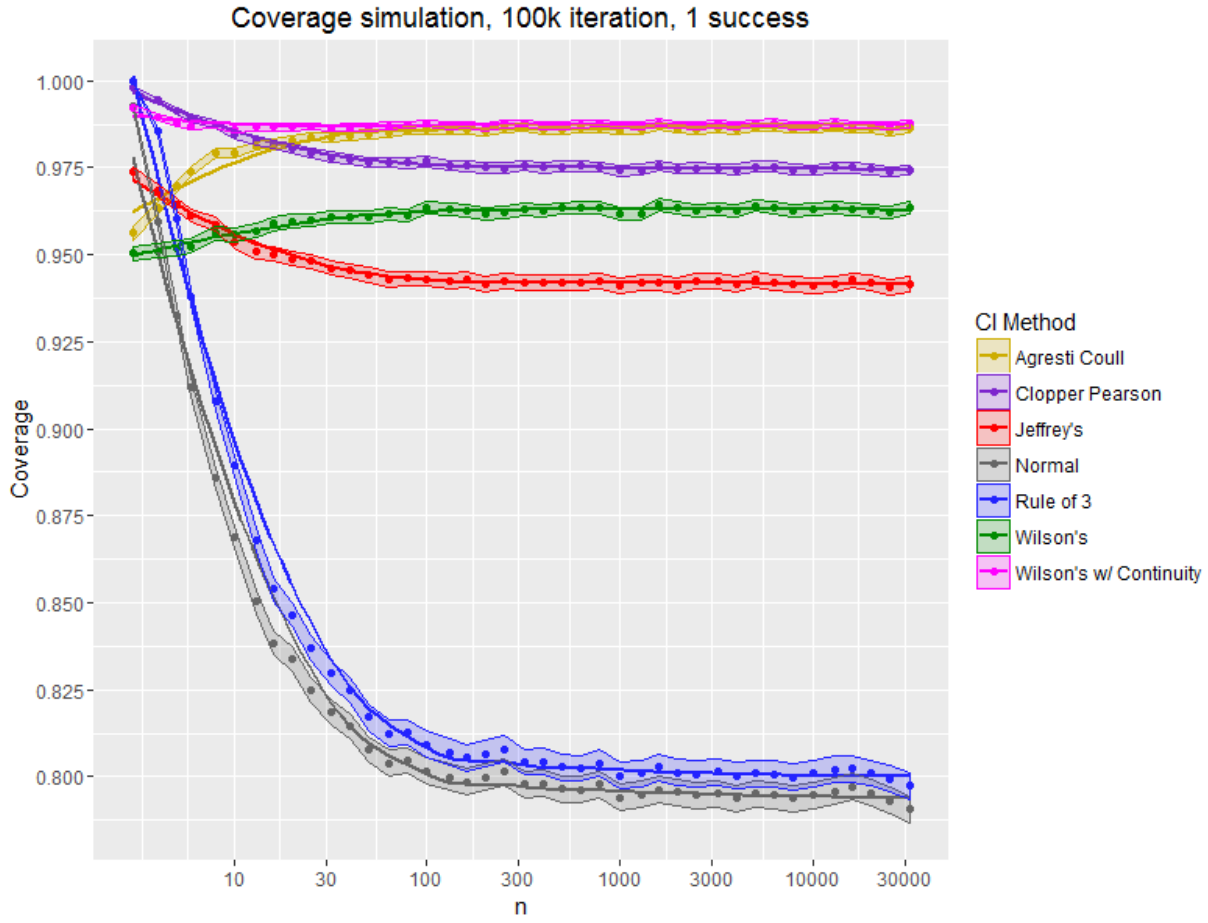
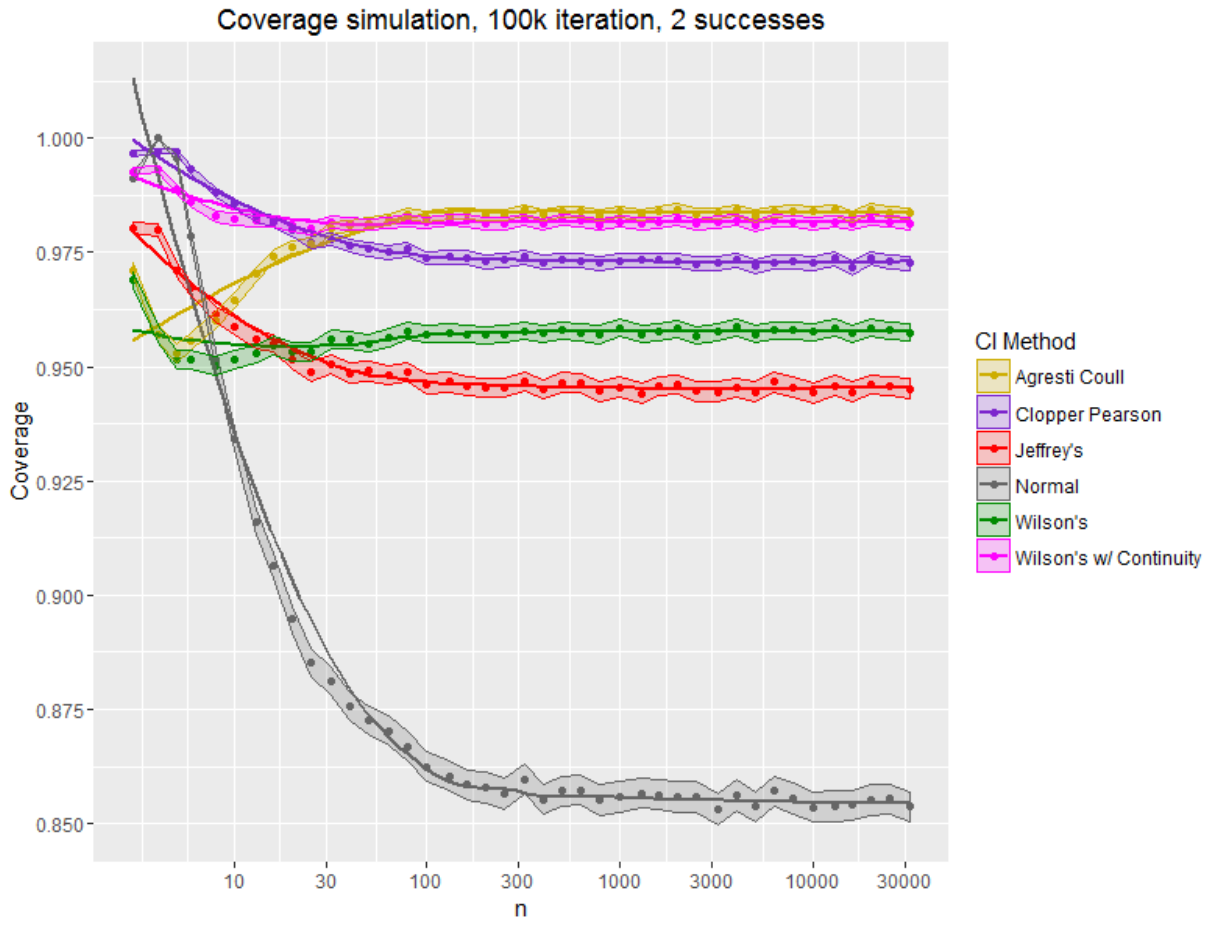


Figure 7. The coverage estimates for the $X = 2$ case.



Chapter 4 - Estimating *n* Ahead of Time

What *n* are We Estimating?

We can't be 100% certain of having a width less than our nominal width unless we are willing to assume the worst-case scenario of $p = 0.5$ which will be massively conservative most of the time for small p. We want to set some probability that we will get a width at or less than the width we want. For these examples, we use a probability of 0.8 (which will be referred to as the narrowness rate). Note that the method should work equally well for other values. If we assume our initial guess for p is correct, we can simulate a large number of ad-hoc CI for a large combination of selected p and nominal widths.

Simple Case: *X* will be 0 (Almost Certainly)

Here the only situation we need to consider is the one where we use the Rule of Three. Rearranging and solving for n gives:

$$n = \frac{-\ln(\alpha)}{width} \approx \frac{3}{width}$$

where $\alpha = 0.05$. If we estimate that the probability of getting 0 successes for that sample size is less than our nominal narrowness rate, then this sample size is sufficient and we get a width precisely equal to the nominal width.

Trickier case: *X* may be Nonzero

Despite Jeffrey's method being arguably preferable to Wilson's, it does not easily give us a way of calculating the width of the resulting confidence interval, so we'll base our initial estimate for n on Wilson's interval. We can just rearrange the equation for the width of Wilson's interval (w) to solve for n.

$$w = 2 * \frac{1}{1 + \frac{z_{\alpha/2}^2}{n}} \left[z \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}} \right]$$

which can be rearranged to form a quadratic equation in terms of n:

$$w\left(1 + \frac{z_{\alpha/2}^2}{n}\right) = 2z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}$$

then

$$w^2\left(1+\frac{z_{\alpha/2}^2}{n}\right)^2 = 4z_{\alpha/2}^2\left(\frac{\hat{p}(1-\hat{p})}{n}+\frac{z_{\alpha/2}^2}{4n^2}\right)$$

and

$$w^2 + 2w^2 z_{\alpha/2}^2\left(\frac{1}{n}\right) + w^2 z_{\alpha/2}^4\left(\frac{1}{n^2}\right) = 4z_{\alpha/2}^2\hat{p}(1-\hat{p})\left(\frac{1}{n}\right) + z_{\alpha/2}^4\left(\frac{1}{n^2}\right)$$

which can be written in the quadratic form

$$0 = w^2 n^2 + 2z_{\alpha/2}^2(w^2 - 2\hat{p}(1-\hat{p}))n + z_{\alpha/2}^4(w^2-1).$$

Using the quadratic formula results in a solution for *n* (note: we are interested in only the positive result):

$$n = \frac{-2z_{\alpha/2}^2\left(w^2-2\hat{p}(1-\hat{p})\right)+\sqrt{\left(2z_{\alpha/2}^2(w^2-2\hat{p}(1-\hat{p}))\right)^2-4w^2z_{\alpha/2}^4(w^2-1)}}{2w^2}.$$

A bit of algebra to make a cleaner form gives us:

$$n = \frac{z_{\alpha/2}^2}{w^2}\left((2\hat{p}(1-\hat{p})-w^2)+\sqrt{(w^2-2\hat{p}(1-\hat{p}))^2+w^2(1-w^2)}\right).$$

This should give an *n* that will result in the nominal width for a given $\hat{p}$. This won't give the smallest possible *n*, as can be seen when we allow $\hat{p}$ to go to 0:

$$n_{p\to 0} = \frac{z_{\alpha/2}^2}{w^2}\left((2(0)(1-(0))-w^2)+\sqrt{\left(w^2-2(0)(1-(0))\right)^2+w^2(1-w^2)}\right)$$

$$n_{p\to 0} = \frac{z_{\alpha/2}^2}{w^2}\left(-w^2+\sqrt{w^4+w^2(1-w^2)}\right)$$

$$n_{p\to 0} = \frac{z_{\alpha/2}^2}{w^2}(-w^2+w) = \frac{z_{\alpha/2}^2}{w} - z_{\alpha/2}^2$$

which (for small *w*) is significantly larger that the estimate of *n* provided by the Rule of 3, which does result in the smallest *n* needed to achieve the nominal width in that case.

Does It Work?

Let's consider a couple of hypothetical scenarios to illustrate how this works. First consider a situation where we would like our final confidence interval (α=0.05) to have a width of less than 0.01, and our initial estimate of *p* is also 0.01. Our minimum estimate of *n* is therefore:

$$n = \frac{2.996}{width} = \frac{2.996}{0.01} = 299.6 \approx 300.$$

However, this holds true only if our experiment results in zero successes and that unlikely because

$$P(X = 0|n = 300; p = 0.01) = (1 - 0.01)^{300} = 0.049.$$

Any other value of *X* will give us a larger width, as shown in Figure 8, where the lines represent different numbers of successes in the experiment. The average width is therefore far above nominal ($\overline{w} = 0.02476$).

Unless we are content with a ~5% chance of having a width less than our nominal width, we require a larger sample size, and should use the formula derived from Wilson's interval, i.e.

$$n = \frac{1.96^2}{0.01^2}\left((0.02(1 - 0.01) - 0.01^2) + \sqrt{(0.01^2 - 0.02(1 - 0.01))^2 + 0.01^2(1 - 0.01^2)}\right)$$

which becomes

$$n = \frac{3.8416}{0.0001}(0.0197 + 0.02209) = 1605.4 \approx 1606.$$

Using this sample size drastically decreases the chances of getting zero successes (to about one in ten million), but now there are many likely values of *X* that can give a width below the nominal value, bringing the average width down to nominal ($\overline{w} = 0.00992$).

This is much closer to what we would like; however, our "nominal" value is probably not the average we are hoping for. It is far more likely that we would like to control the probability of getting a width less than "nominal". For example, we might want only a 20% chance of getting a wider confidence interval than desired. Fortunately there is a relatively straightforward method of achieving such control. The equations above are based on $\hat{p}$, not p. We just assumed that the two would be close for simplicity and as a rough estimate. Once we have an estimate for *n*, we can use that estimate to consider what values $\hat{p}$ might take and with what probabilities. We can then select a $\hat{p}$ that represents a larger quantile of the distribution, and re-use our equation to estimate what *n* is needed to give an average width equal to the nominal width at this higher value of $\hat{p}$. This is an imperfect method, and does not always give the exact result desired, but it is an improvement upon the original Wilson's estimate.

Let's return to our hypothetical situation where the true value of *p* is 0.01 and we wish to have an 80% likelihood of the width of our resulting 95% confidence interval being less than 0.01. Assuming that $p = \hat{p}$ for our initial estimate, the previous method gave an estimated

sample size of 1606 (rounded up from 1605.443). At that size and probability of success, as long as the number of successes is 16 or less the resulting width will be less than 0.01. This should occur about 56% of the time (as calculated directly from a binomial distribution). There is an 80% chance that x will be 19 or less, so we can recalculate n using 19/1606=0.01183 as our new $\hat{p}$. This in turn gives us an estimated n of 1861. At our original p of 0.01, we will be under our width limit for any number of successes less or equal to than 21, which will occur 75.4% of the time. We thusly have an estimate of n that is much more closely controlled by the likelihood of achieving a CI narrower than desired. This can be improved upon, as seen in the next example.

Other Hypothetical Situations

The previous example had p and w equal to each other, and if p is much greater than w, we are in a situation where zeros are exceeding unlikely. So let's also consider the case where p is much smaller than the nominal width.

This is the situation for which the Rule of 3 estimate is best suited, i.e. where the likelihood of no successes is very high. In this particular case, the probability of zero successes is fairly large (~74.4%), but that is still below our predetermined 80% narrowness rate. We are also in the strange situation where using the second estimate (from Wilson's interval) actually decreases the probability of getting a width less than nominal, as the sample size is small enough that the width of the $X = 1$ case is still larger than nominal, but the probability of $X = 0$ decreases as the sample size increases. The third method ($n = 533$) does not provide us with a perfect estimate (it is clear that we have not quite captured the $X = 1$ possibility), but is very close, about 15 trials too small. This also occurred in the previous example, where the final capture rate was only about 75%. This is due to the fact that we found our second estimate of the sample size by using a value of $\hat{p}$ that was calculated using a quantile of the binomial distribution that depended on the first estimate of the sample size. We can produce an improved estimate by recalculating the needed sample size using our second improved estimate for n. This methodology is imperfect, e.g. if we continue repeating we can end up jumping between two estimates of n, but it does improve our estimate of n somewhat. In this case our third estimate is $n = 548.5$, which does get us to the minimum of 80% likelihood of achieving the nominal width, as $X = 1$ now gives a sufficiently narrow CI. This third iteration is the preferred estimate for n.

In General

We have considered two specific cases and have been able to produce an estimate for n that will result in a confidence interval with a pre-determined likelihood of having a set maximum width in both cases. We can now consider a larger collection of possibilities, looking only at the probability of reaching the maximum width as a response variable for a pair of p and w. The p and w are selected similarly as the sample sizes in previous simulations, i.e. to be independently evenly spaced on a logarithmic scale with all combinations of p and w considered. For any given p, the final sample size (and thus the probability of reaching nominal width) depends upon the choice of w, and w depends on p similarly, so it is perhaps most useful to consider the ratio between p and w in each case as a single predictor variable.

For a sufficiently small p relative to width, the Rule of 3 estimate for n gives the smallest estimate that has a sufficiently high likelihood of being below the nominal width, with the Wilson's method estimate using a quantile of $\hat{p}$providing a useful estimate for larger values of p (or small widths).

The Elephant in the Room

Obviously, the problem is that we do not necessarily have a good estimate for p ahead of time. (That's what this whole process is for – estimating p!) A single estimate for p can be replaced with an overall prior for p, with only the estimation of the quantile changing, but that is beyond the scope of this paper, and left as an exercise for the reader.

Figures

Figure 8. The distributions of widths generated with p = 0.01, nominal width = 0.01, and the sample size generated from the Rule of3 method (n = 300).

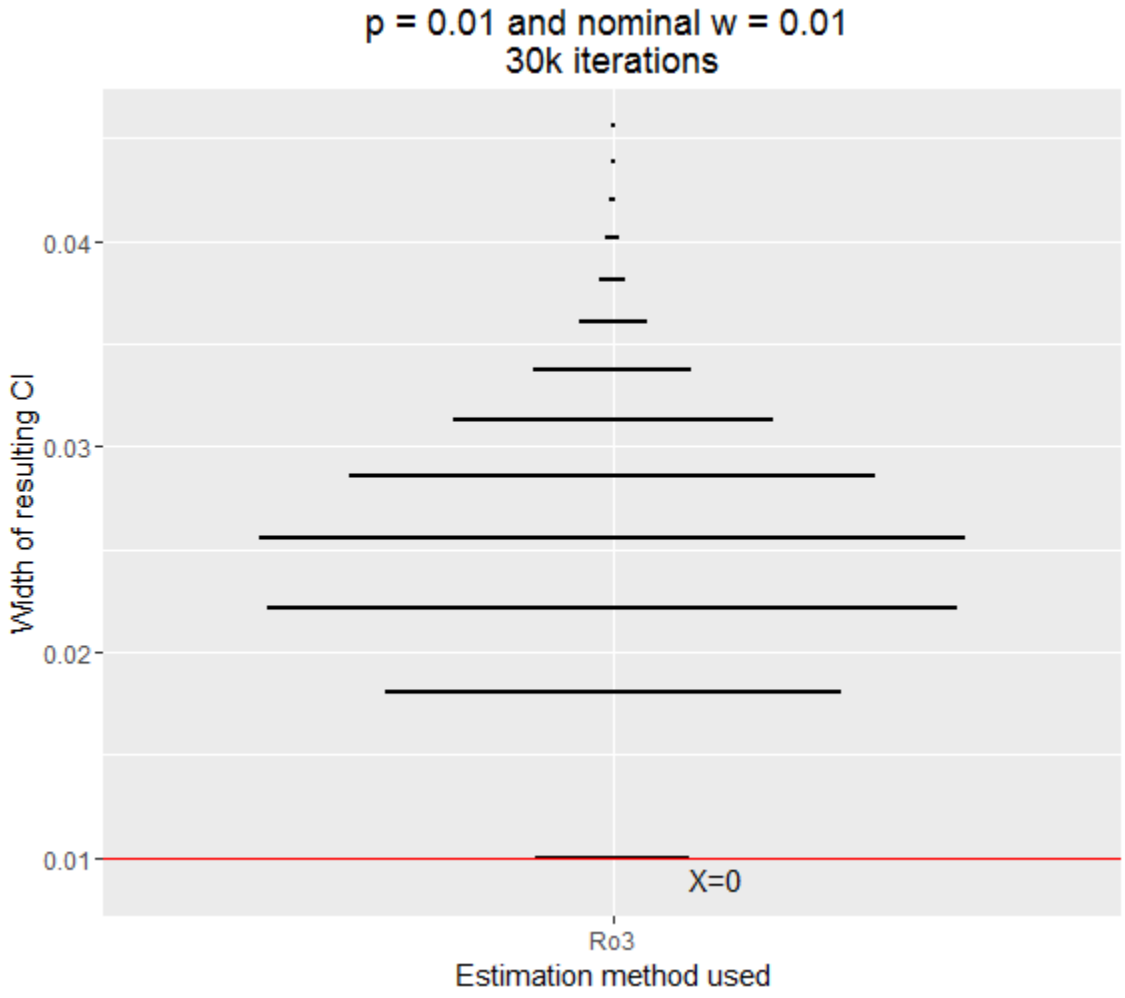


Figure 9. The distributions of widths generated with $p = 0.01$, nominal width = 0.01, and the sample size generated from Wilson's method.

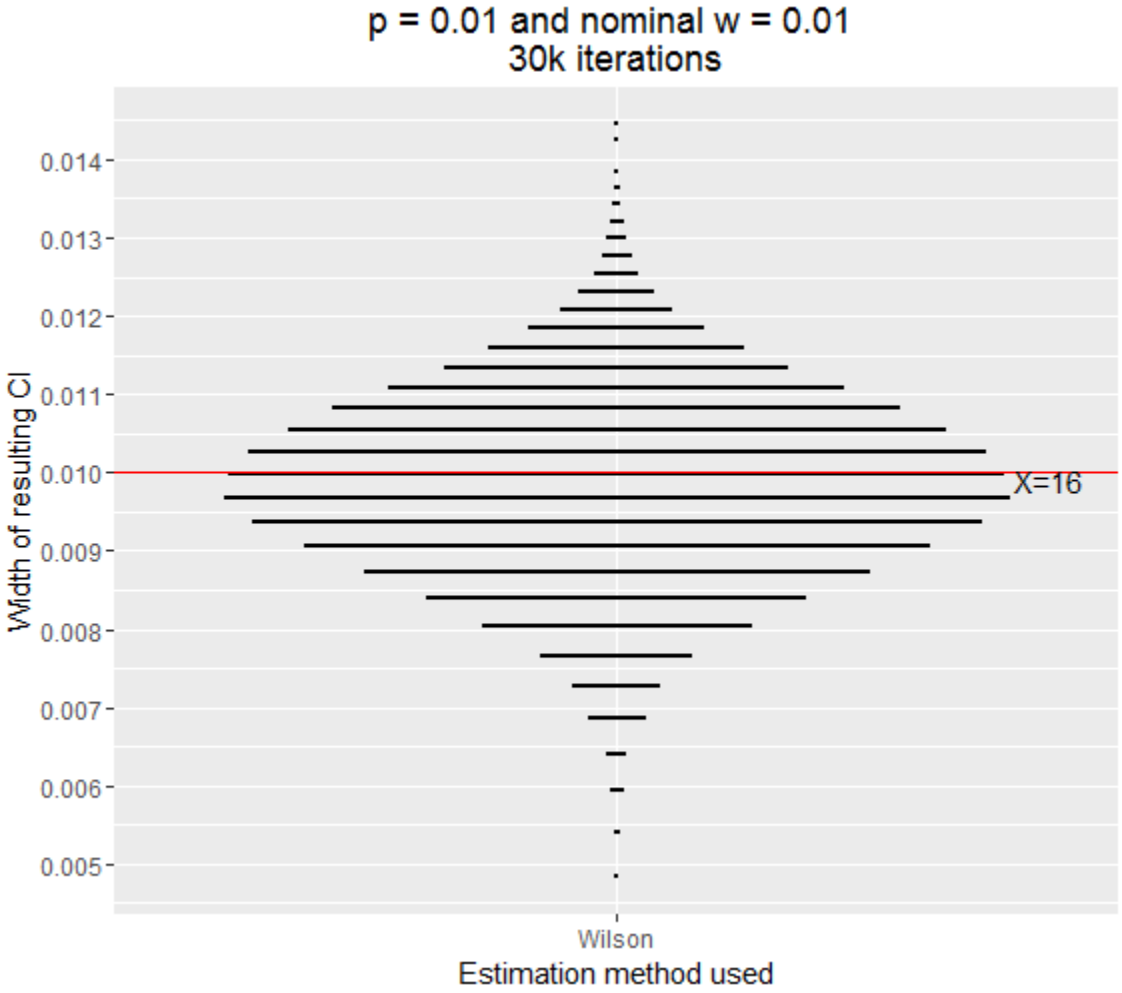


Figure 10. The distributions of widths generated with p = 0.01, nominal width = 0.01, and the sample size generated from Wilson's method using a set quantile of $\hat{p}$.

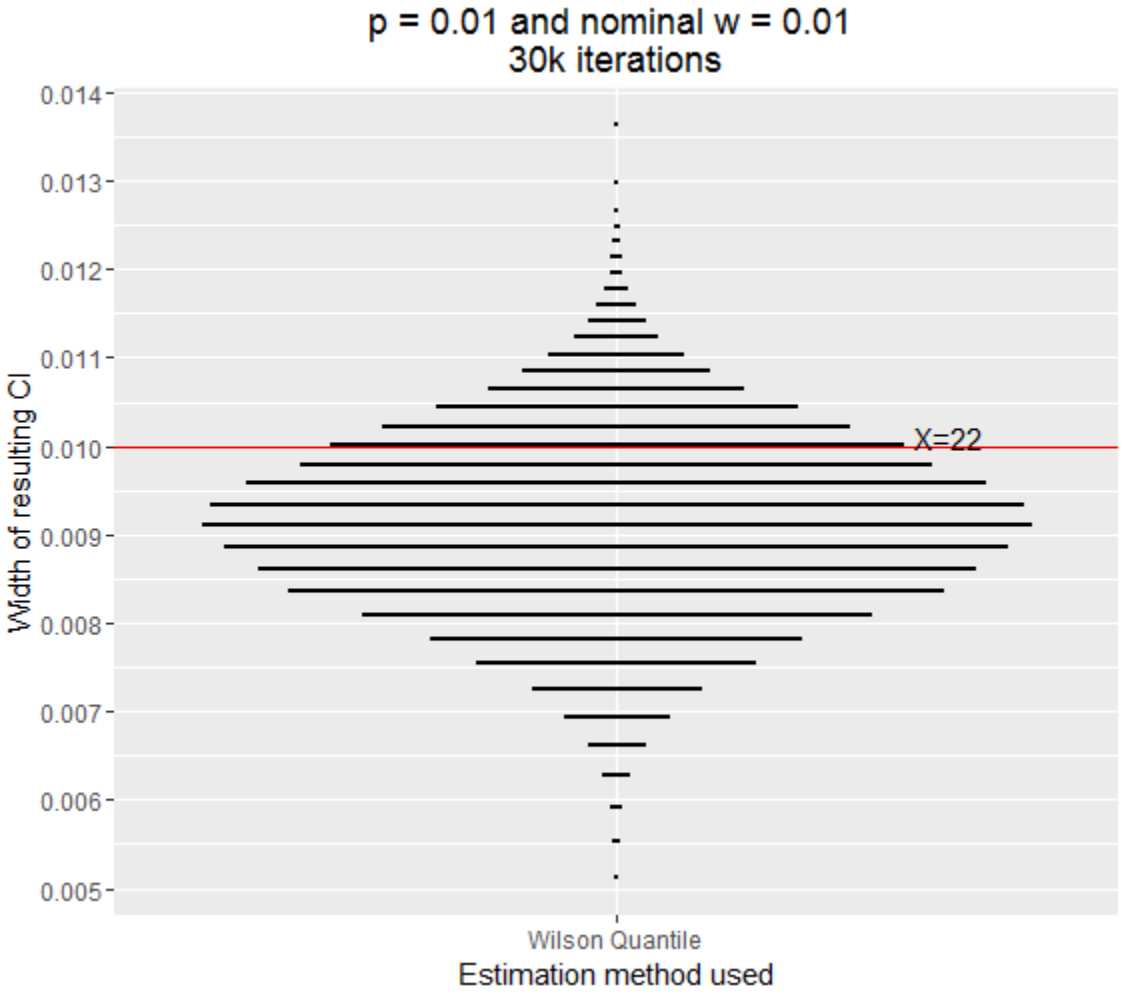


Figure 11. Comparison of 3 estimates of *n*: Rule of 3 (*n* = 300), Wilson's (*n* = 464), and Wilson's with quantile (*n* = 533).

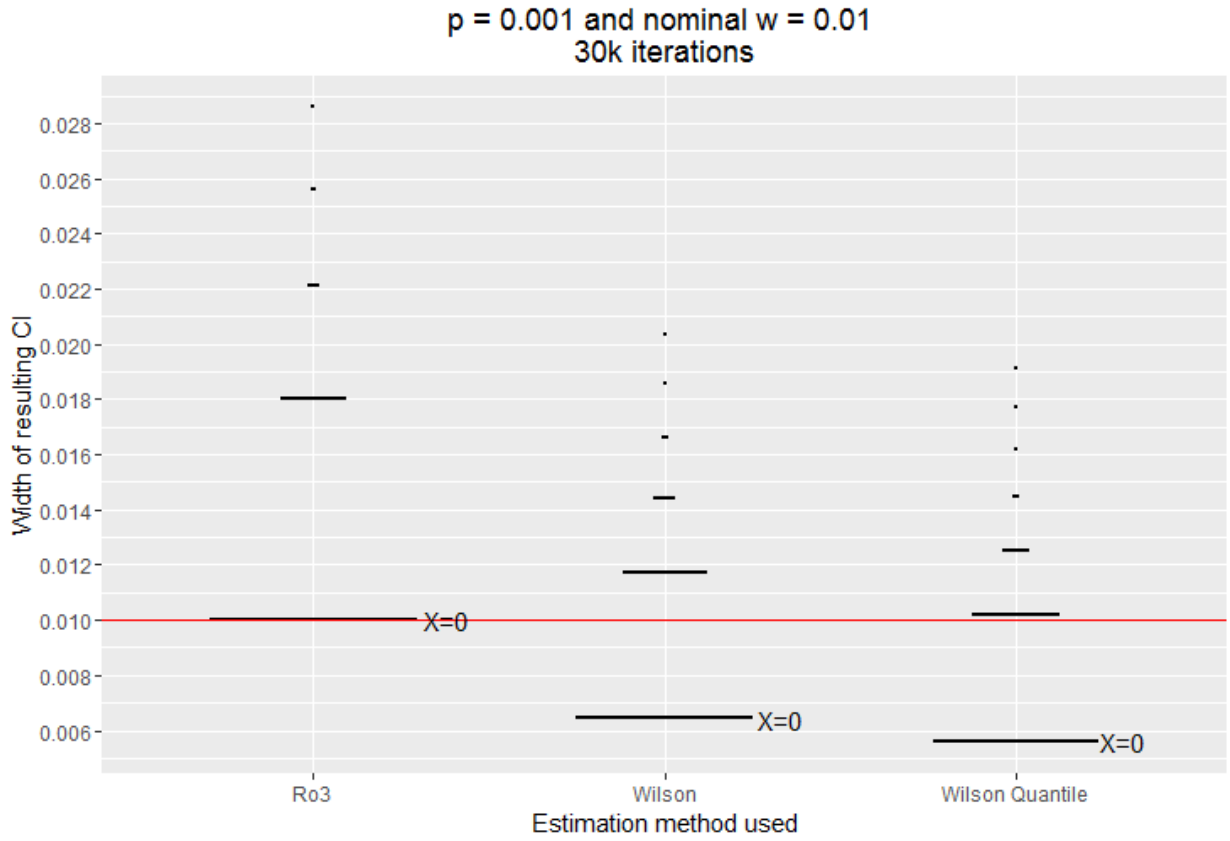
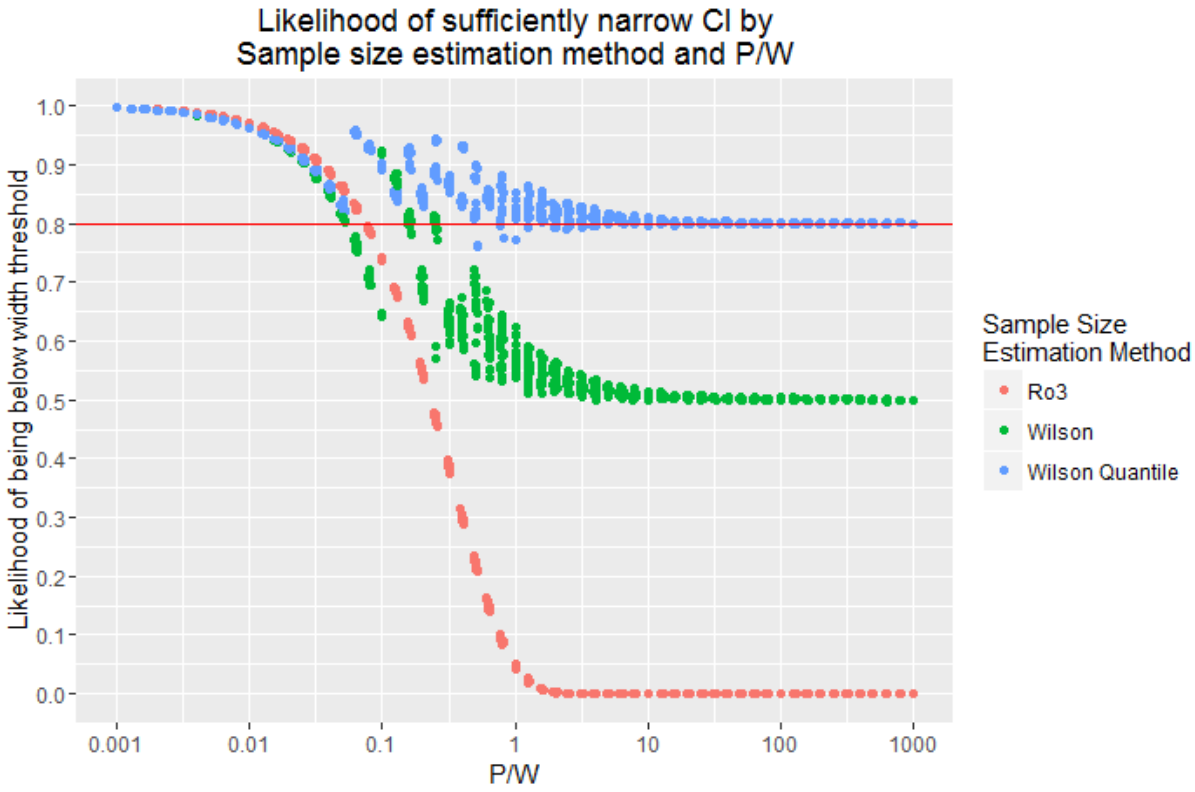


Figure 12. Rates at which estimated values of *n* result in confidence intervals narrower than the nominal width with respect to the ratio between the estimate of *p* and the nominal width.



Chapter 5 - Conclusion

Confidence intervals are exceptionally powerful tools for estimating the unknown parameters of the universe, and their proper and accurate construction is of utmost importance. When a simple approximation falls on its face, something is needed to fill the void. This paper considers the void left by the normal distribution in a common occurrence when studying rare (or exceedingly common) events. Using an ad-hoc combination of previously determined methods we can construct confidence intervals in these situations with self-assurance, knowing that while not perfect, we have a more reasonable estimate than before. The use of this method allows for more precise planning of experiments beforehand, allowing researchers to budget more effectively not only by setting an upper bound on their final CI width, but by allowing for the possibility of failure by setting an acceptable likelihood of surpassing that upper bound.

Bibliography

Agresti, Alan; Coull, Brent A. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician,* 52:2, pp. 119-126.

Belle, G. V. (2008). Statistical rules of thumb. Hoboken, NJ: Wiley.

Cai, T.T. (2005). One-sided confidence intervals in discrete distributions. *Journal of Statistical Planning and Inference,* 131:1, 1, pp. 63-88.

Clopper, C.; Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26:4, pp. 404-413.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine,* Vol.17, pp. 857-872.

Simon, Steve (2010) Confidence interval with zero events, The Children's Mercy Hospital, Kansas City, MO, http://www.pmean.com/01/zeroevents.html (last accessed 06-26-2017).

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22:158, pp. 209-212.

Appendix: R code

```
####BINOMIAL PROPORTION CONFIDENCE INTERVAL.R####
library(ggplot2)
library(dplyr)
library(gridExtra)

setwd("M:/Master's Report/wd")

alpha.level<-0.05 #confidence level

wilsons.interval<-function(alpha.level,n,num.successes)
{
  z.alpha<-qnorm(1-alpha.level/2)

  p.hat<-num.successes/n


  center<-p.hat+z.alpha^2/(2*n)
  MoE<-z.alpha*sqrt(p.hat*(1-p.hat)/n+z.alpha^2/(4*n^2))
  multiplier<-1/(1+z.alpha^2/n)
  lower<-multiplier*(center-MoE)
  upper<-multiplier*(center+MoE)

  return(c(wilson.lower=lower,wilson.upper<-upper))
}

jeffreys.interval<-function(alpha.level,n,x)
{
  p.hat<-x/n

  alpha<-0.5+p.hat*n
  beta<-0.5+n-p.hat*n

  lower<-qbeta(alpha.level/2,alpha,beta)
  upper<-qbeta(1-alpha.level/2,alpha,beta)

  return(c(lower,upper))
}

run.sim<-function(alpha.level,n,num.successes)
{
  z.alpha<-qnorm(1-alpha.level/2)

  p.hat<-num.successes/n

  #1 Normal approximation
  center.1<-p.hat
  MoE.1<-z.alpha*sqrt(p.hat*(1-p.hat)/n)
  lower.1<-center.1-MoE.1
  upper.1<-center.1+MoE.1

  #2 Wilson's interval
  center.2<-p.hat+z.alpha^2/(2*n)
  MoE.2<-z.alpha*sqrt(p.hat*(1-p.hat)/n+z.alpha^2/(4*n^2))
```

```
  multiplier.2<-1/(1+z.alpha^2/n)
  lower.2<-multiplier.2*(center.2-MoE.2)
  upper.2<-multiplier.2*(center.2+MoE.2)

  #3 Wilson's interval w/ Continuity correction
  lower.3<-max(0,(2*n*p.hat+z.alpha^2-(1+z.alpha*sqrt(z.alpha^2-
1/n+4*n*p.hat*(1-p.hat)+4*p.hat-2)))/(2*(n+z.alpha^2)))
  upper.3<-min(1,(2*n*p.hat+z.alpha^2+(1+z.alpha*sqrt(z.alpha^2-
1/n+4*n*p.hat*(1-p.hat)-4*p.hat+2)))/(2*(n+z.alpha^2)))

  #4 Jeffrey's interval
      ##WARNING, The following is a highest density CI, but **ONLY** when
x=0,
  ##otherwise it's just an equal tails CI
  alpha.4<-0.5+p.hat*n
  beta.4<-0.5+n-p.hat*n
  if(num.successes==0)
  {
    lower.4<-0
    upper.4<-qbeta(1-alpha.level,alpha.4,beta.4)
  }
  else
  {
    lower.4<-qbeta(alpha.level/2,alpha.4,beta.4)
    upper.4<-qbeta(1-alpha.level/2,alpha.4,beta.4)
  }

  #5 Clopper-Pearson
  lower.5<-binom.test(num.successes,n)[4]$conf.int[1]
  upper.5<-binom.test(num.successes,n)[4]$conf.int[2]


  #6 Agresti-Coull
  n.tilde.6<-n+z.alpha^2
  p.tilde.6<-(1/n.tilde.6)*(num.successes+z.alpha^2/2)
  lower.6<-p.tilde.6-z.alpha*sqrt(p.tilde.6*(1-p.tilde.6)/n.tilde.6)
  upper.6<-p.tilde.6+z.alpha*sqrt(p.tilde.6*(1-p.tilde.6)/n.tilde.6)

  #7 Rule of 3
  lower.7<-0
  upper.7<-3/n

  result<-
data.frame(lower.1,upper.1,lower.2,upper.2,lower.3,upper.3,lower.4,upper.4,lo
wer.5,upper.5,lower.6,upper.6,lower.7,upper.7)

  return(result);
}

set.seed(1113)
start<-proc.time()
data<-c(n=1,round(run.sim(alpha.level,1,0),10))
n.check<-
c(3,4,5,6,8,10,13,16,20,25,32,40,50,63,79,100,130,160,200,250,320,400,500,630
,790,1000,1300,1600,2000,2500,3200,4000,5000,6300,7900,10000,13000,16000,2000
0,25000,32000)
```

```
####The number of Succeses being simulated for####
X<-0                                                ####
################################################

for (n in n.check)
{
  data<-rbind(data,c(n,round(run.sim(alpha.level,n,X),10)))
}

proc.time()-start
data<-data[-1,]
#View(data)

widths<-data.frame(n=0,method=0,width=1,low=0,high=1)
methods<-c("Normal","Wilson's","Wilson's w/ Continuity","Jeffrey's","Clopper
Pearson","Agresti Coull","Rule of 3")

for(i in 1:length(data[,1]))
{
  n<-data[i,1]
  for(j in 1:7)
  {
    widths<-rbind(widths,c(n,
                           methods[j],
                           #the widths are either the diff between bounds, or
stopped at 0 or 1
                           max(as.numeric(data[i,2*j+1])-
as.numeric(data[i,2*j]),0),
                           as.numeric(data[i,2*j]),
                           as.numeric(data[i,2*j+1])))
  }
}

widths<-widths[-1,]
#View(widths)


#Coverage

min.trials<-100000

coverage<-data.frame(method=c(),n=c(),cover=c(),trials=c())

start<-proc.time()
set.seed(10102)
count<-1

for(n in n.check)
{
  num.successes<-rep(0,7)
  num.trials<-0
  while(num.trials<min.trials)
  {
    upper.bound<-min(1,qbeta(1-1e-5,2,n+2))
    p<-runif(1,0,upper.bound)
    samp<-rbinom(1,n,p)
    if(samp==X)
```

```
    {
      num.trials<-num.trials+1
      interval<-run.sim(alpha.level,n,samp)
      for(i in 1:7)
      {
        if(p>interval[2*i-1] && p<interval[2*i])
        {
          num.successes[i]<-num.successes[i]+1
        }
      }
    }
  }
  for(i in 1:7)
  {
    coverage<-rbind(coverage,c(i,n,num.successes[i],num.trials))
  }
  print(paste("N =",n,"in", round((proc.time()-
start)[3],1),"seconds,",length(n.check)-count,"iterations left"))
  count<-count+1
}

coverage[,1]<-as.factor(methods[coverage[,1]])

colnames(coverage)<-c("method","n","cover","trials")

#View(coverage)

dat<-cbind(coverage,widths[,3:5])

write.csv(dat,file=paste0("sim_data_",X,"_",min.trials,".csv"))
dat<-read.csv(file=paste0("sim_data_",X,"_",min.trials,".csv"))


#Let's make some graphs!

#Graph of lower bound (in terms of p-hat) over n
gg.low<-ggplot(dat,aes(x=n,y=low,color=method)) +
  geom_line(size=1) +
  scale_x_log10(breaks=c(3,10,30,100,300,1000,3000,10000,30000)) +
  labs(y="Lower Bound",title="Comparing CI Lower bounds using various
methods",color="CI Method")

#Graph of widths (in terms of p-hat) over n
gg.width.p<-ggplot(dat,aes(x=n,y=width,color=method)) +
  geom_line(size=0.8) +
  scale_x_log10(breaks=c(10^(1:4))) +
  scale_y_log10(breaks=c(10^-(1:4)),labels=c(0.1,0.01,0.001,"0.0001")) +
  labs(y="Width (p-hat)",title="Comparing CI widths using various
methods",color="CI Method")

#Graph of widths (in terms of x) over n
gg.width.x<-ggplot(dat,aes(x=n,y=width*n,color=method)) +
  geom_line(size=0.8) +
  geom_point() +
  scale_x_log10(breaks=c(10^(1:4))) +
  scale_y_continuous(breaks=c(0:10)) +
```

```
  labs(y="Width (x)",title="Comparing CI widths using various
methods",color="CI Method")

#Graph of widths (in terms of x) over n, correcting for negative lower bounds
gg.width.x.corr<-ggplot(dat,aes(x=n,y=(width+low)*n,color=method)) +
  geom_line(size=0.8) +
  #geom_point() +
  scale_x_log10(breaks=c(10^(1:4))) +
  scale_y_continuous(breaks=c(0:10)) +
  labs(y="Width (x)",title="Comparing CI widths using various methods
(corrected)",color="CI Method")

#Graph of Coverage
gg.cover<-dat %>%
  #filter(method!="Normal") %>%
  filter(method!="Rule of 3") %>%
  ggplot(aes(x=n,y=cover/trials,color=method)) +
  #geom_point() +
  geom_smooth(se=F,span=0.2) +
  scale_x_log10(breaks=c(3,10,30,100,300,1000,3000,10000,30000)) +
  labs(y="Coverage",color="CI Method",title="Coverage simulation, 100k
iterations") +
  geom_ribbon(aes(ymin=cover/trials-3*sqrt((cover/trials*(1-
cover/trials))/trials),
                  ymax=cover/trials+3*sqrt((cover/trials*(1-
cover/trials))/trials),
                  fill=method),
              alpha=0.1) +
  guides(fill="none")


gg.low
gg.width.p
gg.width.x
gg.width.x.corr
gg.cover

grid.arrange(gg.low+guides(color="none")+labs(title=""),
             gg.width.x.corr+labs(title=""),
             gg.cover+guides(color="none")+labs(title=""),
             nrow=3,
             top=paste0("Case X=",X))









##Determining an ad-hoc rule for confidence intervals
##the Ro3 is clearly preferable for moderate to large n when x=0,
##but it's a toss up between Wilson's and Jeffrey's in the nonzero case
##let's make a specific test to compare the two for various x and n
```

```
min.trials<-1000

coverage.wj<-
data.frame(method=c(),n=c(),x=c(),low=c(),high=c(),width=c(),cover=c(),trials
=c())

start<-proc.time()
set.seed(10105)
count<-1

for(n in n.check)
{
  for(x in 1:min(10,round(n/2-1e-3)))
  {
    w.interval<-wilsons.interval(alpha.level,n,x)
    j.interval<-jeffreys.interval(alpha.level,n,x)

    num.trials<-0
    num.successes<-c(0,0)
    while(num.trials<min.trials)
    {
      lower.bound<-qbeta(1e-5,x-1,n-x+1)
      upper.bound<-qbeta(1-1e-5,x+1,n-x-1)
      p<-runif(1,lower.bound,upper.bound)
      samp<-rbinom(1,n,p)
      if(samp==x)
      {
        num.trials<-num.trials+1

        if(p>=w.interval[1] &&p<=w.interval[2])
          num.successes[1]<-num.successes[1]+1

        if(p>=j.interval[1] &&p<=j.interval[2])
          num.successes[2]<-num.successes[2]+1
      }
    }

    coverage.wj<-
rbind.data.frame(coverage.wj,c(2,n,x,w.interval[1],w.interval[2],w.interval[2
]-w.interval[1],num.successes[1],num.trials))
    coverage.wj<-
rbind.data.frame(coverage.wj,c(4,n,x,j.interval[1],j.interval[2],j.interval[2
]-j.interval[1],num.successes[2],num.trials))
  }
  print(paste("N =",n,"in", round((proc.time()-
start)[3],1),"seconds,",length(n.check)-count,"iterations left"))
  count<-count+1
}

coverage.wj[,1]<-as.factor(methods[coverage.wj[,1]])

colnames(coverage.wj)<-
c("method","n","x","low","high","width","cover","trials")

View(coverage.wj)
```

```
write.csv(coverage.wj,file="sim_data_100000iter_Will_vs_Jeff_Partial.csv")

##In general Jeffrey's method is closer to 95% confidence
#than Wilson's for large n, and any x. However, it is non-conservative...
#However, I believe that even for equivalent coverage, Jeffrey's
#will give a narrower interval, even as an equal tailed interval

#For the moment, let's find the n for each x where the coverages
#are roughly equivalent between methods, and compare widths.

#This shows the absolute difference in widths (Will-Jeff)
wj.diff<-(coverage.wj[coverage.wj[,1]=="Wilson's",-1]-
coverage.wj[coverage.wj[,1]=="Jeffrey's",-1])
wj.diff[,1:2]<-coverage.wj[coverage.wj[,1]=="Wilson's",-1][,1:2]
wj.diff<-wj.diff[,c(-3,-4)]
View(wj.diff)

#For x=1: n=10; abs. diff=1.597072e-02; Relative diff: a 4.13% decrease from
Will
#For x=2: n=16; abs. diff=7.988008e-03; Relative diff: a 2.46% decrease from
Will
#For x=3: n=25; abs. diff=6.146978e-03; Relative diff: a 2.38 decrease from
Will
#For x=4: n=32; abs. diff=4.338469e-03; Relative diff: a 1.88% decrease from
Will
#For x=5: n=40; abs. diff=0.003378172; Relative diff: a 1.64% decrease from
Will
#For x=6: n=50; abs. diff=0.0028358826; Relative diff: a 1.56% decrease from
Will

#So it would appear that Jeffrey's method tends to give narrower intervals
for the same coverage
#This should be relatively unsurprising: since it's based on the Beta
distribution
#it can account for some skewness in the final sampling distribution. This
could potentially
#be improved by using a HDPI instead, but we'll limit ourselves to equal-
tailed CI for now

#To summarize, Wilson's interval is an all around good interval, and is
probably the best
#to use for small n (less than 15, unless x>2, then less than 10*x)
#as it will more closely match the desired coverage,but for larger n,
#the Rule of 3 is preferable when x=0, and Jeffrey's is best when x>0,
#Even though I use Wilson's interval in my paper

####N ESTIMATION.R####

library(ggplot2)
library(dplyr)
library(reshape2)

setwd("M:/Master's Report/wd")

alpha.level<-0.05
```

```
n.check<-
c(3,4,5,6,8,10,13,16,20,25,32,40,50,63,79,100,130,160,200,250,320,400,500,630
,790,1000,1300,1600,2000,2500,3200,4000,5000,6300,7900,10000,13000,16000,2000
0,25000,32000)

wilsons.interval<-function(alpha.level,n,num.successes)
{
  z.alpha<-qnorm(1-alpha.level/2)

  p.hat<-num.successes/n


  center<-p.hat+z.alpha^2/(2*n)
  MoE<-z.alpha*sqrt(p.hat*(1-p.hat)/n+z.alpha^2/(4*n^2))
  multiplier<-1/(1+z.alpha^2/n)
  lower<-multiplier*(center-MoE)
  upper<-multiplier*(center+MoE)

  return(c(wilson.lower=lower,wilson.upper<-upper))
}

jeffreys.interval<-function(alpha.level,n,x)
{
  p.hat<-x/n

  alpha<-0.5+p.hat*n
  beta<-0.5+n-p.hat*n

  lower<-qbeta(alpha.level/2,alpha,beta)
  upper<-qbeta(1-alpha.level/2,alpha,beta)

  return(c(lower,upper))
}

##estimating sample size needed, given a necessary width, confidence level,
and a prior on p

#Using a Wilson's/Jeffrey's/Ro3 ad-hoc rule for constructing a confidence
interval:
#If n<=15, use Wilson's
#Otherwise, use Ro3 when x=0 and Jeffrey's when x>1
#If p.hat is >5%, possibly use Wilson's anyway (now we're out of our specific
case of interest)

ad.hoc<-function(alpha.level,n,x)
{
  if(n<=15)
    c(wilsons.interval(alpha.level,n,x))
  else
  {
    if(x==0)
      c(0,-log(alpha.level)/n)
    else
      c(jeffreys.interval(alpha.level,n,x))
  }
}
```

```
ad.hoc.w<-function(alpha.level,n,x)
{
  if(n<=15)
    c(wilsons.interval(alpha.level,n,x))
  else
  {
    if(x==0)
      c(0,-log(alpha.level)/n)
    else
      c(wilsons.interval(alpha.level,n,x))
  }
}


#Great, now we need to determine a necessary n for a given width and prior on
p

#We can do this two ways: such that the average width is equal to the width
we want,
#or that we can be some % confident of getting a width under this maximum
width

#Begin with an n calculated from the Ro3, which will give a lower bound of n.
#n=3/width
#if x>0, the width will be larger, so n must be larger as well,
#but if n>15 we can ignore the use of Wilson's interval, as a sample of
fifteen will
#always result in a width of more than 20% (basically worthless)

for(i in 0:8)
{
  print(ad.hoc.w(alpha.level,15,i)[2]-ad.hoc.w(alpha.level,15,i)[1])
}

#So the only consideration is between the Ro3 and Jeffrey's Rule

#the Ro3 makes it easy to find n, but Jeffrey's does not.
#I don't really want to mess about with the incomplete beta function
#to estimate a width directly from an n and p-hat. However,
#I can use the fact that Wilson's interval is wider (for sufficiently small
p)
#to find an upper bound on n. Let's try it!

est.width.wilsons<-function(alpha,n,p)
{
  z<--qnorm(alpha/2,0,1)
  (2/(1+z^2/n))*z*sqrt(p*(1-p)/n+z^2/(4*n^2))
}

#only for x=0
n.est.Ro3<-function(alpha.level,w,p=0.5)
{
  return(-log(alpha.level)/w)
}

#Solves Wilson's interval equation for n in terms of p and w
n.est.wilsons<-function(alpha,w,p)
```

```
{
  z<-qnorm(1-alpha/2,0,1)
  ##equation is in form (a +/- sqrt(b))/c
  a<-2*p*(1-p)-w^2
  b<-(w^2-2*p*(1-p))^2+w^2*(1-w^2)
  c<-w^2/z^2
  result<-(a+sqrt(b))/c
  return(result)
}

n.est.iterated<-function(alpha,w,p,prob.less,iter)
{
  n<-n.est.wilsons(alpha,w,p)
  for(i in 1:iter)
  {
    p.hat<-qbinom(prob.less,ceiling(n),p)/ceiling(n) #occasionally  will give
capture rates less than expected
    n<-n.est.wilsons(alpha,w,p.hat)
  }

  return(n)
}

#A quick check to make sure everything is working

hyp.p<-0.001

hyp.width<-0.01

n.est.Ro3(0.05,hyp.width)
n.est.wilsons(0.05,hyp.width,hyp.p)
n.est.iterated(0.05,hyp.width,hyp.p,0.8,10)

#let's test these values of n, assuming the estimate of p is correct,
#resulting in the distribution of widths and the percent of which are less
than required
#first we'll test for just one set of p and w...

iter<-30000

n.est<-c(n.est.Ro3(0.05,hyp.width),
         n.est.wilsons(0.05,hyp.width,hyp.p),
         n.est.iterated(0.05,hyp.width,hyp.p,0.8,1))

zeros<-rep(0,3*iter)

data.width<-data.frame(n=zeros,x=zeros,low=zeros,high=zeros,width=zeros)

set.seed(20314)

start<-proc.time()

use.wils<-T

count<-0
for(n in ceiling(n.est))
{
```

```
  samp<-rbinom(iter,n,hyp.p)
  for(i in 1:iter)
  {
    if(use.wils)
      interval<-ad.hoc.w(alpha.level,n,samp[i])
    else
      interval<-ad.hoc(alpha.level,n,samp[i])
    data.width[iter*count+i,]<-c(n,samp[i],interval,interval[2]-interval[1])
    if(i%%(iter/100)==0) {print(paste(i,(proc.time()-start))[3])} #This lets
me see how it's progressing
  }
  print(n)
  count<-count+1
}

proc.time()-start

data.width<-cbind(data.width,as.factor(data.width[,1]))

colnames(data.width)<-c("n","X","low","high","width","method")

levels(data.width$method)<-c("Ro3","Wilson","Wilson Quantile")

View(data.width)

summarise(group_by(data.width,n),p=sum(width<hyp.width)/iter,m=mean(width))

write.csv(data.width,file=paste0("n_est_result_p",hyp.p,"_w_",hyp.width,"_ite
r_",iter,if(use.wils){"w"},".csv"))
data.width<-
read.csv(file=paste0("n_est_result_p",hyp.p,"_w_",hyp.width,"_iter_",iter,if(
use.wils){"w"},".csv"))

data.width %>%
  #filter(method=="Ro3") %>%
  #filter(method=="Wilson") %>%
  #filter(method=="Wilson Quantile") %>%
  ggplot(aes(x=method,y=width)) +
  geom_dotplot(binaxis="y",stackdir =
"center",dotsize=.1,binwidth=hyp.width/(iter/10)) + #Requires tweaking to
look nice
  #scale_y_log10(breaks=c(0.001*(1:10),0.01*(1:10))) + #Can't decide if this
should really be on a log scale...
  scale_y_continuous(breaks=seq(0,1,0.002)) + #...or a linear scale
  geom_hline(aes(yintercept=hyp.width),color="red") +

annotate(geom="text",x=c(1.38,2.33,3.3),y=c(0.01,0.0064,0.0056),label="X=0")
+ #For all three
  #annotate(geom="text",x=1.33,y=0.0101,label="X=22") + #for just one
  labs(title=paste("p =",hyp.p,"and nominal w =",hyp.width,"\n30k
iterations"),x="Estimation method used",y="Width of resulting CI")




#Now we'll test over a series of p and w
```

```
#We won't keep every sample this time, just the overall resulting probability
#of a width lower that what we want

#let's define some p's and width's to test over (roughly log spaced 10, 5,
and 3 per power of 10)
#10^-(10:40/10)
#10^-(5:20/5)
#10^-(3:12/3)

p.check<-
c(0.10,0.079,0.063,0.050,0.040,0.032,0.025,0.020,0.016,0.013,0.010,0.0079,0.0
063,0.0050,0.0040,0.0032,0.0025,0.0020,0.0016,0.0013,0.001,0.00079,0.00063,0.
00050,0.00040,0.00032,0.00025,0.00020,0.00016,0.00013,0.0001) #31 long
#p.check<-
c(0.10,0.063,0.040,0.025,0.016,0.010,0.0063,0.0040,0.0025,0.0016,0.001) #11
long
#p.check<-c(0.10,0.046,0.022,0.01,0.0046,0.0022,0.001) #7 long

w.check<-
c(0.10,0.079,0.063,0.050,0.040,0.032,0.025,0.020,0.016,0.013,0.010,0.0079,0.0
063,0.0050,0.0040,0.0032,0.0025,0.0020,0.0016,0.0013,0.001,0.00079,0.00063,0.
00050,0.00040,0.00032,0.00025,0.00020,0.00016,0.00013,0.0001) #31 long
#w.check<-
c(0.10,0.063,0.040,0.025,0.016,0.010,0.0063,0.0040,0.0025,0.0016,0.001) #11
long
#w.check<-c(0.10,0.046,0.022,0.01,0.0046,0.0022,0.001) #7 long


iter<-30000

data.width.general<-
data.frame(p=c(),w=c(),n=c(),method=c(),ave.width=c(),perc.lower=c())

set.seed(20542)

start<-proc.time()

for(p in p.check)
{
  for(w in w.check)
  {
    n.est<-c(n.est.Ro3(alpha.level,w),
             n.est.wilsons(alpha.level,w,p),
             n.est.iterated(alpha.level,w,p,0.8,3))

    method<-1
    for(n in ceiling(n.est))
    {
      samp<-rbinom(iter,n,p)

      sum.width<-0
      count<-0
      for(i in 1:iter)
      {
        interval<-ad.hoc.w(alpha.level,n,samp[i])
        sum.width<-sum.width+interval[2]-interval[1]
        if(interval[2]-interval[1]<=w)
```

```
          count<-count+1
      }
      data.width.general<-
rbind(data.width.general,c(p,w,n,method,sum.width/iter,count/iter))
      method<-method+1
    }
  }
  print(paste(p,(proc.time()-start)[3]))
}

proc.time()-start

colnames(data.width.general)<-c("p","w","n","method","ave.width","capture")

data.width.general$method<-as.factor(data.width.general$method)

levels(data.width.general$method)<-c("Ro3","Wilson","Wilson Quantile")

#View(data.width.general)

write.csv(data.width.general,file=paste0("ad_hoc_widths_Wils_",iter,".csv"))
data.width.general<-read.csv(file=paste0("ad_hoc_widths_",iter,".csv"))

data.width.general %>%
  ggplot(aes(x=method,y=capture)) +
  geom_boxplot() +
  scale_y_continuous(breaks=c(0:10/10))

data.width.general %>%
  #filter(method!="1"Ro3") %>%
  ggplot(aes(x=p/w,y=capture,color=as.factor(method))) +
  geom_point() +
  scale_y_continuous(breaks=c(0:10/10)) +

scale_x_log10(breaks=c(0.001,0.01,.1,1,10,100,1000),labels=c("0.001","0.01","
0.1","1","10","100","1000")) +
  geom_hline(aes(yintercept=0.8),color="red") +
  labs(x="P/W",
       y="Likelihood of being below width threshold",
       title="Likelihood of sufficiently narrow CI by\nSample size estimation
method and P/W",
       color="Sample Size\nEstimation Method")
```