**Confidence intervals for the parameters of the negative hypergeometric distribution**

Rachel Roggenkemper

**Introduction**

The Negative Hypergeometric Distribution (NHG) is a discrete probability distribution that is useful when sampling from a finite population without replacement, specifically where the population can be split up into two mutually exclusive groups (successes and failures).

A simple example involves a bag that contains a total number of *N* marbles, where there are *M* red marbles (which we’ll consider successes) and *X = N - M* blue marbles (which we’ll consider failures). We will conduct a random sample from the bag of marbles and keep track of how many red and blue marbles are in the sample. Our goal is to estimate the number of red marbles in the bag when *N* is known, but *M*,and therefore *X*, are unknown; or to estimate the total number of marbles, *N*, when the total number of red marbles, *M,* is known. For this distribution, the sampling will continue without replacement until a fixed number of successes, *m*, is observed, where . Let denote the number of blue marbles that are drawn to get *m* red marbles. Note that since we stop sampling until the marble is chosen, some combination of red marbles and blue marbles will be chosen in the first draws, and the last marble chosen will always be a red marble.

The probability that is:

which simplifies to:

**Notation**

Before delving into applied examples, it is essential to establish the notation and definitions that will be used throughout this paper. The following symbols represent the key parameters and statistics of the negative hypergeometric distribution:

: Total number of items where

: Number of successes in the population where

: Number of failures in the population where

: Fixed number of successes to be observed where

: Number of failures observed before the success where

*:* Number of items being sampledwhere

**Point Estimates**

Before constructing confidence intervals for the parameters, it is important to first compute point estimates. These point estimates serve as initial approximations for the unknown parameter values and provide a foundation for the subsequent interval estimation methods.

Maximum likelihood estimator (MLE) for *M*:

is the greatest integer such that

Unbiased estimator for *M*:

Point Estimate for N:

**Applied Example**

Confidence intervals may be desired for two situations: (i) for the unknown parameter *N* if *M* is known; and (ii) for the unknown parameter *M* if *N* is known.

**(i) Example when estimating *N* if *M* is known**

Imagine a conservationist is tasked with estimating the population size of a rare species of penguin within a nature reserve in Antarctica. The reserve is large, and the penguins are scattered unevenly throughout, making a full census impractical. To estimate the population size, the conservationist decides to use a sampling method that ensures a sufficient number of encounters with the rare species to make a reliable estimate.

In a typical capture-recapture scenario, the conservationist would capture a random sample of penguins, mark them, release them, and then capture another sample to see how many marked penguins are recaptured. However, this method does not guarantee that any marked penguins will be recaptured if they are very sparse, leading to possible inaccuracies in the population estimate.

Instead, the conservationist uses a negative hypergeometric sampling approach, where they continue to capture penguins until they have recaptured a fixed number of penguins. This approach ensures that each sampling effort results in a sufficient number of successes (recaptured marked penguins), which is vital for accurate and reliable population estimates.

In this context, confidence intervals for parameters of the negative hypergeometric distribution are crucial because they provide a range of plausible values for the parameters like the total number of failures before achieving a fixed number of successes. These intervals help in understanding the variability and reliability of the estimates under conditions where a fixed number of successes is guaranteed. These intervals are particularly important in planning and resource allocation in ecological and public health research, ensuring that decisions are based on data that includes a predefined number of successes, therefore increasing the reliability of the outcomes.

For instance, suppose the conservationist decides to stop the capture-recapture process after recapturing 10 marked penguins. During this process, they note 30 unmarked penguins before achieving this goal. Using negative hypergeometric, they can construct a confidence interval to estimate the range of the number of unmarked penguins in the total population, thereby gaining insights about the overall biodiversity and population density of the nature reserve.

**(ii) Example when estimating *M* if *N* is known**

Imagine a public health official at California Polytechnic State University, San Luis Obispo is tasked with estimating the number of students who have contracted COVID-19. The total student population is known, but the number who have had the virus is unknown. Traditional survey methods might not yield accurate results due to possible non-response or inaccurate self-reporting by students. To address this challenge, the official uses a negative hypergeometric sampling approach, ensuring a more reliable estimate by focusing on a specific number of confirmed cases.

The official decides to conduct a survey that continues until a pre-determined number of COVID-19 cases, , is reported, rather than randomly polling a fixed number of students. This method avoids the potential shortfall of not gathering enough case data. This approach guarantees that each sampling effort results in a sufficient number of successes (in this case COVID-19 cases), which is crucial for making accurate and reliable population estimates. The survey only stops after cases are collected, at which point the data gathered will allow the official to construct a confidence interval for . These intervals are incredibly useful as they provide a range of plausible values for the number of students who have had COVID-19, which is crucial information for university health planning and resource allocation and for ongoing or future health challenges.

We will now apply this methodology to real data. In January 2022, the total student population at Cal Poly is 21,778. It is known that 1,293 students have tested positive since October 20, 2021, but this number is treated as unknown in the estimation process. The goal is to estimate the total number of students who have tested positive for COVID-19 () by continuing to survey until 20 positives are found. During the data collection, a random sampling process was conducted to simulate the selection of students until 20 positive COVID-19 cases were identified. Throughout the process, the total number of students sampled and the number of negatives observed before reaching 20 positives were recorded. The results of the sampling showed that the total number of samples taken to get 20 positives was 468 , with 448 negatives observed before obtaining the 20 positives. Using the observed data (), the analog to the Clopper-Pearson method was applied to calculate the confidence interval for . Debany (2023) applied this well-known reliable method to the negative hypergeometric distribution, and the methodology will be introduced in detail shortly. The calculated 95% confidence interval for was found to be [577 , 1,362]. Thus, we are 95% confident that the number of Cal Poly students who have had COVID-19 is between 577 and 1,362. Although in practice we do not know the true value of , in this scenario, we do know that = 1,293, which falls in our confidence interval.

**Methods**

In this section, we present various methods used to construct confidence intervals for . Beginning with large sample procedures that use a normal approximation. We will discover that these methods provide poor coverage, especially with small sample sizes. Thus, we turn to methods with strict coverage, starting with the analog to the Clopper-Pearson method. However, while this approach ensures coverage, it tends to be overly conservative and does not minimize cardinality. Therefore, we explore multiple minimal cardinality procedures, in addition to Blaker’s Method and the Conditional Minimal Cardinality (CMC) method.

**Approximate Confidence Intervals (Zhang & Johnson)**

Zhang & Johnson (2011) proposed using large sample approximations to approach providing confidence intervals for the negative hypergeometric distribution when estimated the number of successes, , assuming the total population size, , is known. To do this, they investigated using the maximum likelihood estimator and an unbiased estimator. They also used the method of Taylor’s series expansions to find approximations for the variance.

**Maximum Likelihood Estimator for**

One approach to approximate confidence intervals is to use the MLE.

The maximum likelihood estimator (MLE) for is:

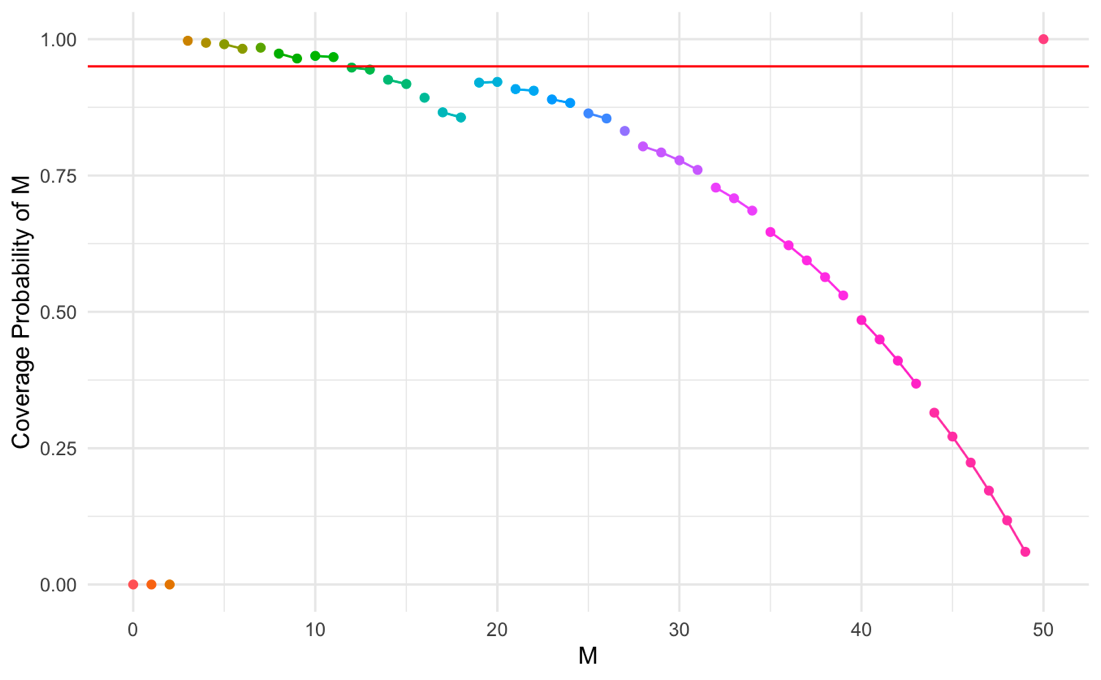
Using the Taylor series method to find an estimator for the standard error of the MLE for gives the following:

Putting the MLE and standard error together, a 100(1 - )% confidence interval for the MLE of is:

where represents the critical value from the standard normal distribution corresponding to a cumulative probability of .

With this confidence interval equation, we examine the coverage probability function.

Figure 1 illustrates the coverage probability function using Zhang’s approximate confidence interval when approximating with the MLE. However, it is evident that the performance of this method is very poor. Starting at = 13 trough 49, the coverage probability is always below the confidence level, which is 0.95, and continues to decrease as increases. The coverage probability even drops to 0.06 when = 49.



*Figure 1: CPF of NHG (Zhang - MLE) where N = 50, m = 3, Confidence Level = 0.95*

**Unbiased Estimator for**

Another approach to approximate confidence intervals is to use the unbiased estimator.

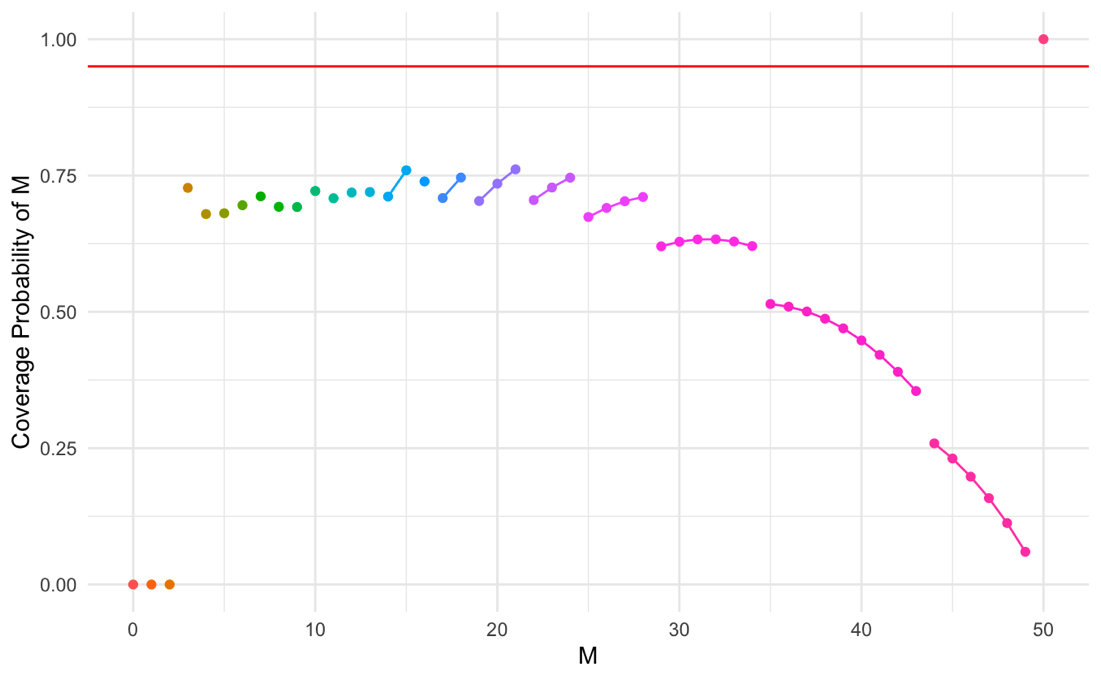
An unbiased estimator for is:

Using the Taylor series method to find an estimator for the standard error of the unbiased estimator for gives the following:

Putting the unbiased estimator and standard error together, a 100(1 - )% confidence interval for the unbiased estimator of is:

With this confidence interval equation, we calculate coverage probabilities and examine the coverage probability function.

Figure 2 illustrates the coverage probability function using Zhang’s approximate confidence interval when approximating with the unbiased estimator. However, it is evident that the performance of this method is very poor, even worse than the MLE performance. Using the unbiased estimator, the coverage probability is always below the confidence level except at = 50. At every other in the plot, with a confidence level of 0.95, the coverage probability is rarely above 0.75, and often far below 0.75. However, the fact that the unbiased estimator is performing worse in terms of coverage probability than the MLE is not surprising. Zhang notes that although the confidence intervals based on the unbiased estimator tend to be narrower, this comes at a cost in terms of coverage probability.



*Figure 2: CPF of NHG (Zhang - Unbiased) where N = 50, m = 3, Confidence Level = 0.95*

**Analog to the Clopper-Pearson Method**

Since the approximate methods have poor coverage, we will now consider methods with strict coverage. Clopper & Pearson (1934) proposed the Clopper-Pearson method, which is a well-established approach for calculating exact confidence intervals for the binomial distribution. This method is known for its robustness and reliability, making it a foundational building block. Drawing inspiration from the Clopper-Pearson method, we propose an analogous approach for constructing confidence intervals for parameters of the negative hypergeometric distribution. This method provides a way to estimate the total numbers of successes () or the total population size ().

The proposed method, which we refer to as the “Analog to the Clopper-Pearson Method,” involved calculating the probability mass function (pmf) for the negative hypergeometric distribution and determining bounds such that the cumulative probabilities are close to the desired confidence level.

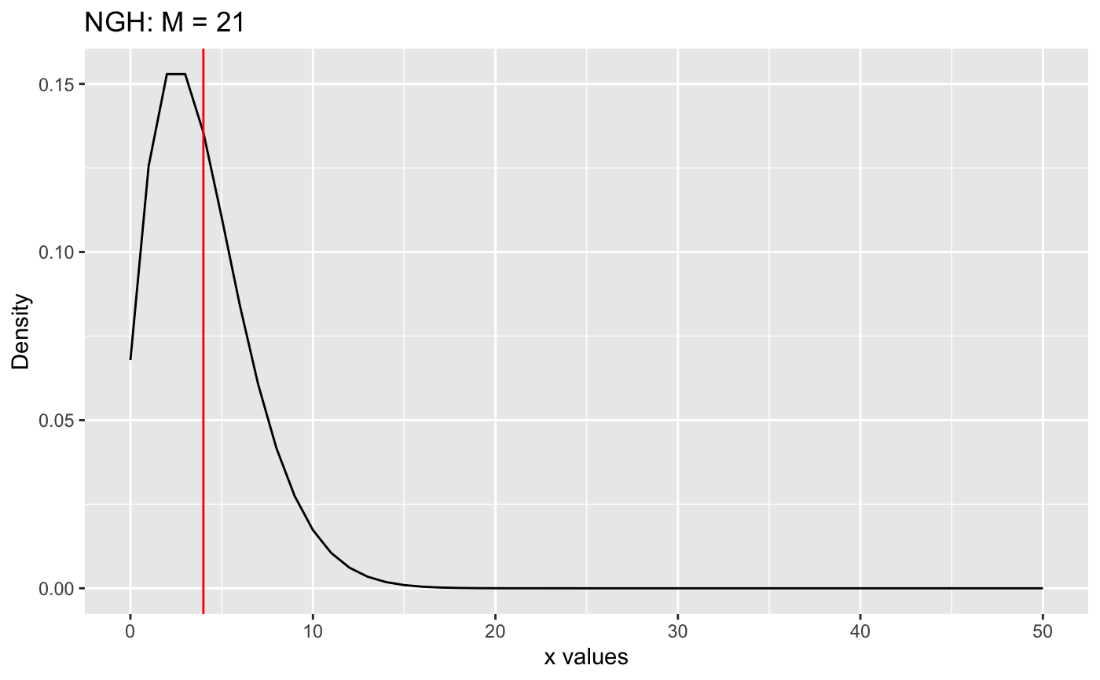
To determine the confidence interval for , we need to calculate the lower and upper bounds. The idea behind it is for a fixed and , we find the values of , where for the distribution with that , the value of is unlikely to occur. Thus, the that are found as lower and upper bounds represent the value has to be to make the chosen value of unlikely to occur. This gives us all the values such that the is in the middle 100(1 - )% of the distribution. We do this on both ends to find the lower and upper confidence bounds. In other words, for a fixed , we push as many into the rejection region as long as the probability of either the upper or lower tail does not exceed . These become a part of the rejection region, and all the that are not in either tail are in the acceptance region.

To calculate the lower bound, we first start with the smallest possible value of (i.e. ). We then increment until the cumulative probability is just below (without going over) , where is the significance level. In contrast, to calculate the upper bound, we start with the largest possible value of (i.e. ) and decrement until the cumulative probability is just below (without going over) .

**Example of Finding a 95% Confidence Interval for using the Analog to the Clopper-Pearson Method:**

To illustrate the process of determining the confidence interval for using the Analog to the Clopper-Pearson Method, let’s consider an example where and

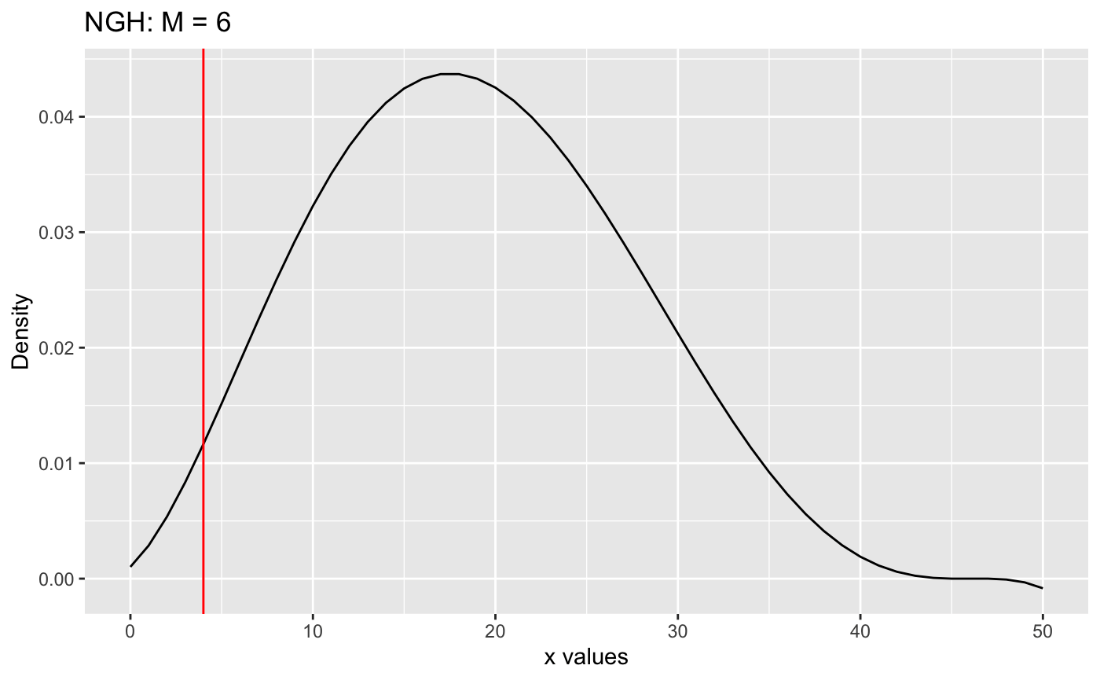
We start with an initial point estimate for . In this case, the point estimate is . Figure 3 shows the plot below the pmf for the NHG distribution where . The red vertical line indicates the observed number of failures ().



*Figure 3: PMF of NHG where M = 21*

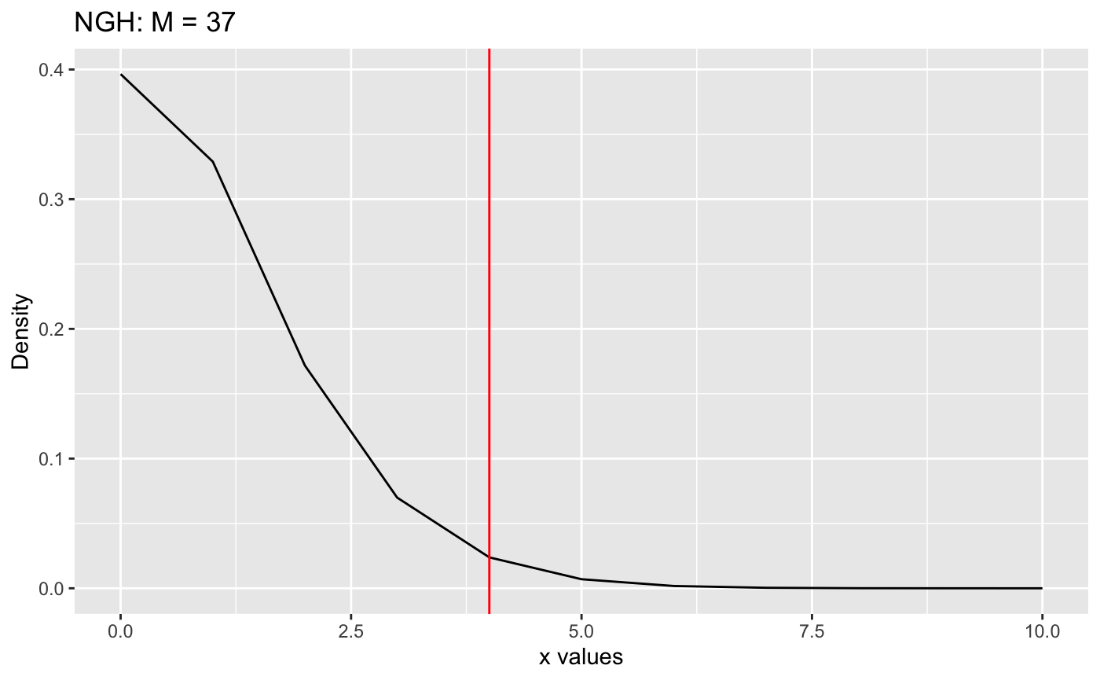
We can observe the distribution of the number of failures before reaching 3 (fixed) successes when there are 21 successes in a population of 50.

To find the lower bound, we start with the smallest possible value of (i.e. ). We increment and calculate the cumulative probability until we have reached the highest value of before the left tail probability exceeds . In this case, we find that the cumulative probability for is the highest value where the cumulative probability (equivalent to the area to the left of the red line) is still less than 0.025, specifically equaling 0.0292. Therefore, the lower bound is .



*Figure 4: PMF of NHG where M = 6*

To find the upper bound, we start with the largest possible value of (i.e. ). We decrement and calculate the cumulative probability until we have reached the smallest value of before the right tail probability exceeds . In this case, we find that the cumulative probability for is the lowest value where the cumulative probability (equivalent to the area to the right of the red line) is still less than 0.025, specifically equaling 0. 0.0331. Therefore, the upper bound is .



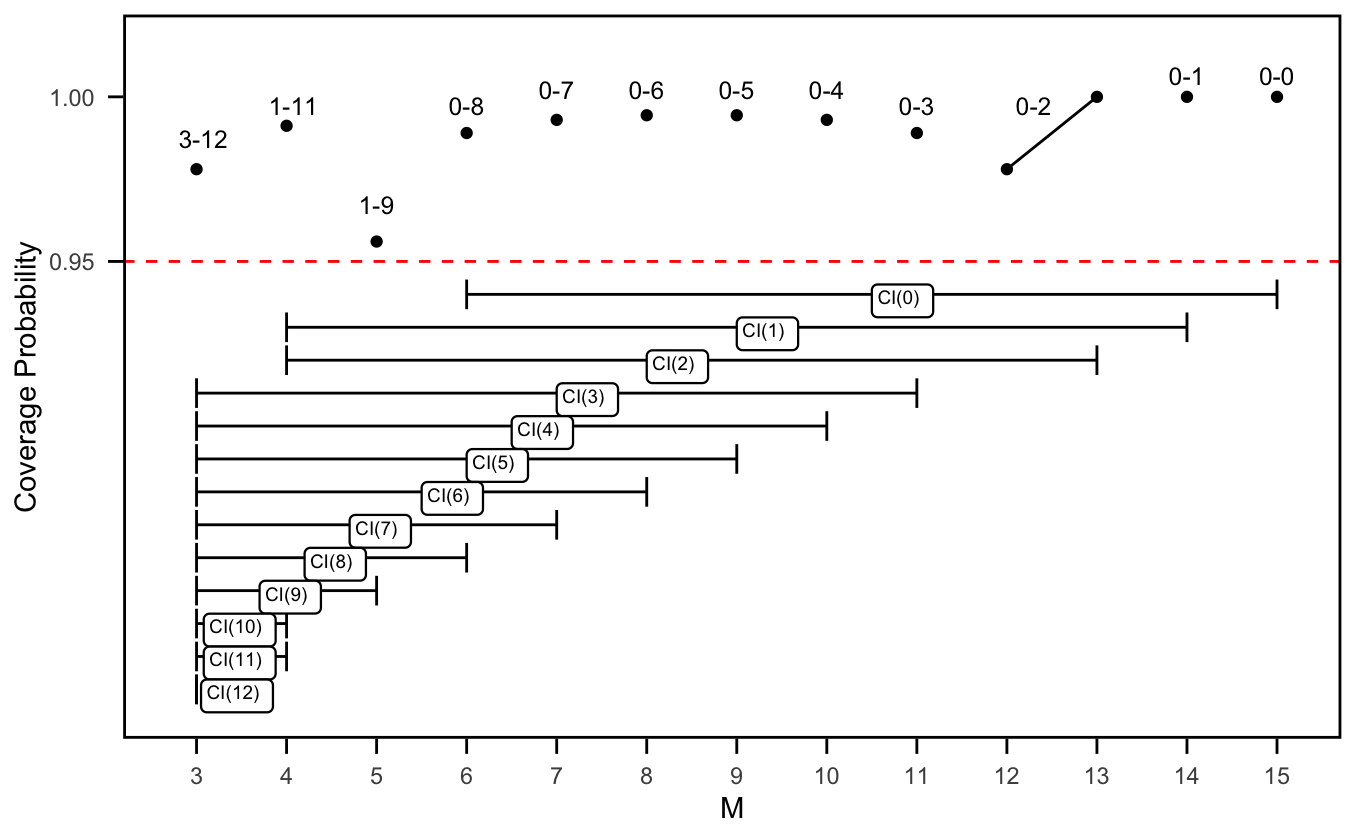
*Figure 5: PMF of NHG where M = 37*

Based on the Analog to the Clopper-Pearson Method, the 95% confidence interval for given and is [6, 37]. This interval provides a range of plausible values of the total number of successes in the population when the total population size is 50, and there are 3 fixed successes that are sampled and 4 sampled failures that are needed to reach the 3 fixed number of successes.

**Interplay Between CPF and Resulting Confidence Intervals**

Figure 6 displays the relationship between the graphical properties of a CPF and the resulting confidence intervals, showing how one can work backwards and reverse engineer confidence intervals from a given CPF. To do this, for an observed x, find the first and last place acceptance sets in the CPF that observed x is present, and those are the resulting lower and upper confidence bounds. To obtain all confidence intervals, repeat this step for each x.

For example, to find the confidence interval for x = 1, we must find the first and last place acceptance sets in the CPF that include 1. So, starting from the right, we see that the first place the CPF uses such a set is at M = 14, and the last place is at M = 4. Thus, the confidence interval for x = 1 is [4, 14].



*Figure 6: Interplay between 95% NHG CPF and resulting confidence intervals when N = 15, m = 3*

**Minimal Cardinality Procedures**

When constructing confidence intervals, the primary focus is minimizing interval length while still achieving a coverage probability greater than or equal to the confidence level. With the two above methods, the Approximate Method has the issue of undercoverage, where the coverage probability is often far below the confidence level. With the Analog to the Clopper-Pearson Method, although there is not an issue of undercoverage, one might even say that there is overcoverage. The intervals often have coverage far above the confidence level, leading to excessively wide intervals which lowers precision.

Thus, we turn to minimal cardinality procedures where we reverse engineer our confidence intervals, guaranteeing coverage that is above the confidence level but as small as possible, while providing minimal cardinality regarding the cardinality of acceptance sets, therefore narrower intervals. The procedure can be contained in three main steps, and can be summarized as follows:

1. Find all minimal cardinality acceptance curves

For each value of , starting from the total population size and decrementing to 0, we calculate all possible acceptance curves. This is broken into two cases, since special consideration is given when , the total number of successes in the population, is less than the fixed number of successes, . In this case, the bounds of the acceptance curves, and , are both set to .

For other values of (i.e. ), we then loop through each value of starting at and ending with . For each value of , we iterate through all potential values of and . Holladay (2019) proves that and must be non-decreasing to obtain gapless intervals. To ensure that and are both non-decreasing so gaps cannot occur, we keep track of the previous values of and , and make sure to that the iteration for only begins at its current minimum so it cannot be less than any previous values of . The iterations for a continue up to . Then, for each , the iteration for starts at the maximum for and its current minimum, to ensure that and that is also non-decreasing, and continues up to . For each pair of and , the corresponding coverage probability and cardinality is recorded.

To ensure that the acceptance curves are valid, we filter the results to include only those with a coverage probability greater than or equal to the specified confidence level. Additionally, for each value of , we keep only the acceptance curves with the smallest cardinality.

This approach results in all minimal cardinality acceptance curves and guarantees that the acceptance curves are as precise as possible while maintaining the desired coverage probability.

1. Choose acceptance curves based on the desired procedure

After finding all minimal cardinality acceptance curves, each has at least one corresponding acceptance curve. To choose which acceptance curve to use depends on the method. So, we iterate through each to choose an acceptance curve. When there is only one acceptance curve, then regardless of method, we must choose that acceptance curve. However, if there is more than one acceptance curve for that , these are coincidental endpoints and then it depends on the method which one will be chosen. It should be noted that similar to above, when iterating through each , to ensure there are no gaps, we must make sure and are both non-decreasing, so in addition to choosing the acceptance curve based on procedure, we also ensure that every acceptance curve that is chosen so that and are non-decreasing.

**Modified Sterne Method (MST)**

When there is more than one minimal cardinality acceptance curve for a chosen , using the Modified Sterne Method, the acceptance curve with the highest coverage probability is chosen. This method is adapted from Sterne (1954). The modification made to Sterne’s original approach is to ensure there are no gaps by ensuring and are both non-decreasing. This procedure that modifies Stene’s original procedure is from the Schilling & Doi (2014) approach, and is named Modified Sterne Method (MST).

Should I mention the case when coverage is the same? For example: N = 20, m = 3, CL = 0.95: M = 5, 2-14 and 1-13

To visualize this, Figure 6 displays all 95% minimal cardinality acceptance curves when = 50 and = 3. Coincidental endpoints occur at = 14, 13, 12, 10, 9, 7, 6, and 4. Since MST chooses coincidental endpoints to maximize coverage, when looking at = 10 specifically, acceptance set 1-22 would be chosen over 2-23 because 1-22 has higher coverage. This process is repeated for all coincidental endpoints.

**Crow & Gardner Method (CG)**

When there is more than one minimal cardinality acceptance curve for a chosen , using the Crow & Gardner (1959) method, the transition must be as early as possible, thus the acceptance curve with the largest possible and is chosen. Note that this is flipped compared to applications with Poisson and Binomial because in those cases, the transitions are from left to right, however with the NHG, we are transitioning from right to left. Thus, when transitioning as early as possible, this results in the largest possible , while with Poisson and Binomial, it results in the smallest possible .

For example, looking at the example in Figure 6, there is a coincidental endpoint at = 7, where there are 3 possible acceptance sets: 2-30, 3-31, and 4-32. Applying CG to the NHG when estimating , to transition as early as possible, CG chooses coincidental endpoints to be the largest possible value. Thus, at = 7, 4-32 would be the acceptance set that is chosen. This process is then repeated for all coincidental endpoints.

**Bryne and Kabaila Method (BK)**

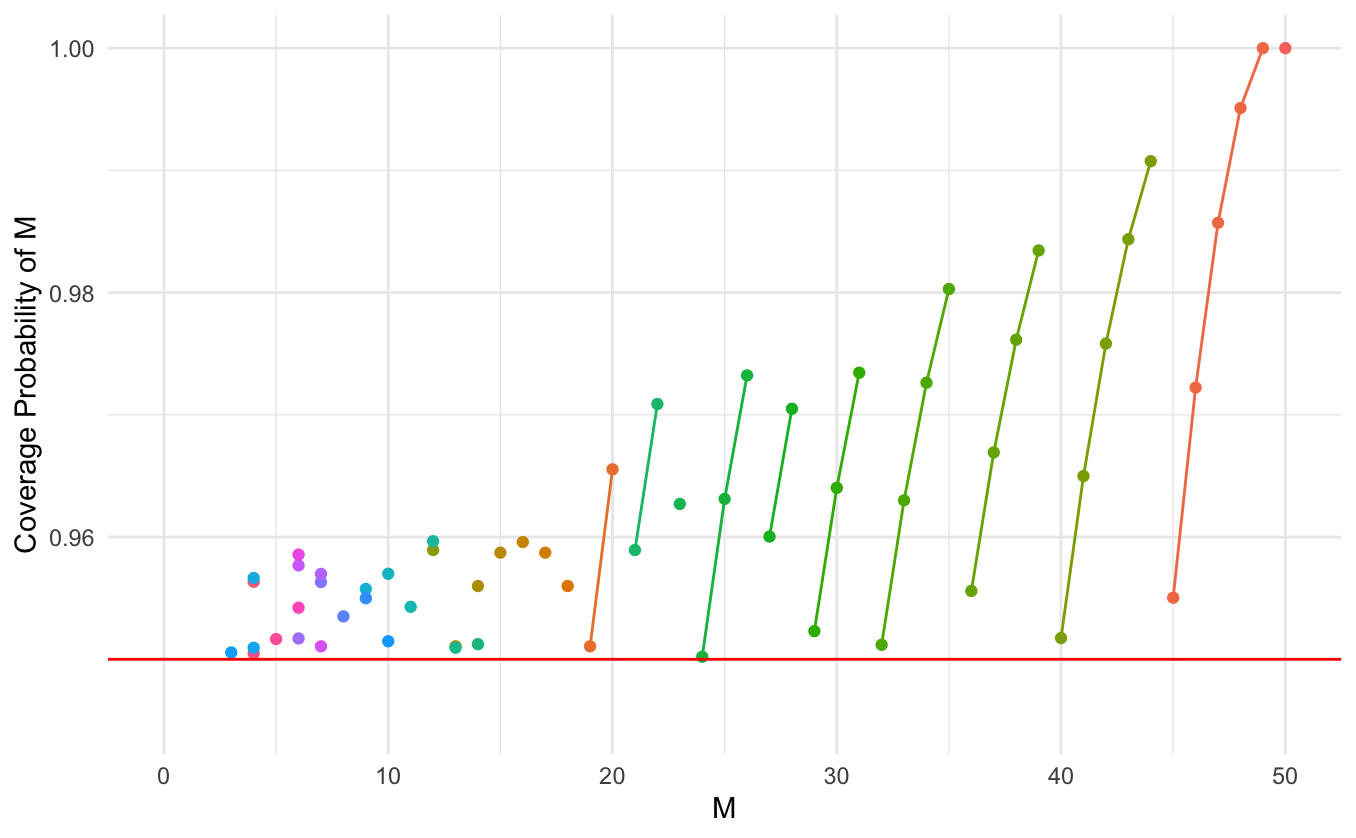
When there is more than one minimal cardinality acceptance curve for a chosen , using the Bryne and Kabaila (2005) method, the transition must be as late as possible, thus the acceptance curve with the smallest possible and is chosen. This approach is the stark opposite of CG. Note that this is flipped compared to applications with Poisson and Binomial because in those cases, the transitions are from left to right, however with the NHG, we are transitioning from right to left. Thus, when transitioning as late as possible, this results in the smallest possible , while with Poisson and Binomial, it results in the largest possible .

An example of the application of BK can be using Figure 7. There is a coincidental endpoint at = 6, where there are 4 possible acceptance sets: 2-33, 3-34, 4-36, and 5-36. When applying BK to the NHG when estimating , to transition as late as possible, BK chooses coincidental endpoints to be the smallest possible value. Therefore, at = 6, 2-33 would be the acceptance set that is chosen. This process is then repeated for all coincidental endpoints.

1. Find confidence interval bounds using acceptance curves

Now, for each we have acceptance curves. To find the confidence interval bounds using these acceptance curves, for each possible value of , we determine the bounds of the confidence intervals by identifying the first and last acceptance curves where appears. Specifically, we find the first and last occurrences of acceptance curves that include .

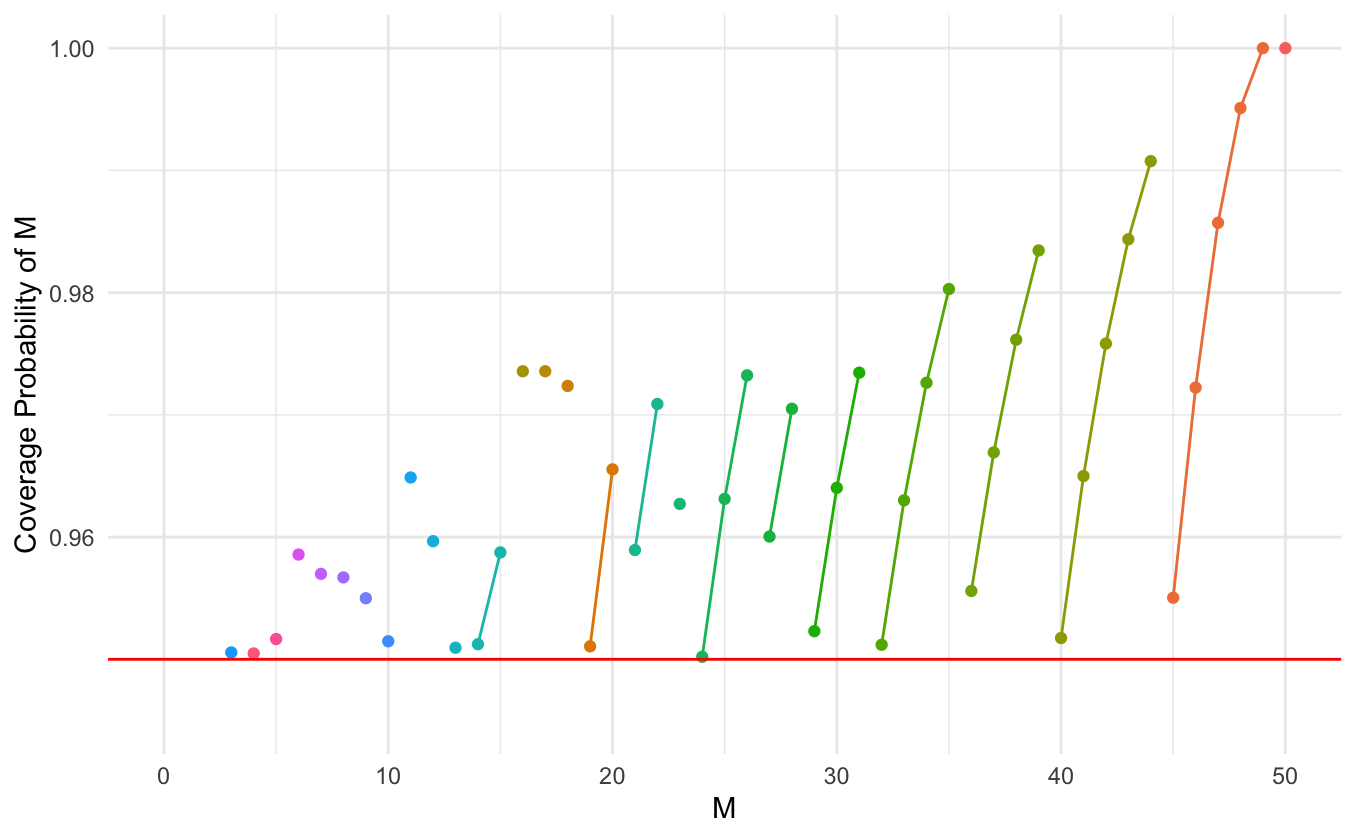
The lower bound of the confidence interval is the value of from the last occurrence acceptance curve, and the upper bound of the confidence interval is the value of from the first occurrence acceptance curve.

*Figure 7: All 95% Minimal Cardinality Acceptance Curves when N = 50, m = 3*

**Blaker’s Method**

Blaker’s (2000) approach to constructing confidence intervals opens the possibility of asymmetric rejection region tail probabilities. Blaker’s Method is as follows: iterate through each and , fix and . Define the min-tail probability of a fixed value of to be the minimum of and . Then find all the that have a min-tail probability as small as that , including the fixed . Then, sum up the probability of observing these . If the sum of these probabilities is greater than , which is defined as 1 – confidence level, include the fixed in the acceptance set for the fixed . Once we have done this for all M, we have the acceptance sets for all , and like above, can work backwards to find the confidence intervals.

The 95% CPF for Blaker’s Method for the NHG() is shown in Figure 8. Blaker’s method performs well in regard to coverage probability. The coverage probability is always above the confidence level. However, it is important to note that Blaker’s Method does not possess the minimal cardinality property.

*Figure 8: CPF of Blaker’s Method where , Confidence Level = 0.95*

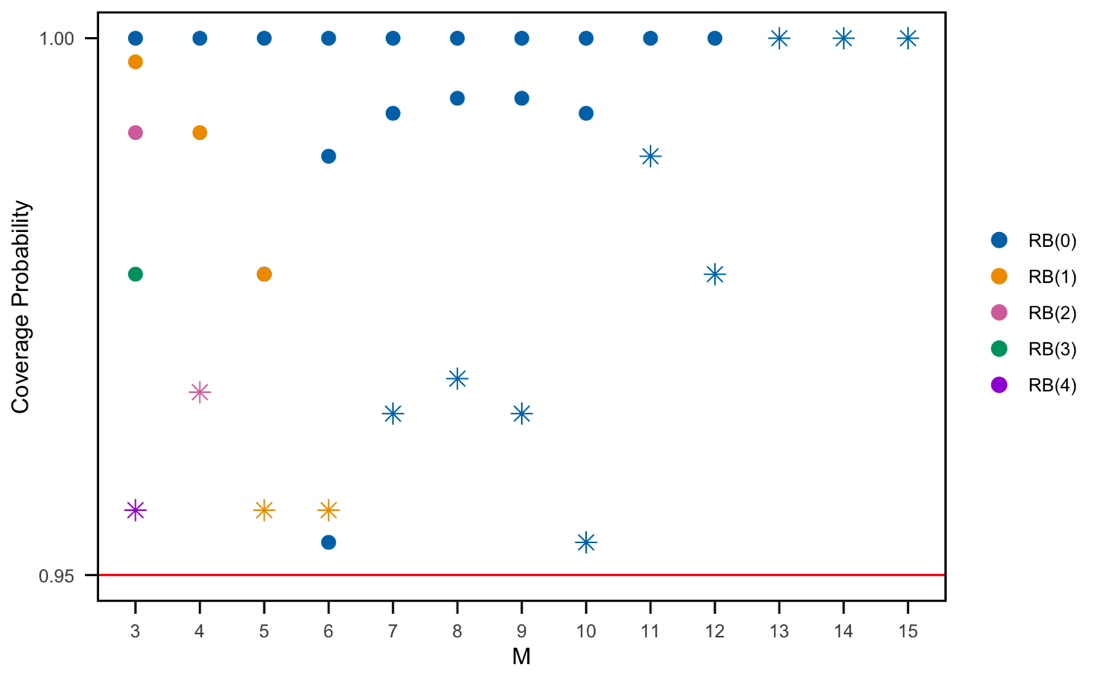
**Conditional Minimal Cardinality (CMC)**

The CMC method is a newly introduced approach from Schilling, Holladay, & Doi (2023), originally developed for constructing intervals for negative binormal parameters. This method leverages a coverage probability approach. Given its success in the negative binomial context, we extend the CMC method to the negative hypergeometric distribution.

The CMC procedure from a CPF perspective is as follows: For each starting from the right and working to the left, use only acceptance curves with = 0, and switch to using = 1 as soon as there is an acceptance curve with = 1 that has coverage higher than the confidence level. Then for each , choose the lowest possible curve that has coverage higher than the confidence level. Continue this process for all .

Note that we are working from right to left, versus left to right like Schilling, Holladay, & Doi (2023) due to the nature of the negative hypergeometric distribution when M is unknown. Additionally, unlike Schilling, Holladay, & Doi (2023), the rainbows are not quite curves in this case because is integer valued, so they are dots instead.

Figure 9 illustrates how the cpf of the CMC method transitions between rainbows when = 15, = 3, and confidence level = 0.95. Starting from the right to the left, for cases = 13 to = 15, there is only one possible rainbow curve so that is the rainbow that is chosen for the cpf for those ’s. For = 7 to = 12, we can see for each , there are multiple points but they all belong to the same rainbow, so the inner-most or lowest possible point is chosen. For = 3 to = 6, there are now multiple points belonging to different rainbows. Thus, for these ’s, we choose the highest possible rainbow, and if there are multiple points for that given rainbow, we chose the lowest possible point within that specific rainbow.



*Figure 9: Sample Rainbow Plots for N = 15, m = 3, Confidence Level = 0.95. Stars represent points of the rainbows belonging to the cpf.*

In practice, here is how the CMC approach was applied to the negative hypergeometric distribution when is unknown. First find all possible acceptance curves that have non-decreasing and . Then, loop through each and filter within the loop. So, for each , starting from the right (when = ), after finding all possible acceptance curves for that , first filter out all acceptance curves with coverage probability below the confidence level. So now for that , we have all non-decreasing / acceptance curves with coverage above the confidence level. Then, pick the acceptance curve with the highest , which is doing the step of choosing the next rainbow once it rises above the confidence level. Then, to choose the core / inner most curve / lowest possible curve or dot in this instance, choose the acceptance curve with the lowest . This is the same as doing the lowest curve because a is already fixed, so choosing the acceptance curve with the lowest is the same as choosing the lowest possible curve/dot. Then repeat this step for all , making sure that it is still forcing and to be non-decreasing.

**Length and Coverage Comparison**

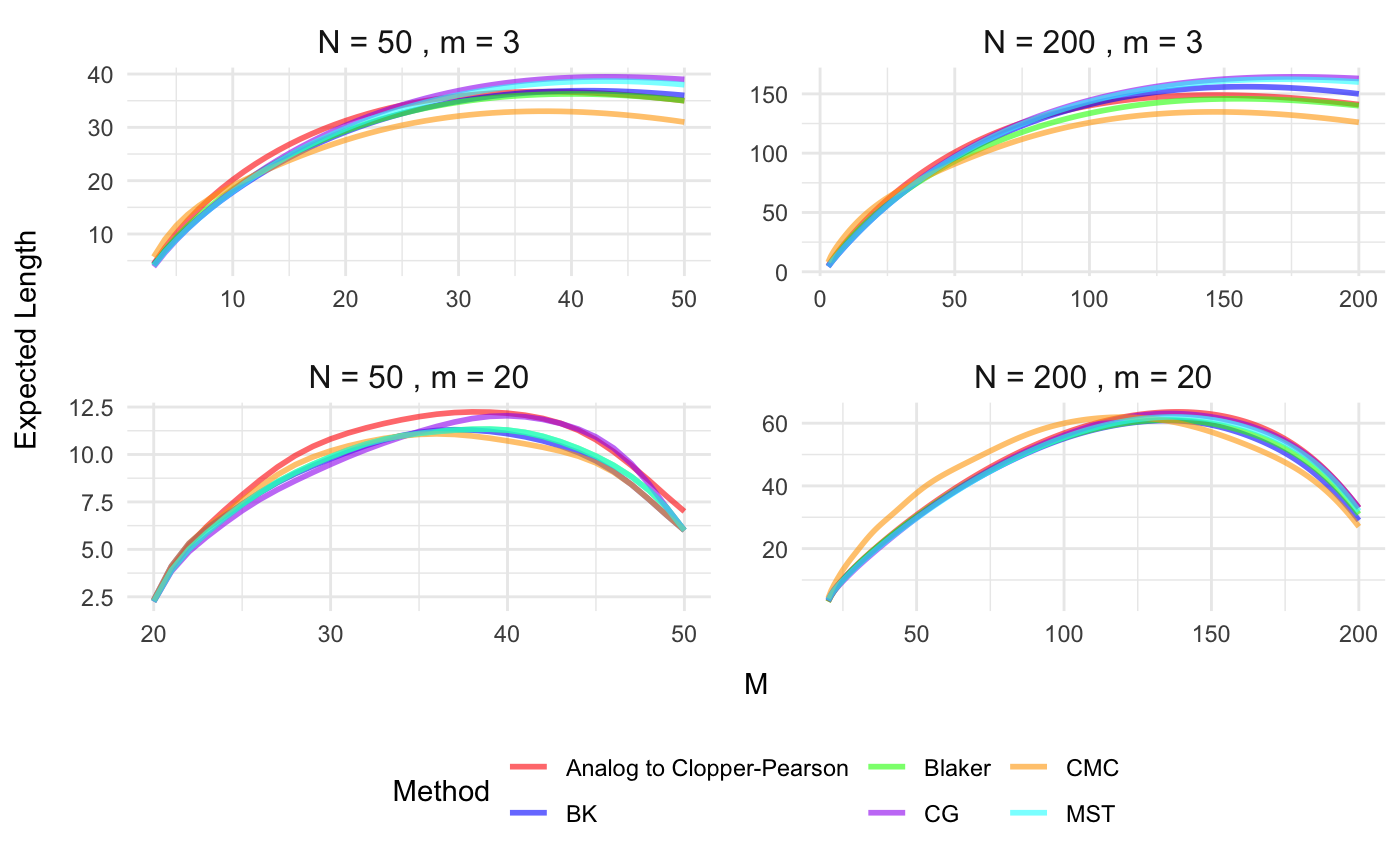
In order to assess the performance of the various confidence interval methods, we will do a detailed comparison, focusing on expected length, average length, and coverage probability of the intervals. This comparison is crucial to understand the trade-offs between the methods, particularly in terms on interval length and coverage probability. For a thorough analysis, we tested across combinations of different sample sizes ( = 50 and = 200) and fixed numbers of successes ( = 3 and = 20).

One way to compare length is using the expected length which is the expected value of the length. To calculate this, we use Law of the Unconscious Statistician (LOTUS) and find that:

Figure 10 displays the expected length of the confidence intervals as a function of for each method, excluding the two normal approximation methods because although the length of their intervals is often shorter than the other methods, they don’t deliver the coverage promised and often have coverage below the confidence level, so those two methods have been excluded from this plot.

The plot reveals several important trends. Across all test cases and methods, the expected length of the intervals increases as approaches , reflecting the growing uncertainty in estimation as the number of successes in the population nears the total population size. However, when = 20, the expected length starts to decrease once nears , while it plateaus when = 3.

Comparing the methods, all methods behave similar. It appears that the CMC method almost consistently produces intervals with the shortest expected length, especially when = 3. CMC does appear to perform the worst when = 200 and = 20 for ’s less than 100.



*Figure10: Plot of expected confidence interval length*

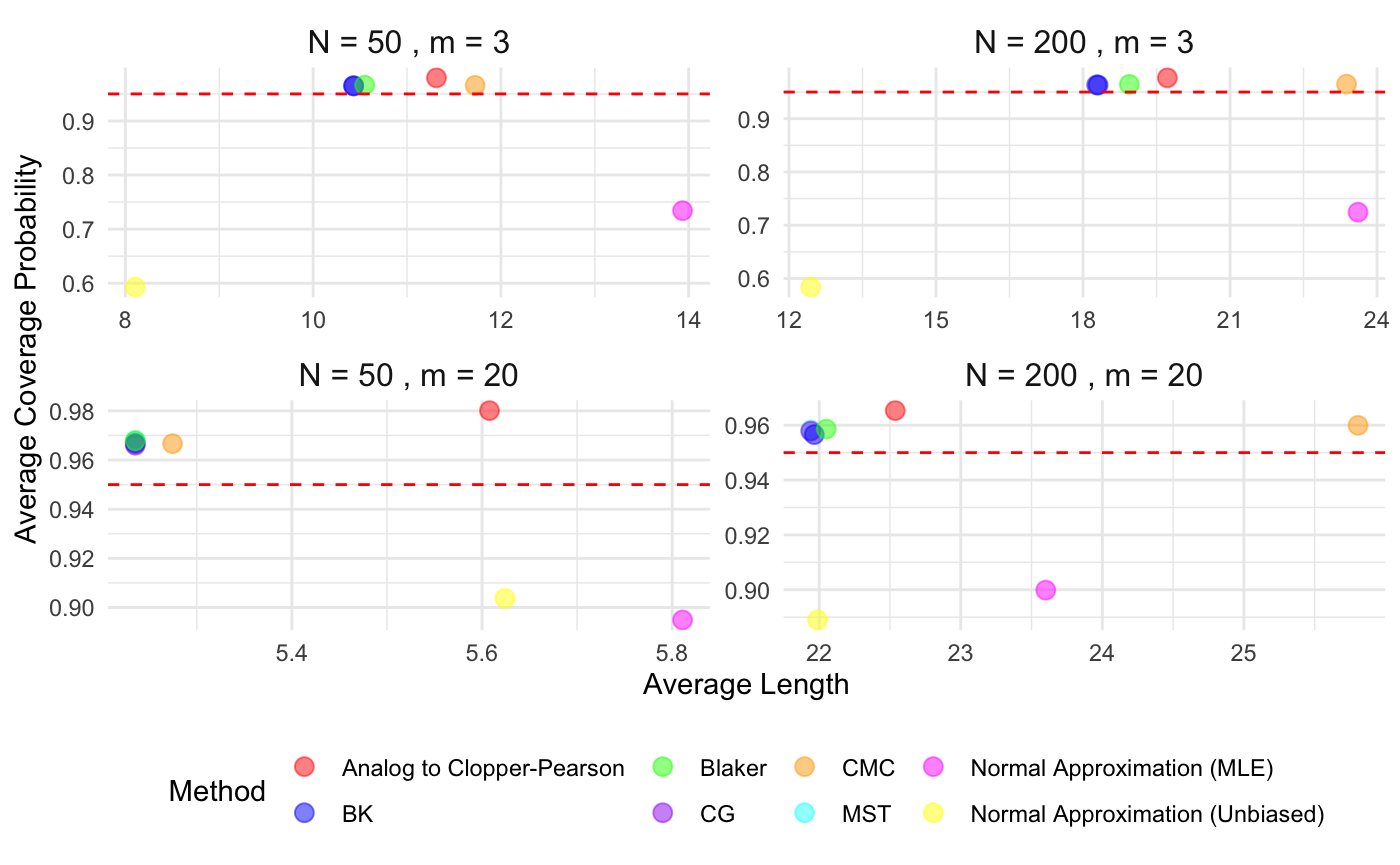
Another way to compare length of intervals is to compare average length. To calculate this, for any combination of N and m, calculate intervals for all and take the average of all these intervals.

Figure 11 displays the difference between the confidence intervals methods by examining their average coverage probability versus the average length for the same four cases as Figure 10. This plot evaluates the efficiency and reliability of each method. Ideally, we want a method with shorter average lengths and high coverage probabilities.

The two normal approximation methods are included in this plot because we can see that their average coverage probability is far below the confidence level. For the unbiased method, we see that they do provide shorter intervals, but really lack in coverage. For the MLE method, they often provide the longest intervals with horrible coverage.

Note that the average length is the same for the minimal cardinality procedures due to how coincidental endpoints work which is proved in Holladay (2019). This is because they are choosing between acceptance sets of minimal cardinality (so the same minimal length). The only difference is between which minimal cardinality sets are being chosen, but the length of them is the same since they are minimal cardinality. Thus, the average length is the same for the minimal cardinality procedures. Therefore, Modified Sterne (MST) is considered superior because they all have the same average length, but MST chooses the acceptance sets with the highest coverage probability, so with the same length as the other minimal cardinality procedures, MST is getting the most coverage. Since the average length for the three minimal cardinality procedures are the same and the average coverage probability is also very similar for each method, so they are stacked upon each other in the plots and difficult to differentiate between in the plot.

Overall, the minimal cardinality procedures perform the best in terms of average length and average coverage probability out of all the methods, and out of the three minimal cardinality procedures, MST performs the best in terms of coverage.



*Figure11: Plot of average confidence interval width versus mean coverage*

**Future Work**

My current research has centered on developing confidence intervals for the negative hypergeometric distribution with as the unknown parameter. To continue this, I would like to do a more comprehensive comparison of the MST and CMC procedures, using a larger collection of combinations of and pairs. In the future, I plan to extend this work to when , the total population size, is unknown. Additionally, I am currently developing a Shiny app to make these methods more accessible to practitioners. Beyond this, I would like to create an R package to further streamline the application of these techniques and aim to publish these findings in a peer-reviewed journal, contributing to the broader statistical community.

**Conclusion**

In this study, we have explored various methodologies for constructing confidence intervals for the negative hypergeometric distribution with the parameter being unknown. Through a comparative analysis of different methods, including the CMC, MST, and others, we evaluated their performance based on interval length and coverage probability. The results indicate that the CMC method generally provides shorter expected lengths, particularly in cases where m is small. However, when considering average length and coverage probability, the MST method consistently delivers intervals with shorter average lengths and higher coverage.

**References**

Blaker, Hans. "Confidence Curves and Improved Exact Confidence Intervals for Discrete Distributions." \*The Canadian Journal of Statistics\*, vol. 28, 2000, pp. 783–798, doi:10.2307/3315916.

Byrne, Jason, and Paul Kabaila. "Comparison of Poisson Confidence Intervals." \*Communications in Statistics — Theory and Methods\*, vol. 34, no. 3, 2005, pp. 545–556, doi:10.1081/STA-200052109.

Clopper, Caleb J., and Egon S. Pearson. "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial." \*Biometrika\*, vol. 26, no. 4, 1934, pp. 404–413.

Debany, Warren H. \*Improved Methods for Exact Confidence Intervals When Sampling With and Without Replacement for Rare Events\*. Air Force Research Laboratory, August 2023.

Holladay, Bret Andrew. \*Contributions to Interval Estimation for Parameters of Discrete Distributions\*. University of California, Santa Barbara, 2019, PhD dissertation.

Jones, Steven Norman. \*A Gaming Application of the Negative Hypergeometric Distribution\*. Master’s thesis, University of Nevada, Las Vegas, 2013, https://digitalscholarship.unlv.edu/thesesdissertations/1846.

Rashad, Omar, and Roselyn Romero. "Cal Poly Updates COVID-19 Dashboard to Show Number of COVID-Positive Students, Not Just Total Positive Tests." \*Mustang News\*, 19 Jan. 2022, https://mustangnews.net/cal-poly-updates-covid-19-dashboard-to-show-number-of-covid-positive-students-not-just-total-positive-tests/. Accessed 22 Sept. 2024.

Schilling, Mark F., and Alyssa Stanley. "A New Approach to Precise Interval Estimation for the Parameters of the Hypergeometric Distribution." \*Communications in Statistics - Theory and Methods\*, Taylor & Francis, 2020, https://doi.org/10.1080/03610926.2020.1737879.

Schilling, Mark F., Bret A. Holladay, and Jimmy A. Doi. "Improved Interval Estimation of Negative Binomial Parameters: A Coverage Probability Approach." \*Journal of Statistical Computation and Simulation\*, vol. 94, no. 1, 2024, pp. 126-141, https://doi.org/10.1080/00949655.2023.2235046.

Sterne, Theodor. "Some Remarks on Confidence or Fiducial Limits." \*Biometrika\*, vol. 41, no. 1/2, 1954, pp. 275–278.

Zhang, Lei, and William D. Johnson. "Approximate Confidence Intervals for a Parameter of the Negative Hypergeometric Distribution." \*Proceedings of the Survey Research Methods Section, Joint Statistical Meetings (JSM)\*, 2011, pp. 1753-1767.

"California Polytechnic State University--San Luis Obispo Student Life." \*U.S. News & World Report\*, https://www.usnews.com/best-colleges/california-polytechnic-state-university-san-luis-obispo-1143/student-life. Accessed 22 Sept. 2024.