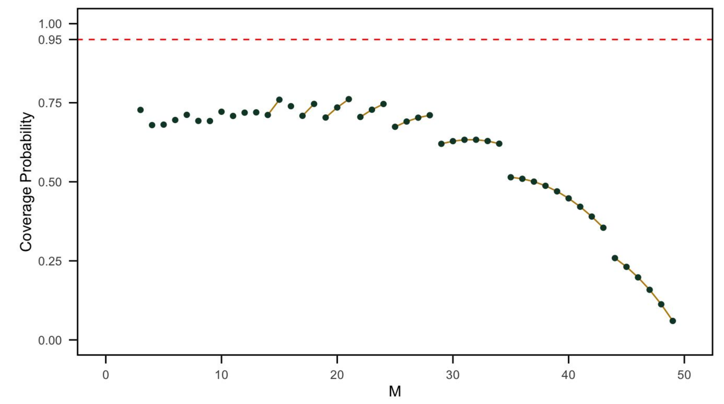
Contributions to confidence intervals for parameters of discrete distributions

Hannah Pawig & Rachel Roggenkemper. Dr. Bret Holladay, Department of Statistics,
California Polytechnic State University, San Luis Obispo

CAL POLY
Frost Summer Undergraduate Research Program
BAILEY COLLEGE OF SCIENCE & MATHEMATICS

Motivation

- In essence, a confidence interval offers a range of plausible values for an unknown parameter. Many commonly used methods for constructing confidence intervals, especially those that rely on large sample sizes, tend to perform poorly when estimating parameters from discrete distributions. These methods can capture the parameters of interest only approximately at the desired rate—highlighting the need for more accurate approaches.
- While the Poisson distribution has been widely studied, newer methods like Conditional Minimal Cardinality (CMC) method have yet to be applied. In contrast, the Negative Hypergeometric (NHG) distribution has been overlooked, with no existing exact confidence methods.
- Our research addresses these limitations by applying CMC to the Poisson case and comparing it to established methods, while developing several new methods for the NHG case.
- Making these methods both accessible and easy to use is essential, as many researchers may continue relying on outdated approaches out of convenience or simply due to a lack of awareness of better alternatives.



Coverage probability plot for the Normal Approximation Method for NHG

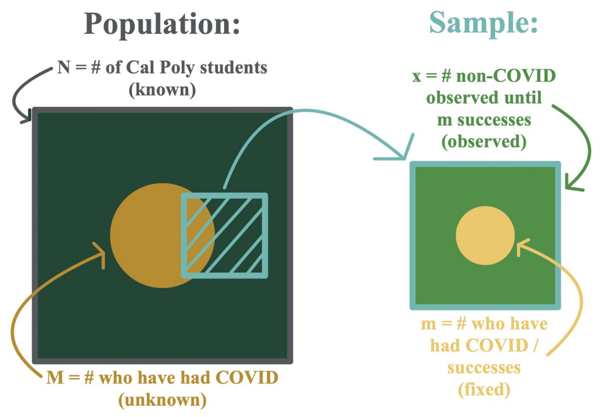
Application

Poisson:

- The Poisson Distribution is highly applicable to real-world data. It models the average number of rare, independent events over a fixed time period, with no upper bound.
- Has a single parameter: the mean number of successes.
- For example, suppose we're interested in the number of visible meteorites in San Luis Obispo during the peak of the Perseids meteor shower and obtain a 95% confidence interval of (120.7, 249.4). This interval provides a plausible range of values for the expected number of meteorites visible per year during the shower's peak. The confidence level (95%) indicates how likely it is that the method used will produce correct intervals in repeated trials.

Negative Hypergeometric:

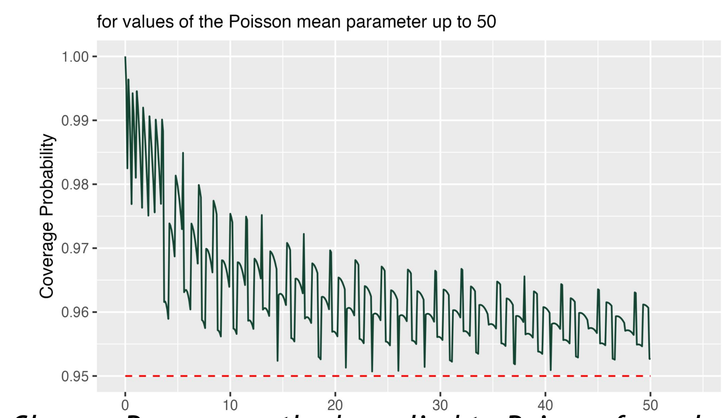
- The Negative Hypergeometric Distribution models sampling without replacement until a fixed number of successes is observed, focusing on the number of failures before reaching that target.
- It is useful in biological and industrial contexts for estimating population parameters through sequential sampling.
- Unlike the Hypergeometric Distribution, which stops after a set number of draws, the NHG stops after a set number of successes.
- Unknown parameter: number of successes in the population



(Left) Application for Poisson: Estimating the number of visible meteorites during Perseid meteor shower peak
(Right) Application for NHG: Estimating number of students who have had COVID out of all Cal Poly students

Goal

Our goal is to make advanced confidence intervals accessible to the general public, particularly to non-coders and those without a statistical background, through a online (Shiny) app. Below is a coverage probability plot for the Clopper-Pearson method in the Poisson case. A coverage probability function (cpf) determines the long-run probability that a method will accurately capture the corresponding parameter value in repeated trials. While this method is widely regarded as the "gold standard" for exact confidence procedures and guarantees the desired confidence level, it tends to be conservative, resulting in wider intervals than necessary. This underscores the importance of making newer methods accessible. In addition to providing alternatives to the previously mentioned approximate methods, these newer methods offer greater precision (narrower intervals) than the current gold-standard approach.



Coverage probability plot for the Clopper-Pearson method applied to Poisson for values of the mean parameter up to 50.

Results & Conclusions

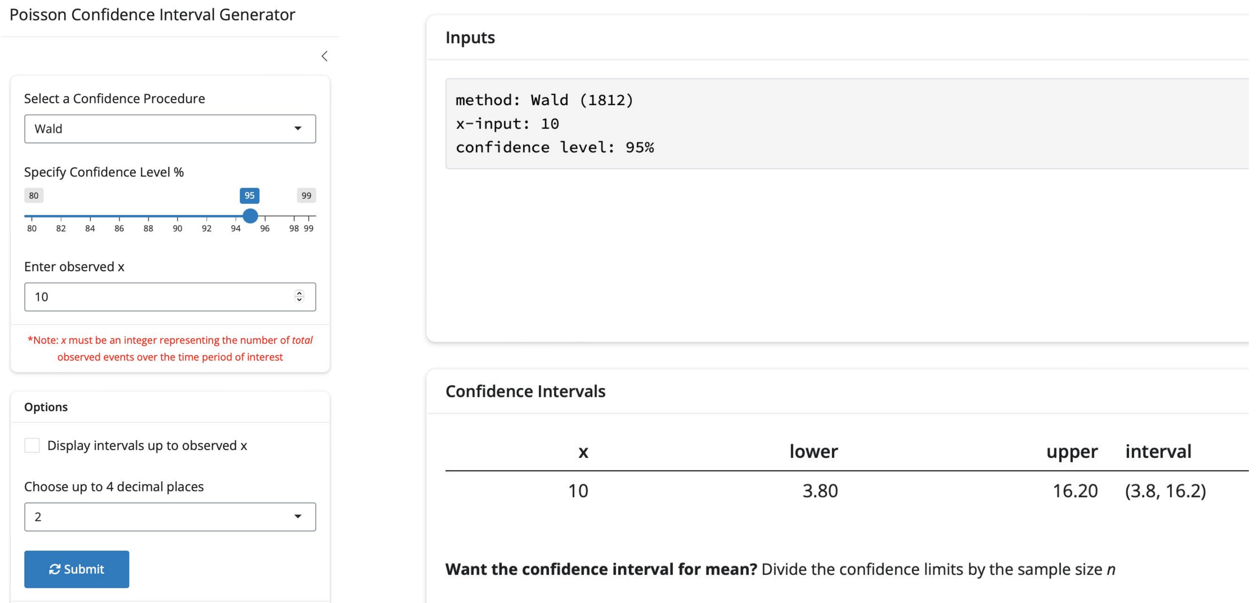
- We developed R functions that generate confidence intervals for the top methods for estimating the Poisson mean and applied CMC to the Poisson distribution for the first time. All methods introduced for the NHG case are novel, inspired by techniques for estimating parameters of other discrete distributions.
- We then conducted comparative analyses of the various methods to identify the "best" method. To compare the methods, we used expected and average confidence interval width as our metrics. For the Poisson distribution, Crow-Gardner and CMC performed the best. For the NHG distribution, Modified Sterne and CMC performed the best.

Distribution	Poisson	NHG
Exact Methods	Analog to Clopper-Pearson (1934)	Analog to Clopper-Pearson (1934)
	Modified Sterne (2014)	Modified Sterne (2014)
	Crow & Gardner (1959)	Crow & Gardner (1959)
	Byrne & Kabaila (2005)	Byrne & Kabaila (2005)
	Blaker (2000)	Blaker (2000)
	Conditional Minimal Cardinality - CMC (2023)	Conditional Minimal Cardinality - CMC (2023)

Table: Results of comparative analysis using expected width and average width. Methods in blue performed best.

Shiny App

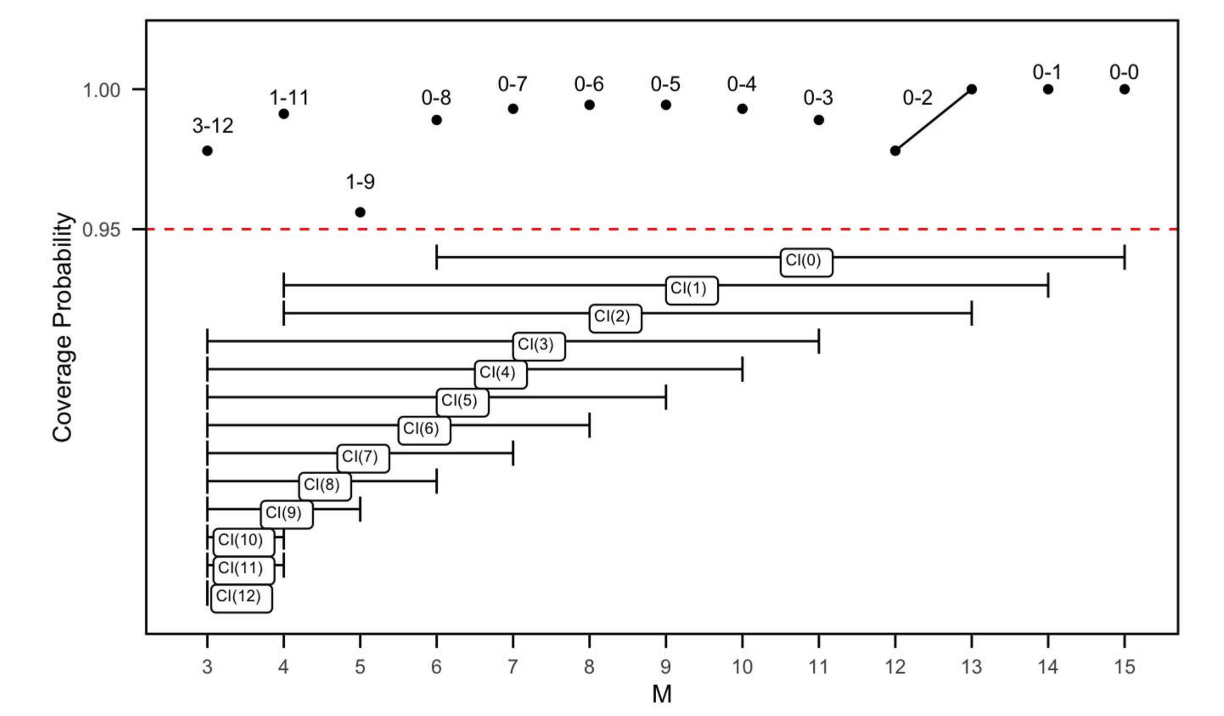
We developed a free public user-friendly Shiny app to make our methods accessible. Users can input their data to easily compute confidence intervals. The app serves as a bridge for researchers who lack the time or coding experience to implement these algorithms on their own.



Pictured: Shiny app interface for computing confidence intervals for the Poisson mean

Methodology

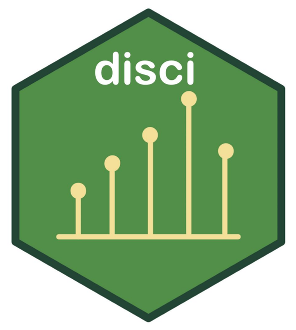
- Each set of confidence intervals corresponds to a unique coverage probability function (cpf). This one-to-one relationship allows us to "reverse engineer" a set of intervals by beginning with a high-performing cpf and working backward to create confidence intervals with desirable properties.
- In particular, this approach allows us to develop methods that provide coverage close to the nominal confidence level without falling below it, resulting in more precise intervals than traditional approaches.



Interplay between a 95% NHG cpf and resulting confidence intervals when population size is 15 and sampling continues until three successes are observed.

Future Directions / Next Steps

- For NHG, develop methods for estimating population size (N).
- Expand the Shiny apps and develop an R package to make these tools accessible to R users.



An idea for R package named "disci" (derived from discrete Confidence Intervals)

References

Debany, Warren H. *Improved Methods for Exact Confidence Intervals When Sampling With and Without Replacement for Rare Events*. Air Force Research Laboratory, August 2023.
Doi, J. A., Holladay, B., & Schilling, M. F. (2023). Improved interval estimation of negative binomial parameters: a coverage probability approach. Journal of Statistical Computation and Simulation, 94(1), 126–141. https://doi.org/10.1080/00949655.2023.2235046
Holladay, B. A. (2019). Contributions to Interval Estimation for Parameters of Discrete Distributions. UC Santa Barbara. Retrieved from https://escholarship.org/uc/item/5mz623j1
Sadovski, Vadim. Meteor shower. Digital Image. Discover Magazine. Kalmbach Media, 4 April 2018. Web. Accessed 5 October 2024.
Schilling, M. F., & Doi, J. A. (2014). A Coverage Probability Approach to Finding an Optimal Binomial Confidence Procedure. The American Statistician, 68(3), 133–145. https://doi.org/10.1080/00031305.2014.899274
Schilling, M. F., & Holladay, B. (2016). Length minimization for Poisson confidence procedures. Communications in Statistics - Theory and Methods, 46(2), 861–873. https://doi.org/10.1080/03610926.2015.1006782
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.
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.

Acknowledgements

This research was generously supported by the William and Linda Frost Fund in the Cal Poly Bailey College of Science and Mathematics.