CALIFORNIA STATE UNIVERSITY, NORTHRIDGE

A New Approach to Precise Interval Estimation for the Parameters of the
Hypergeometric Distribution

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science in Mathematics

By

Alyssa Stanley

December 2017

The thesis of Alyssa Stanley is approved:

Majid Mojirsheibani, Ph.D.

Date

Ann Watkins, Ph.D.

Date

Mark Schilling, Ph.D., Chair

Date

California State University, Northridge

ACKNOWLEDGEMENTS

I would first and foremost like to thank Dr. Mark Schilling for my newfound love and passion in the field of statistics, as well as giving me the opportunity to work with him for this thesis. He has a wonderful talent for making students with a distaste for statistics find an appreciation for it. His dedication and compassion for his students is admirable and has set the example of the kind of educator I would like to be in the future. I am indebted to him for his kindness in allowing me to work with him. This thesis and my graduate career would not have been possible without his guidance.

I would also like to thank Dr. Majid Mojirsheibani and Dr. Ann Watkins for being part of my thesis committee. Time is valuable and I would like to thank the both of them for giving theirs to me.

Thank you to all my family and friends for always believing in me. I am extremely fortunate to have such an amazing support system.

A special thank you to my parents who have always encouraged my academic success. I could not dream of achieving any of this without their unwavering support and love. Thank you for giving me the confidence and courage to do things I'm afraid of.

Finally, I would like to thank my loving fiancé, Timmy Todisco. He has experienced first-hand the sleepless nights, endless hours of work and frustration, and loved me through it all. You never doubted me for one second and I don't think I'll ever be able to express to you how grateful I am to have your support.

DEDICATION

This thesis is dedicated to my incredible parents, John and Jayne Stanley, who taught me from a very young age that I can do anything I set my mind to.

TABLE OF CONTENTS

Signature Page	ii
Acknowledgements	iii
Dedication	iv
List of Figures	vii
List of Tables	ix
Abstract	x
 Chapter 1. Introduction	 1
Chapter 2. Estimating M	3
2.1 Introduction	3
2.2 Acceptance Functions	4
2.3 Selection of Acceptance Functions	7
2.4 Resolving the Issue of Gaps	13
2.5 Formal Description for Estimating M	16
2.6 Alternate Methods	17
2.6.1 A Description of Wang's Method	18
2.6.1.1 One-Sided Confidence Intervals	19
2.6.1.2 Two-Sided Confidence Intervals	21
2.7 A Comparison	22
2.8 Other Criteria.....	26
2.9 Applications	28
Chapter 3. Estimating N.....	31
3.1 Introduction	31
3.2 Acceptance Functions	32
3.3 Selection of Acceptance Functions	34
3.4 Resolving the Issue of Gaps	37
3.5 Formal Description for Estimating N.....	39
3.6 Alternate Methods	40
3.6.1 A Description of Wang's Method	40
3.7 A Comparison	43
3.8 Other Criteria.....	47

3.9 Applications	49
Chapter 4. Conclusion.....	51
References	52
Appendix.....	54

LIST OF FIGURES

Figure 1. Probability of accepting a lot of size 100, given that 5 or fewer defective units in a sample size of 25 units is acceptable.	4
Figure 2. Hypothetical confidence sets obtained when Ω at $M = 10$ is not a proper interval for the case when $n = 10$ and $N = 20$	6
Figure 3. Coverage probabilities of acceptance curves eligible for selection based on a 90% confidence level for $N = 20$ and $n = 7$	7
Figure 4. Confidence sets for the case when $n = 10$ and $N = 20$ for various hypothetical acceptance sets.	8
Figure 5. The set of minimal acceptance functions eligible for selection for an exact minimal span confidence procedure for $N = 20$, $n = 7$, $1 - \alpha = 90\%$	9
Figure 6. Coverage probability function with associated acceptance functions and confidence intervals for the minimal span procedure method with $N = 20$, $n = 7$ and 90% confidence level.	11
Figure 7. Confidence sets for the case when $n = 20$, $N = 75$, and $\alpha = 0.10$. Red points represent gaps.	14
Figure 8. (a) – (b) Acceptance functions selected for the case $N = 75$, $n = 20$, and 90% confidence level.	15
Figure 9. Type O acceptance functions for the case when $n = 10$, $N = 20$ with a reference line at $1 - \alpha = 0.95$	20
Figure 10. Coverage probability of the minimal span procedure, along with the coverage probability of Wang's method shown in red when it differs, for the case $n = 20$, $N = 50$, $1 - \alpha = 90\%$	25
Figure 11. 90% confidence intervals for $x = 12$ at each $n \leq 60$ for the case when $N = 60$. A violation to monotonicity is shown in red.	27
Figure 12. Acceptance functions that lie above $1 - \alpha = 0.90$ for the case when $n = 7$ and $M = 20$	33
Figure 13. Acceptance functions eligible for selection for $N \geq 100$ for the case when $n = 7$, $M = 20$, and $1 - \alpha = 0.90$	35
Figure 14. Acceptance functions above the 90% confidence level for all $N \in \{20, 21, \dots, 79, 80\}$ for the case $n = 7$ and $M = 20$. Functions with minimal span at select values of N are highlighted in blue.	36
Figure 15. (a) – (b) Acceptance functions selected for the case $M = 15$, $n = 10$, and 90% confidence level.	38
Figure 16. Type O acceptance functions of the form $AF(7 - l)$ for the case when $n = 7$, $M = 20$, and $1 - \alpha = 0.90$	41
Figure 17. Acceptance functions of the form $AF(0 - u)$ between $20 \leq N \leq 170$ that lie above $1 - \alpha = 0.95$ for the case $n = 7$ and $M = 20$	42

Figure 18. The ratio of length of the minimal span procedure versus Wang's procedure for $x > 0$ for the case when $n = 25$, $M = 50$, and 90% confidence level with a reference line shown when the ratio is equal to one.	46
Figure 19. (a) – (b) Coverage Probability of the minimal span procedure along with Wang's shown in red where it differs for the case $n = 25$ and $M = 50$	47

LIST OF TABLES

Table 1. Confidence intervals for $n = 7$, $N = 20$, and $1 - \alpha = 90\%$, 95% , and 99% using the method of minimal span acceptance functions with highest coverage probability.....	12
Table 2. 90% confidence intervals for Wang's, Konijn's, and the minimal span procedure for $n = 20$, $N = 50$	23
Table 3. Confidence intervals for M and M / N when $n = 7$ and $N = 20$ at the 90% confidence level.	29
Table 4. 90% confidence intervals for N when $n = 7$ and $M = 20$	37
Table 5. 90% confidence intervals for Wang's and the minimal span procedure when $n = 25$ and $M = 50$	44

Abstract

A New Approach to Precise Interval Estimation for the Parameters of the Hypergeometric Distribution

By

Alyssa Stanley

Master of Science in Mathematics

We study interval estimation for both parameters of the hypergeometric distribution: (i) the number of successes in a finite population and (ii) the size of the population. In contrast to traditional methods that specify intervals via a formula, our approach is to first establish the coverage probability function of an ideal procedure. This in turn determines the confidence intervals.

In the case when the population is known and we wish to estimate the number of successes, we find that our approach performs better than preexisting methods in terms of coverage and average length. The precise nature of our confidence procedure for the success parameter near small values of success (failure) indicates a particular usefulness, for example, in the field of medicine. In the case of estimating the population size, we also find that our procedure performs better in terms of coverage probability and length in comparison to a significant existing method. Applications are given to ecology and medicine.

Chapter 1. Introduction

A problem in statistics that often occurs in practice is to estimate the proportion of a finite population having a certain characteristic. One common model is the well-known binomial distribution. This distribution models a scenario in which a sample of size n is taken with replacement and the number of units, x , having a certain characteristic is observed. We can then use this to estimate the proportion ($\hat{p} = x / n$) of the population of size N having a certain characteristic, which we may also refer to as a *success*.

Often, we must sample from a finite population *without* replacement. The *hypergeometric distribution* gives the probability of taking a sample without replacement and obtaining x successes. Examples where the hypergeometric distribution applies include polls, the lottery, acceptance sampling, and capture-recapture. Where n , x , and N are as described previously and M is the number of units in the population possessing the characteristic, the hypergeometric distribution has the following probability mass function:

$$P(X=x; M, N, n) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}}, \text{ where } \max\{0, n + M - N\} \leq x \leq \min\{M, n\}.$$

Notice that estimating M when N is known is equivalent to estimating the proportion of successes in the population $p = M / N$, and therefore is often approximated by the binomial distribution when n is small in comparison to N (e.g. Kinney 2014, p. 330). However, in the event that the sample size is not small compared to the population size, or if a precise estimation is necessary, using the binomial distribution as an approximation is not ideal.

The goal of this thesis is to obtain exact level $1 - \alpha$ confidence intervals $C(X)$

(i) for the parameter M when N is known and (ii) for the parameter N when M is known.

Here, *exact* confidence intervals are those in which the coverage probability of $C(X)$ is at least $1 - \alpha$. In the past, it was desirable to provide interval estimators that are based on a formula, see, for example Katz (1953) and Thompson (2002). Although this may have been preferable when computers and statistical programs were not readily available, nowadays we can take advantage of these technological tools to easily generate exact confidence intervals that perform better than the previously mentioned formulas in terms of coverage and interval length.

In the following chapters, we will provide a confidence procedure for both N and M using a coverage probability approach similar to that of Schilling and Doi (2014), Schilling and Holladay (2017) and Choi (2015). We provide an in-depth comparison to a recently developed procedure due to Wang (2015), a method that also uses a coverage probability approach. We will also discuss the length superiority of our confidence procedure in comparison to preexisting methods, in addition to demonstrating the new procedure's high coverage performance. We also examine to what extent this method achieves other previously studied criteria.

Chapter 2. Estimating M

2.1 Introduction

As discussed in Sahai and Khurshid (1995), the hypergeometric distribution plays an important role in modern biological and biomedical applications when it is reasonable to assume a finite population N . For example, suppose a researcher knows that N people have been exposed to a disease and wants to estimate the number of people M that are infected by it. If it is not practical to determine the disease status of every individual in the population, then in order to obtain an estimate for M , the researcher can observe a random sample n of those exposed and determine the number of people x from the sample that are infected. The sample proportion $\hat{p} = x / n$ gives a natural point estimate for the true proportion $p = M / N$. Confidence intervals in the context of the proportion will be discussed later in this chapter.

Another application of the hypergeometric distribution is *acceptance sampling*. Suppose a company wants to determine whether to accept or reject a lot for having too many defective products. Perhaps testing the entire lot is too expensive, takes too much time, or will ruin the product, so testing only a random sample of n products is feasible. Now suppose that it is acceptable to have x or fewer defective products in the random sample, otherwise the lot cannot be distributed to consumers. The probability of accepting the lot is given by a function of the true number of defective products M :

$$P(X \leq x) = \sum_{i=0}^x \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}}$$

The function above is shown in Figure 1 for all M in $[0, 100]$, although it is actually defined only on the set of integers within this range. The curve created by this function is called an *acceptance curve* (Schilling and Doi, 2014). In this particular

example, we have a lot of 100 units that needs to be accepted or rejected. We sample 25 units and allow 5 of them to be defective without rejecting the lot. As one would expect, the probability of accepting a lot decreases as the true number of defective units M in the lot increases.

Figure 1. Probability of accepting a lot of size 100, given that 5 or fewer defective units in a sample of size 25 units is acceptable.

2.2 Acceptance Functions

The acceptance curve shown in Figure 1 is referred to as a *Type O* acceptance curve (Schilling and Doi, 2014). In addition to *Type O* curves, there are also *Type I* acceptance curves, and both types are a critical component of this thesis. We explore acceptance curves in detail in this section, as they are the basis of our confidence procedures for the parameters of the hypergeometric distribution.

The confidence procedures discussed in this thesis are determined by their *coverage probability*. The coverage probability $CP(M)$ for a given M for a hypergeometric confidence procedure is the probability of observing any number x of successes for which the associated confidence interval contains M . The set of all $CP(M)$ for $M \in \{0, 1, 2, \dots, N\}$ for a confidence procedure is called the *coverage probability*

function, which is made up of portions of acceptance curves. Since the purpose of this thesis is to find a high performing exact confidence procedure, the selection of appropriate acceptance curves requires that the coverage probability function is always greater than or equal to $1 - \alpha$.

Recall that the hypergeometric distribution is discrete and hence is only defined for $M \in \{0, 1, 2, \dots, N\}$. Thus, it is more appropriate to refer to the acceptance curves of the hypergeometric distribution as *acceptance functions*. The following formula represents these acceptance functions $AF_{\Omega}(M)$ as a function of the parameter M :

$$AF_{\Omega}(M) = P(X \in \Omega) = \sum_{x \in \Omega} \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}},$$

where Ω is any subset of $\{0, 1, 2, \dots, n\}$. In other words, an acceptance function provides the probability that $x \in \Omega$ for each value of M .

Consider the hypothetical case in Figure 2 where $n = 10$ and $N = 20$. From the figure, we can see that $\Omega = \{2, 3, 4\} \cup \{6, 7, 8\}$ at $M = 10$. Sets of this form contain holes horizontally, thus typically creating holes vertically. In this case, the resulting confidence set would be $M \in \{5, 6, 7, 8, 9\} \cup \{11, 12, 13, 14, 15\}$, which is not a proper interval because it is not a set of consecutive integers. Thus, we restrict all subsets to be of the form $\Omega_{lu} = \{l, l+1, l+2, \dots, u-2, u-1, u\}$, since any other form of Ω can result in confidence sets for M that are *not* intervals. Acceptance sets of the form Ω_{lu} are called *acceptance intervals* (Blyth and Still 1983).

Figure 2. Hypothetical confidence sets obtained when Ω at $M = 10$ is not a proper interval for the case when $n = 10$ and $N = 20$.

Acceptance functions generated by acceptance intervals of the form Ω_{0u} and Ω_{ln} are called *Type O* functions. Those generated by acceptance intervals of the form Ω_{lu} , such that $l \neq 0$ and $u \neq n$, are called *Type I* functions. Type O acceptance functions of the form $AF_{\Omega_{0u}}$ have a maximum of 1 at $M = 0$ and a minimum of 0 at $M = N$. Similarly, Type O acceptance functions of the form $AF_{\Omega_{ln}}$ have a maximum of 1 at $M = N$ and a minimum of 0 at $M = 0$. In contrast, Type I curves obtain their minimum of 0 at $M = 0$ and N while obtaining their maximum within $M \in \{1, \dots, N - 1\}$. The proof of the minimum and maximum of acceptance functions of the form $AF_{\Omega_{0u}}$ is shown in the Appendix, with the proof of acceptance functions of the form $AF_{\Omega_{ln}}$ following similarly.

Due to the position of their maxima and minima, acceptance functions of the form $AF_{\Omega_{0u}}$ and $AF_{\Omega_{ln}}$ exceed the $1 - \alpha$ level for values near $M = 0$ and N , respectively. On the other hand, Type I functions of interest will exceed $1 - \alpha$ for values of M closer to $[N/2]$. Therefore, we find that the coverage probability function for a hypergeometric

confidence procedure will consist of Type O acceptance functions near M equal to 0 and N , while consisting of Type I acceptance functions elsewhere.

Figure 3 displays the acceptance functions with coverage probabilities greater than or equal to $1 - \alpha$ for the case of a population size of $N = 20$, a sample of size $n = 7$, and $\alpha = 0.10$ (90% confidence level). For simplicity, the notation of $AF(l - u)$ will refer to the acceptance function generated by the acceptance interval $\mathcal{Q}_{lu}$ for the remainder of this thesis, where the value $u - l$ is called the *span* of the acceptance function. Lines have been added in Figure 3 to connect the consecutive values of each acceptance function for easier visualization, but we did not smooth the resulting piecewise linear graph into a curve in order to emphasize the restriction of M to the domain $\{0, 1, 2, \dots, N\}$.

Figure 3. Coverage probabilities of acceptance functions eligible for selection based on a 90% confidence level for $N = 20$ and $n = 7$. Here, the notation $l - u$ represents $AF(l - u)$.

2.3 Selection of Acceptance Functions

The goal of creating a confidence procedure with superior length in comparison to preexisting methods will drive the selection process of acceptance curves. By superior

length, we mean a procedure for which the *average length* is shorter than that of any existing exact confidence procedures, where the average length is the mean length of all $n + 1$ confidence intervals. The main procedure that comes into competition with ours is one recently developed by Wang (2015), to be discussed later.

Since our confidence procedure is exact, the primary quality that an acceptance function must have in order to be eligible for selection is a coverage probability greater than or equal to $1 - \alpha$. Beyond this criteria, we must determine which selection of functions will give our confidence procedure the shortest average length. Due to a result from Crow (1956), it turns out that choosing acceptance functions with minimal span naturally leads to this result. See, for example, the hypothetical case for $n = 10$ and $N = 20$ shown in Figure 4. The horizontal sets represent the acceptance function selected at each value of M , while the vertical sets represent the confidence intervals for each x that result from the choice of acceptance functions. Notice, moving from left to right, that as the horizontal sets get shorter, so do the vertical sets. Thus, we expect that acceptance functions of minimal span will result in shorter intervals for M .

Figure 4. Confidence sets for the case when $n = 10$ and $N = 20$ for various hypothetical acceptance sets.

Thus, once we have all the acceptance functions that lie above $1 - \alpha$, we choose acceptance functions with minimal span. Figure 5 shows the set of all available minimal span acceptance functions for $M \in \{0, 1, \dots, N\}$ for the case when $N = 20$, $n = 7$, and $1 - \alpha = 90\%$. It can be seen in Figure 5 that the only minimal span acceptance functions eligible for selection up until $M = 5$ are Type O functions. At $M = 5$ we have eight functions to select from: $AF(0 - 3)$, $AF(0 - 4)$, $AF(0 - 5)$, $AF(0 - 6)$, $AF(0 - 7)$, $AF(1 - 4)$, $AF(1 - 5)$, and $AF(1 - 6)$. The shortest span, three, is obtained by the two functions $AF(0 - 3)$ and $AF(1 - 4)$. Figure 5 displays the resulting minimal span acceptance functions in consideration for the case $N = 20$ and $n = 7$ at the 90% confidence level.

Figure 5. The set of minimal acceptance functions eligible for selection for an exact minimal span confidence procedure for $N = 20$, $n = 7$, $1 - \alpha = 90\%$.

Now the question is how we should select the functions when there is a choice between two or more minimal span acceptance functions (as is the case for $M \in \{5, 6, 14, 15\}$). To investigate this, we explore a few methods used in previous papers for estimating the parameters of various distributions.

Crow (1956) proposes to always choose the minimal span acceptance function with the largest l and u values that is above $1 - \alpha$. That is, according to Crow, $AF(1 - 4)$ would be chosen rather than $AF(0 - 3)$ at $M = 5$. Using this method, we would also use $AF(1 - 4)$ at $M = 6$. However, Blythe and Still (1983) have pointed out that Crow's approach results in intervals that are not strictly monotone—often two or more confidence intervals for consecutive values of x have the same lower or upper limit.

On the other hand, Kabaila and Byrne (2001) propose a method that is the opposite of Crow's. Moving through the values of M in an increasing fashion, Kabaila and Byrne retain the same acceptance function for each consecutive integer until it falls beneath $1 - \alpha$. Once the acceptance function is below $1 - \alpha$, the next available acceptance function with minimal span (the one with the *smallest* l and u values above the confidence level) is chosen until it, too, drops below $1 - \alpha$. According to Kabaila and Byrne, $AF(1 - 4)$ would not be selected until $M = 7$ in Figure 5.

Another method, due to Blythe and Still (1983), is an intermediate solution between the previous two methods. Rather than selecting the function furthest to the left or right, Blythe and Still choose to switch to the next acceptance function at the midpoint between the endpoints of the interval for which the parameter has two acceptance functions above the chosen confidence level. If there are more than two acceptance functions to select from, Blythe and Still (1983) recommend taking each confidence interval endpoint to be the midpoint of the interval of possibilities. This method would work ideally when the parameter to be estimated can take on decimal values, but is not practical for our parameter M , which is only defined on the positive integers.

As previously stated, our procedure will be following an approach similar to that of Schilling and Doi (2014). Contrary to the three previously mentioned methods, Schilling and Doi look at the acceptance functions from top to bottom, rather than from left to right. In other words, when faced with a choice between two or more minimal span acceptance functions, this method selects the function with the *highest* coverage probability.

Using this strategy, we choose $AF(0 - 3)$ at $M = 5$ in Figure 5 since it has a higher coverage probability than $AF(1 - 4)$. However, $AF(1 - 4)$ has a higher coverage probability than $AF(0 - 3)$ at $M = 6$, and thus would be selected. Figure 6 shows the resulting coverage probability function, along with the associated acceptance functions and the confidence intervals produced for the case when $N = 20$, $n = 7$, and $1 - \alpha = 90\%$ after choosing minimal span acceptance functions with the highest coverage probability at each M .

Figure 6. Coverage probability function with associated acceptance functions and confidence intervals for the minimal span procedure method with $N = 20$, $n = 7$ and 90% confidence level. Here, for example, $CI(0)$ denotes the confidence interval when $x = 0$.

Table 1 displays the resulting confidence intervals for the case above as well as for two other typical confidence levels, 95% and 99%. This specific example demonstrates some nice characteristics that we would expect to see for a viable confidence procedure. First we note that, for each confidence level, both the lower limits and the upper limits of the confidence intervals increase as x increases. This is known as *monotonicity* and is a natural result—as the number of successes in the sample increases, so does the expected number of successes in the population.

x	Confidence Level		
	90%	95%	99%
0	[0, 5]	[0, 6]	[0, 8]
1	[1, 8]	[1, 10]	[1, 11]
2	[2, 11]	[2, 12]	[2, 13]
3	[4, 14]	[4, 14]	[3, 15]
4	[6, 16]	[6, 16]	[5, 17]
5	[9, 18]	[8, 18]	[7, 18]
6	[12, 19]	[10, 19]	[9, 19]
7	[15, 20]	[14, 20]	[12, 20]

Table 1. Confidence intervals for $n = 7$, $N = 20$, and $1 - \alpha = 90\%$, 95% , and 99% using the method of minimal span acceptance functions with highest coverage probability.

Another desirable characteristic that can be seen with respect to the common confidence levels in Table 1 is *nesting*. A confidence procedure that possesses the nesting property means that for each x , every confidence interval contains any confidence interval with a lower confidence level that is computed from the same data. Nesting is also an expected result for any practical confidence procedure to have.

We can also observe the property of *equivariance* in Table 1. An equivariant confidence procedure is one that satisfies the following: If x generates the confidence interval $[l_x, u_x]$ then $n - x$ generates the confidence interval $[N - l_x, N - u_x]$ (Blyth and

Still, 1983). Equivariance is a property that applies to interval estimation for the hypergeometric distribution (as well as the binomial distribution) since the definition of “success” for a general hypergeometric experiment is arbitrary and can just as easily be switched with “failure,” while still having the same consequence.

Since equivariance is a property of confidence intervals for the hypergeometric parameter, M , it will be a component of our algorithm for finding a confidence procedure of superior length. Nesting and monotonicity, on the other hand, are merely desirable properties that we wish to achieve, although it is not true that both properties will hold for all cases. Blaker (2000) investigated the properties of length optimality, nesting, and confidence sets that are intervals at the same time and proved that every possible confidence procedure necessarily fails to satisfy at least one of these three properties. We will explore the extent to which nesting and monotonicity hold in our minimal span confidence procedure for numerous values of n and N in Section 2.8.

2.4 Resolving the Issue of Gaps

Although the selection of acceptance functions in the example of Figures 3 and 5 works seamlessly, in particular situations the issue of *gaps* can arise. Confidence sets containing gaps occur when the acceptance functions $AF(l - u)$ selected fail to be nonincreasing in both l and u . Figure 7 shows the consequence of selecting acceptance functions with a decreasing l and/or u value, where the solid red points indicate gaps. First note that for optimal viewing, the axes have been switched compared to previous similar figures in this thesis. The vertical sets of x represent the acceptance intervals at each M , hence Figure 7 displays the confidence intervals for each x horizontally. Notice

that the lower limit of the acceptance interval for $M = 9$ is smaller than for $M = 8$. This results in a gap in the confidence interval for $x = 0$. There is a similar occurrence for u when $M = 67$, resulting in a gap in the confidence interval for $x = 20$. For this reason, we see that we must require the l and u sequences to be nondecreasing.

Figure 7. Confidence sets for the case when $n = 20$, $N = 75$, and $\alpha = 0.10$. Red points represent gaps.

Now consider the selection of acceptance functions for the case $n = 20$ and $N = 75$, at the 90% confidence level based on the method described in the previous section. Figure 8(a) shows that coverage probability function jumps from $AF(1 - 4)$ at $M = 8$ to $AF(0 - 4)$ at $M = 9$. Consequently the value of l changes from 1 to 0 and therefore the l sequence is not monotonic increasing, producing a gap. The resulting confidence set for $x = 0$ in this case would be the integers $M \in [0, 7] \cup [9]$ rather than a proper interval. This is clearly undesirable.

Figure 8. (a) – (b) Acceptance functions selected for the case $N = 75$, $n = 20$, and 90% confidence level. (a) Shows the initial procedure leads to a gap at $M = 9$. (b) Resolution to the gap issue.

In order to eliminate such gaps, we follow the method used by Schilling and Doi (2014), who proposed that if the choice of a particular minimal span acceptance function produces a gap, use the minimal span acceptance function with the next highest coverage probability in substitution. Following this method, $AF(0 - 4)$ is the function that disrupts the monotonicity, so we substitute $AF(1 - 5)$ as the acceptance function at $M = 9$. Figure 8(b) shows the graphical interpretation of this replacement of acceptance functions. The dotted line segment in Figure 8(b) indicates the acceptance function in violation of monotonicity. An extension of $AF(1 - 5)$ has been added in bold, which replaces $AF(0 - 4)$ at $M = 9$. After this modification to resolve gaps, we obtain a confidence interval for $x = 0$, namely $M \in [0, 7]$, at the negligible expense of a slightly lower coverage probability at $M = 9$.

Since the confidence intervals for M of the hypergeometric distribution are equivariant, we perform the above modification only for (i) $M \in [0, N/2 - 1/2]$ if N is odd

or (ii) $M \in [0, N/2]$ if N is even and apply the equivariance property to solve for the remaining intervals. Upon investigating all the samples sizes $n \in \{1, 2, \dots, N-1, N\}$ for each population size $N \in \{1, \dots, 99, 100\}$ at the confidence levels 90%, 95%, and 99%, we found that gaps occur for some x in only 54 cases of all possible 15,150 cases (0.356%).

It is clear that gaps are an extremely rare occasion for the hypergeometric case, as opposed to the binomial case where Schilling and Doi (2014) found that gaps occur for some x in approximately 40% of the 300 cases they investigated. The rarity of gaps in the hypergeometric distribution is likely due to the fact that the parameter M is restricted to positive integers, whereas the binomial parameter can take on all real numbers $p \in [0, 1]$.

It is also interesting to note that each time we encountered a gap in the cases examined, it consisted only of a singleton set. In fact, considering the gaps that occur for $M \in [0, [N/2]]$, all cases were analogous to the gap in Figure 8(a), yielding a confidence set of the form $M \in [a, b] \cup [b + 2]$. We also found that gaps never occur when the sample size n is more than 37% the population size N in the 15,150 cases that we investigated.

2.5 Formal Description for Estimating M

In summary, the method of producing confidence intervals of superior length for the hypergeometric parameter M chooses the acceptance function of minimal span with the highest coverage probability, except in the extremely rare case of when we need to resolve a gap, in which we substitute the acceptance function of minimal span with the one having the next highest coverage probability.

The following is a formal description of the algorithm used to find the minimal span confidence procedure for the hypergeometric parameter M :

Step 1: Beginning with $M = 0$, for each $M \in [0, N/2 - 1/2]$ if N is odd or $M \in [0, N/2]$ if N is even, let $AF_M(l - u)$ denote the acceptance function achieving the highest coverage probability above $1 - \alpha$ among all acceptance functions of minimal span at M . If more than one acceptance function achieves the highest coverage probability, select the function $AF_M(l - u)$ with the largest value of l . Assign M to the confidence intervals for each $x \in [l, u]$, except in the case described by Step 2.

Step 2: Whenever $AF_M(l - u)$ and $AF_{M+1}(l' - u')$ from Step 1 are such that $l' < l$ and/or $u' < u$, let k be the largest integer that produces $AF_{M+k}(l' - u')$. Reassign $M + 1, \dots, M + k$ to $x \in [l' + 1, u' + 1]$.

Step 3: Upon constructing the intervals for all $M \in [0, [N/2]]$, apply equivariance to compute the remaining confidence intervals. For x generating the confidence interval $[l_x, u_x]$, $n - x$ generates the confidence interval $[N - l_x, N - u_x]$.

2.6 Alternate Methods

Traditionally, confidence intervals have been provided by means of a formula. There exist a few variations of Wald-type intervals for the hypergeometric parameter M ; for example one popularly used formula

$$C_{Wald}(x) = \frac{Nx}{n} \pm t_{1-\frac{\alpha}{2}, n-1} \left(\frac{(N-n)Nx(n-x)}{n^2(n-1)} \right)^{1/2}$$

is provided in Thompson (2002). Intervals constructed by formulas typically achieve shorter length intervals, but significantly violate coverage probability and can even be shown to have coverage probabilities as low as n / N (Wang, 2015). Formula-based intervals, in general, do not guarantee the predetermined coverage probability and are therefore not strictly legitimate estimates.

There exist only a few methods of interval estimation for M of the hypergeometric distribution which obtain the desired coverage probability condition of never falling below $1 - \alpha$. These exact methods are due to Cochran (1977), Konijn (1973), Buonaccorsi (1987), and Wang (2015). Buonaccorsi (1987) shows that his confidence intervals are the same as Konijn's (1973), both of which were shown to be superior to Cochran's (1977) in terms of length.

The method recently developed by Wang (2015) is a modification to the intervals produced by Konijn (1973), which results in his method being superior to the previous three in terms of length, while maintaining coverage probability above the chosen $1 - \alpha$ level. Since Wang's procedure outperforms the previously studied methods, we provide an explanation of his method, along with a comparison of performance to the minimal span procedure.

2.6.1 A Description of Wang's Method

Wang (2015) recently proposed a confidence procedure also based on coverage probabilities. Wang describes his procedure as *optimal*. A careful reading of his paper shows that what he means by this is that it is "admissible"—if any interval is replaced with a proper subinterval, the resulting confidence procedure will have coverage probability strictly less than $1 - \alpha$. Following this definition, the confidence procedure for M in this thesis is also admissible. However, the performance of these two confidence procedures differ in average length, as will be detailed in the following section. It will become evident that the minimal span procedure yields similar intervals to Wang's, however, our approach often reduces the overall average length of the possible

confidence intervals for the hypergeometric parameter M and can therefore be reasonably considered to be superior to Wang's method.

2.6.1.1 One-Sided Confidence Intervals

One-sided confidence intervals for M have the form $[0, U(X)]$ and $[L(X), N]$ and are called the lower and upper $1 - \alpha$ level confidence intervals, respectively. They satisfy $P(M \leq U(X)) \geq 1 - \alpha$ for all M , and similarly $P(M \geq L(X)) \geq 1 - \alpha$ for all M .

Wang proves that the *one-sided* confidence intervals obtained by his method produce the smallest possible length. Our procedure produces these same intervals. To see how, note that the latter probability above can also be written as

$$P(h(M) \geq X) \geq 1 - \alpha \quad (1)$$

where $h(M)$ is some increasing function of M by solving $M \geq L(X)$ for X . Thus, finding the lower bound of the upper one-sided confidence interval is determined by the probability that $X \leq h(M)$. Here, we notice that $P(X \leq h(M))$ is simply a Type O acceptance function; thus only Type O acceptance functions are involved in the determination of upper one-sided confidence intervals, and due to equivariance for lower one-sided confidence intervals as well.

Figure 9 displays all the Type O acceptance functions when the sample size is 10 and the population size is 20. The shortest possible upper one-sided confidence interval for any value of x has a lower limit $L(X)$ at the smallest value of M for which $AF(0 - (x - 1))$ drops below the specified confidence level. It is important to notice that $AF(0 - (x - 1))$ is simply $P(X \leq x - 1)$, which represents the probability condition described in equation (1). As an example, consider when $x = 6$ in the figure below. We

must find the first value of M for which $AF(0 - 5)$ drops below 0.95. Looking at the graph, we notice that this occurs at $M = 8$. Thus, $L(6) = 8$ and we obtain the upper one-sided confidence interval $[L(6), N] = [8, 20]$. A technicality to mention is that we clearly require $L(0) = 0$.

Figure 9. Type O acceptance functions for the case when $n = 10, N = 20$ with a reference line at $1 - \alpha = 0.95$.

The upper limit of the lower one-sided confidence interval, $U(x)$ is obtained through equivariance by the equation:

$$U(x) = N - L(n - x). \quad (2)$$

For example, say we want to find the lower one-sided confidence interval for the case when $x = 6$. Based on the equation above, we must first find $L(4)$. Following the procedure described above, we locate the smallest M for which $AF(0 - 3)$ drops below 0.95 in Figure 9, which occurs when $M = 5$. Therefore we find $L(4) = 5$, and using equation (2) we find that $U(6) = 15$ is the resulting upper bound of the lower one-sided confidence interval, namely $[0, 15]$.

2.6.1.2 Two-Sided Confidence Intervals

Similarly to calculating Wald-type intervals, Wang's two-sided $1 - \alpha$ level confidence intervals $C(x) = [L(x), U(x)]$ are constructed by first finding $L(x)$ and $U(x)$ from the one-sided $1 - \alpha/2$ level confidence intervals $[L(x), N]$ and $[0, U(x)]$, respectively. As an example, by means of $L(6)$ and $U(6)$ calculated previously from Figure 9 at the 95% confidence level, Wang's corresponding two-sided 90% confidence interval would be $C(6) = [8, 15]$. It is important to note that these initial intervals are the same as Konijn's (1973).

Once $C(x) = [L(x), U(x)]$ has been calculated for each x , Wang proposes an algorithm which involves a modification of each $C(x)$. The algorithm starts by modifying the interval at $x = n/2$ if n is even or $x = [n/2]$ if n is odd, and x goes down by 1 in each step until it reaches 0. Once the construction of these intervals is complete, Wang uses the equivariance property to compute the remaining intervals.

Specifically, suppose that n is even. Starting at $x = n/2$, Wang uses an iterative algorithm to shorten the confidence interval $C(x)$ by an integer on each end, if possible, until the largest integer, a , is found such that the coverage is greater than $1 - \alpha$. Thus, the modified confidence interval for $x = n/2$ becomes $[a, N - a]$. In the next step, Wang modifies the confidence interval for $x = n/2 - 1$. He shortens the confidence interval for $x = n/2 - 1$ by increasing $L(x)$ up to an integer c and decreasing $U(x)$ to an integer d such that $d - c$ is the smallest value that maintains a coverage probability of at least $1 - \alpha$. This process is repeated for each confidence interval until x reaches 0. A similar approach is used for odd n .

In the instance where N or α is very small, the entire set of acceptance functions that lie above the specified confidence level are Type O functions. The resulting two-sided intervals obtained uniquely achieve minimal length. This is not the case when either N or α is not very small.

2.7 A Comparison

The approach that yields the minimal span confidence procedure and Wang's method are quite different, yet they often have similar outcomes. The most significant distinctions between the two methods are (i) that we use acceptance functions to determine confidence intervals while Wang modifies existing confidence intervals, and (ii) the fact that we always use minimal span curves while Wang does not.

We find that Wang's algorithm sometimes produces shorter intervals than ours towards the middle of the range of possible values of x , while ours sometimes produces shorter intervals than his at or near the ends. Table 2 displays a comparison of the confidence procedures for the case when $n = 20$ and $N = 50$ at the 90% confidence level, along with the intervals produced by Konijn's (1973) procedure which was introduced in Section 2.6 as an alternate exact method.

It can be observed in Table 2 that Wang's intervals are a subset of Konijn's. This is expected, since Wang starts with Konijn's intervals and shortens them when possible, while maintaining the minimum coverage probability. It can be seen in Table 2 that Wang is successful in shortening 15 out of 21 of Konijn's intervals. We may also observe that the minimal span procedure is a subset of Konijn's, and in fact is shorter for 19 of the 21 confidence intervals. Clearly, the minimal span procedure and Wang's procedure are

superior to Konijn's in terms of length. Thus, a comparison of the minimal span method to Wang's method is appropriate.

Table 2 is color coded to represent the differences between the minimal span procedure and Wang's procedure. Red represents when Wang's intervals are shorter than those produced by the minimal span procedure. Green represents when the minimal span procedure's intervals are shorter than those produced by Wang. Blue represents when the intervals are different, but both procedures yield an interval of the same length.

x	Wang's Procedure	Minimal Span Procedure	Konijn's Procedure
0	[0, 5]	[0, 4]	[0, 5]
1	[1, 9]	[1, 8]	[1, 9]
2	[2, 11]	[2, 11]	[2, 12]
3	[4, 15]	[4, 14]	[3, 15]
4	[6, 18]	[5, 17]	[5, 18]
5	[7, 20]	[7, 20]	[7, 20]
6	[10, 22]	[9, 22]	[9, 23]
7	[12, 25]	[12, 24]	[11, 25]
8	[13, 27]	[13, 27]	[13, 28]
9	[16, 29]	[15, 29]	[15, 30]
10	[19, 31]	[18, 32]	[17, 33]
11	[21, 34]	[21, 35]	[20, 35]
12	[23, 37]	[23, 37]	[22, 37]
13	[25, 38]	[26, 38]	[25, 39]
14	[28, 40]	[28, 41]	[27, 41]
15	[30, 43]	[30, 43]	[30, 43]
16	[32, 44]	[33, 45]	[32, 45]
17	[35, 46]	[36, 46]	[35, 47]
18	[39, 48]	[39, 48]	[38, 48]
19	[41, 49]	[42, 49]	[41, 49]
20	[45, 50]	[46, 50]	[45, 50]

Table 2. 90% confidence intervals for Wang's, Konijn's, and the minimal span procedure for $n = 20$, $N = 50$. Red represents when Wang's intervals are shorter than those produced by the minimal span procedure. Green represents when the minimal span procedure's intervals are shorter than those produced by Wang. Blue represents when the intervals are different, but both procedures yield an interval of the same length.

Notice that the confidence intervals of Wang's procedure versus the minimal span procedure differ at all the values of x except $x = 2, 5, 8, 12, 15$, and 18 . For the values of x where the two intervals differ, our method produces intervals that are shorter than Wang's in eight cases and longer in five. As a result the minimal span confidence procedure yields shorter average confidence interval length in this example than Wang's procedure does. The average length for our procedure for this case is 10.95, while Wang's is 11.05 and Konijn's is 12.10.

Due to the exclusive use of minimal span acceptance functions, our algorithm produces the smallest average length of any preexisting exact confidence procedure. Wang's procedure is only provided in detail for even n ; our investigation of all even integers $n \in \{2, 4, 6, \dots, N\}$ for each $N \in \{2, 3, 4, \dots, 100\}$ at the 90% confidence level reveals that the average confidence interval length obtained from our method is smaller than that of Wang's procedure in 73 (2.9%) of all the 2,500 cases (and equal in length in the remaining ones). On average, our procedure yields intervals that are 0.71% shorter in average length than Wang's in the cases where the average lengths differ. A similar result would be expected for odd values of n .

Testing the same cases as above, a comparison was run on the minimal span procedure and Konijn's procedure. It was found that the minimal span procedure shortened the average confidence interval length for nearly all the cases that were tested. Specifically, 2,438 of the 2,500 cases were shorter in average confidence interval length when the minimal span procedure was used (97.5%). Of the cases that were shorter, the minimal span procedure shortened the confidence intervals by an average of 8.7%.

The primary reason why the minimal span procedure improves the average length is because Wang and Konijn do not focus on minimal span acceptance functions. While Wang does find the values of c and d such that $d - c$ is the smallest value for which the coverage probability of $[c, d]$ is above $1 - \alpha$, which naturally tends to result in the use of minimal span acceptance functions, he does not include the restriction of smallest span.

Take, for example, the case in Table 2 where $n = 20$, $N = 50$, and $\alpha = 0.10$. Figure 10 shows the coverage probability of our method, along with the coverage probability of Wang's method shown in red when it differs from ours. We observe a peak in Wang's coverage probability at $M = 25$ and find that the acceptance function corresponding to Wang's algorithm at that value is $AF(7 - 13)$, while we use $AF(8 - 13)$ in our method. Note that using $AF(7 - 12)$ in our algorithm would give us the same coverage probability, but our program defaults to the acceptance functions with the larger l and u values when the coverage probability is equal for two or more minimal span acceptance functions.

Figure 10. Coverage probability of the minimal span procedure, along with the coverage probability of Wang's method shown in red when it differs, for the case $n = 20$, $N = 50$, $1 - \alpha = 90\%$.

Every point where Wang’s coverage probability differs from ours in Figure 10 is a value of M where a different acceptance function was used. We can observe that the selection of acceptance functions differs for seven points of M , and consequently fifteen confidence intervals are different for the two methods.

It can be observed in Figure 10 that six of the seven cases where the coverage probabilities differ, the minimal span procedure has higher coverage. This is a typical behavior that we see when comparing the two methods. Not only have we provided intervals that are shorter in average length than Wang’s, but we have also increased the certainty in the approximation, concluding that the minimal span procedure is superior.

2.8 Other Criteria

As mentioned in Section 2.3, we will discuss the occurrence of violations in monotonicity and nesting for the minimal span confidence procedure. In addition to testing monotonicity in the success parameter x discussed in Section 2.3, we would also like to test the monotonicity in n .

Given x and n , we would expect that if an additional trial results in failure, both limits of the confidence procedure would decrease. We must be careful to note that this necessarily fails for $x = 0$ and $x = n$, because the lower limit for the confidence interval of $x = 0$ will always be 0 and the upper limit when $x = n$ will always be N . To account for the boundaries on the endpoints, we will test that both limits of the confidence procedure are nonincreasing in the event of an additional trial resulting in failure. We conclude that the opposite case is fundamentally the same because of equivariance – if an additional trial resulted in a success, both limits of the confidence procedure would be

nondecreasing. Therefore, we only investigated the case when an additional trial results in a failure.

For each $N \in \{1, 2, 3, \dots, 100\}$ at the 90%, 95%, and 99% confidence levels, we observed the behavior of the endpoints of the confidence intervals for each $x \in \{0, 1, 2, \dots, n\}$ at all sample sizes $n \leq N$. Figure 11 shows all the 90% confidence intervals of $x = 12$ at each $n \leq N$ for the case when $N = 60$, one of a total of 15,150 cases we investigated. Of these cases, violations to the monotonicity in n occurred only 58 times (0.383%). The particular example shown in Figure 11 violates monotonicity when $n = 49$, where the points are highlighted in red. When $n = 48$, the resulting confidence interval for $x = 12$ is $M \in [12, 18]$, while the confidence interval for the same x at $n = 49$ is $M \in [13, 17]$. Nonmonotonicity occurs because the lower limit for the confidence interval increases from $l = 12$ to $l = 13$ as n increases. Every other value of n behaves the way we would expect. As n increases, the lower and upper limits of x are monotonically nonincreasing.

Figure 11. 90% confidence intervals for $x = 12$ at each $n \leq 60$ for the case when $N = 60$. A violation to monotonicity is shown in red.

We would also like to observe monotonicity in the success parameter, x . As the number of successes in a sample increases, we naturally expect the lower and upper confidence limits to increase. Upon investigating all the samples sizes $n \leq N$ for each population size $N \in \{1, \dots, 99, 100\}$ at the confidence levels 90%, 95%, and 99%, we found that violations to the monotonicity in x occurred only 54 times out of the 15,150 cases that were tested (0.36%). Of these cases we discovered that the violations occurred when consecutive values of x shared the same lower or upper confidence limit, never when the lower or upper confidence limits decreased.

Another desirable property that was mentioned in Section 2.3 is nesting. We expect that the intervals obtained using a higher confidence level contain the intervals obtained using a lower confidence level for each value of x . Investigation was performed for all $n \leq N$ for each associated $N \in \{1, 2, \dots, 99, 100\}$ at each confidence level $1 - \alpha \in \{0.900, 0.901, 0.902, \dots, 0.997, 0.998, 0.999\}$. It turns out that the only time violations to nesting occurred was after gaps were resolved.

2.9 Applications

Estimating the hypergeometric parameter M applies to a wide range of circumstances. Whenever we have a random sample without replacement from a finite population that obtains a binary outcome (success or failure), our interval estimation method may be applied. It has already been mentioned that the hypergeometric distribution applies to testing defective products in a factory. It may also be applied to any situation that we wish to estimate the number of “successes” or “failures.”

Another clear application of interval estimation for M that has been discussed briefly is its correspondence to the proportion of successes in the population. In order to obtain interval estimates for proportions, we must simply divide the lower and upper limits of the confidence intervals for M by the population size N .

Recall the case that was first introduced in Section 2.2 where $n = 7$ and $N = 20$ at the 90% confidence level. Table 3 shows the confidence intervals that are produced by the minimal span confidence procedure, along with the corresponding intervals of the proportion that were found by making the previously mentioned modification.

x	M	M / N
0	[0, 5]	[0.00, 0.25]
1	[1, 8]	[0.05, 0.40]
2	[2, 11]	[0.10, 0.55]
3	[4, 14]	[0.20, 0.70]
4	[6, 16]	[0.30, 0.80]
5	[9, 18]	[0.45, 0.90]
6	[12, 19]	[0.60, 0.95]
7	[15, 20]	[0.75, 1.00]

Table 3. Confidence intervals for M and M / N when $n = 7$ and $N = 20$ at the 90% confidence level.

The hypergeometric parameter M also corresponds to odds. Recall that odds are simply a ratio $X : Y$, where X represents success (failure) and Y represents failure (success) such that $X + Y$ is equal to the population size. Thus, in the context of the hypergeometric distribution we can rewrite the odds ratio as $M : N - M$. Therefore, providing an estimate for M also applies to the estimation of odds ratios.

Since our procedure improved Wang's significantly for small (large) values of x , we find that our method applies especially well in the field of medicine, where small (large) proportions are a common occurrence. In general, proportions are useful in medicine when estimating the prevalence of a disease among a particular ethnicity/region, the proportion of population infected by a disease after exposure, as well as use in scoring systems such as CENTOR, HEART, WELLS and more. See for example Engelberger et al. (2011), Fine et al. (2012), and Streitz et al. (2017).

For example, an article by Xu et al. (2014) conducted a study on women seeking combined hormonal contraception; their goal was to estimate the proportion that have medical contraindications. They found that the proportion of women in their sample with contraindications was approximately 2%. This particular article used binomial confidence intervals as an approximation. Since accuracy is of importance in medicine, the hypergeometric distribution would have also been appropriate had the authors of this article known the size of the population represented by their sample, as well as if they had a simple way of calculating an exact confidence procedure for this distribution. Our method in particular would provide short confidence intervals with the desired confidence level, specifically for this case where the number of women having contraindications is small.

Chapter 3. Estimating N

3.1 Introduction

Although it is a common problem to estimate M in the case where the hypergeometric distribution applies, sometimes we instead wish to estimate N when M is known. In particular, this situation occurs often in ecology when the *capture-recapture* method is used. Suppose an ecologist wants to estimate the population of a particular species. Using the capture-recapture method, he/she would randomly *capture*, mark, and release M units from an unknown population of N units. After a sufficient amount of time has passed, a *recapture* of a random sample of size n is performed from the same population and the number of units, x , that were previously marked is observed. The proportion of the marked sample is then matched to the proportion of the marked population, $x / n = M / N$, to obtain a point estimate of N , namely $\hat{N} = Mn / x$.

Our goal is to create a procedure that will yield an *interval estimate* of N . We notice that the minimum size the population could possibly be is the maximum size of either the capture or recapture sample, therefore $N \geq \max\{n, M\}$. For this reason, lower one-sided confidence intervals will be of the form $[\max\{n, M\}, U(X)]$. On the other hand, there is no limit to how large the population can be and thus we obtain unbounded upper one-sided confidence intervals of the form $[L(X), +\infty]$. We will also determine conventional two-sided intervals $[L(X), U(X)]$, using a similar procedure as in Chapter 2.

There are a few assumptions of the capture-recapture method that need to hold in order to ensure the validity of an interval estimate for N in such studies. According to Southwood and Henderson (2009) the following are the basic assumptions that apply to capture-recapture: (i) the animals are not affected by the marks and the marks are not lost,

(ii) the marked animals become completely mixed in the population, (iii) the probability of being caught is the same for any member of the population in both the capture and recapture, and (iv) sampling must be at discrete time intervals and the actual time involved in taking the samples is small in comparison to the overall time.

The purpose of these assumptions is to ensure that the statistical model is adequate and that the hypergeometric distribution provides a reliable estimate. Notice that assumption (i) guarantees accuracy in the number of marked animals in the recapture. If marks have fallen off, or make an animal more susceptible to death, the recapture will be a biased representation of the proportion of marked animals. In order to guarantee assumption (ii), enough time must be allotted between the capture and recapture to allow the animals to become completely dispersed throughout the population. Assumption (i)-(iii) together reflect the requirement that the sample is random. It is also important to note that, for simplicity, our case assumes the population has not changed significantly due to births or deaths, as well as immigration or emigration.

3.2 Acceptance Functions

The formula for the acceptance functions for estimating N is the same as the one used in Chapter 2, however M is now fixed and they become functions of N :

$$AF_{\mathcal{Q}}(N) = P(X \in \mathcal{Q}) = \sum_{x \in \mathcal{Q}} \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}},$$

such that $\mathcal{Q}$ is any subset of $\{0, 1, 2, \dots, n-2, n-1, n\}$ of the form $\mathcal{Q} = \{l, l+1, l+2, \dots, u-2, u-1, u\}$.

The acceptance functions behave much differently as functions of N as opposed to when they are functions of M . We first notice that small values of x , the number of

marked units in our recapture, yield large values of N . Therefore, the order of our acceptance functions is switched in comparison to the case of estimating M . Rather than acceptance functions of the form $AF_{\Omega_{0l}}$ starting from the left and moving to the right as l increases, they now start from the right and move to the left. On the contrary, acceptance functions of the form $AF_{\Omega_{un}}$ now start from the left and move to the right as u decreases. Figure 12 shows the acceptance functions generated by the formula above, along with the labeling of a few Type O acceptance functions.

Figure 12. Acceptance functions that lie above $1 - \alpha = 0.90$ for the case when $n = 7$ and $M = 20$.

Recall that the case of estimating M , the success parameter, was equivariant because the definition of “success” and “failure” were arbitrary and could be switched, while obtaining the same consequence. The parameter N has no such property and therefore will not have symmetry between the acceptance functions. Figure 12 shows this lack of symmetry and, in fact, we see that the functions are very narrow and close together for small values of N and become more spread apart as N increases. This is the typical nature that we see of acceptance functions as a function of N .

3.3 Selection of Acceptance Functions

The selection process of acceptance functions for estimating N is the same as in Chapter 2, with the only difference being that we start from large values of N and move to the left until we reach $\max\{n, M\}$. The reason for doing so is because we wish to obtain intervals for x starting at $x = 0$. Since the order of acceptance functions is switched from the case in Chapter 2, we must start from the right. Then we can apply the same method as in Chapter 2 to determine confidence intervals for the hypergeometric parameter N . Recall that the minimal span procedure chose acceptance functions with minimal span and highest coverage probability.

Since N may go to infinity, we must determine where to start examining the acceptance functions that are eligible for selection at each N . The most obvious value of N to start at would be the largest value of N for which one of the acceptance functions crosses the predetermined confidence level. In fact, $AF(0 - 0)$ is always the last acceptance function to cross the confidence level, at say N_0 , and remains above the confidence level for all $N \geq N_0$. Since $AF(0 - 0)$ clearly has minimal span, we will use this acceptance function at all $N \geq N_0$. Thus, the resulting confidence interval for the minimal span procedure of $x = 0$ will contain $[N_0, +\infty)$.

Take Figure 13, for example, which shows a different portion of acceptance functions for the same case as Figure 12, but for $100 \leq N \leq 6,000$. We see that $AF(0 - 0)$ is indeed the last acceptance function to cross the confidence level. This function becomes greater than or equal to the predetermined confidence level when $N_0 = 1,342$. Thus, $AF(0 - 0)$ will be the acceptance function selected for all $N \geq 1,342$ because no

other acceptance function can have a smaller span, and it maintains sufficient coverage probability.

Figure 13. Acceptance functions eligible for selection for $N \geq 100$ for the case when $n = 7$, $M = 20$, and $1 - \alpha = 0.90$.

After determining N_0 , we follow the same minimal span confidence procedure as Chapter 2 for all $N \in \{N_0 - 1, N_0 - 2, \dots, \max\{n, M\} - 1, \max\{n, M\}\}$. Figure 14 shows the section where $N \in \{20, 21, \dots, 89, 90\}$, which was highly compressed in Figure 12, with points where the function is defined and lines connecting the defined points for each acceptance function for clarity. The minimal span acceptance functions have been highlighted in blue.

Figure 14. Acceptance functions above the 90% confidence level for all $N \in \{20, 21, \dots, 79, 80\}$ for the case $n = 7$ and $M = 20$. Functions with minimal span at select values of N are highlighted in blue.

Referring to Figure 14, we see that $AF(0 - 3)$, with span three, is the only minimal span acceptance function between $80 \geq N \geq 70$ and switches to $AF(1 - 4)$ for $69 \geq N \geq 51$. Notice in both these cases we are not considering $AF(1 - 5)$ because it has a span of four, while $AF(0 - 3)$ and $AF(1 - 4)$ have a span of three. Since we prioritize minimal span, $AF(1 - 5)$ is not selected until $N = 50$, when it is the only option for a minimal span acceptance function until $AF(2 - 6)$ comes into play at $N = 46$. However, $AF(1 - 5)$ obtains a higher coverage probability than $AF(2 - 6)$ for $46 \geq N \geq 40$, so we continue to use it until the function $AF(2 - 5)$, with smaller span, rises above the confidence level. Acceptance functions for the remaining values of N are selected in a similar fashion. The resulting confidence intervals are shown in Table 4.

x	Lower Limit	Upper Limit
0	70	$+\infty$
1	43	1,341
2	35	250
3	28	114
4	25	69
5	23	50
6	21	36
7	20	27

Table 4. 90% confidence intervals for N when $n = 7$ and $M = 20$.

3.4 Resolving the Issue of Gaps

Gaps occur for the parameter N for the same reason as they do in Chapter 2, and will be resolved in the same manner. We find that gaps occur much more often in the case of estimating N than they do in the case of estimating M . Testing for gaps was performed for all $n \leq M$ for each associated $M \in \{1, 2, \dots, 49, 50\}$ at the $1 - \alpha = 0.90$, 0.95 , and 0.99 levels and it was found that gaps occurred 262 times for some x out of the total 3,825 cases (6.85%). Gaps may occur more often in the case of estimating N due to the asymmetry of the acceptance functions, as well as the fact that the acceptance functions span over a large set of M values.

Recall from Chapter 2 that decreasing values of l or u resulted in confidence sets for M that are not intervals. In the case of estimating N , we must observe each l and u value starting at N_0 and moving in a decreasing order of N until $\max\{n, M\}$ is reached. We will again require that the l and u sequences are nondecreasing, otherwise a gap arises and we must resolve it.

To illustrate the gap issue in the case of estimating N , consider Figure 15, which shows the acceptance functions that lie above $1 - \alpha = 0.90$ for the case $n = 10$ and $M = 15$. That is, the first capture consisted of 15 units while the recapture consisted of 10. Figure 15(a) displays the acceptance functions selected by our initial procedure for the values of N such that $75 \geq N \geq 50$. Looking at the graph from right to left, we see that $AF(1 - 4)$ is selected for $N = 63$ and 64. The next acceptance function that is selected is $AF(0 - 4)$, which violates the required monotonicity in l , leading to a gap at $N = 62$.

When the choice of an acceptance function results in a gap, we again use the minimal span acceptance function with the next highest coverage probability in substitution. Figure 15(b) shows the resolution to the gap issue, where the dotted portion of $AF(0 - 4)$ represents the acceptance function in violation of monotonicity and an extension of $AF(1 - 5)$ has been added in bold, which now becomes the assigned acceptance function.

Figure 15. (a) – (b) Acceptance functions selected for the case $M = 15$, $n = 10$, and 90% confidence level. (a) Shows the initial procedure leads to a gap at $N = 62$. (b) Resolution to the gap issue.

3.5 Formal Description for Estimating N

In general, the method of producing confidence intervals for the hypergeometric parameter N selects acceptance functions of minimal span with highest coverage probability beginning at N_0 and then proceeds in the same fashion in a decreasing order of N until $\max\{n, M\}$ is reached. In the event that gaps occur, we substitute the acceptance function of minimal span with the one having the next highest coverage probability.

The following is a formal description of the algorithm used to find the $1 - \alpha$ level minimal span confidence procedure for the hypergeometric parameter N :

Step 1: Locate the starting point, N_0 , of evaluating acceptance functions. Here, $N_0 = \min\{N \mid AF(0 - 0) \geq 1 - \alpha\}$.

Step 2: For each N , beginning with N_0 and continuing in a decreasing fashion until $\max\{n, M\}$ is reached, let $AF_N(l - u)$ denote the acceptance function achieving the highest coverage probability above $1 - \alpha$ among all acceptance functions of minimal span at N . If more than one acceptance function achieves the highest coverage probability, select the function $AF_N(l - u)$ with the largest value of l . Assign N to the confidence intervals for each $x \in [l, u]$, except in the case described by Step 3.

Step 3: Whenever $AF_N(l - u)$ and $AF_{N-1}(l' - u')$ from Step 3 are such that $l' < l$ and/or $u' < u$, let k be the largest integer that produces $AF_{N-k}(l' - u')$. Reassign $N - 1, \dots, N - k$ to $x \in [l' + 1, u' + 1]$.

3.6 Alternate Methods

There are several formula-based confidence procedures for estimating N . For example one formula provided by Thompson (2002) is

$$C(x) = \frac{nM}{x} \pm z_{\alpha/2} \left(\frac{Mn(M-x)(n-x)}{x^3} \right)^{1/2}.$$

However, it is clear that an issue arises in the event that $x = 0$. The default confidence interval in this case would be $[\max\{n, M\}, +\infty)$, which is the widest the interval could possibly be and effectively useless. Many attempts have been made to improve confidence intervals for the parameter N , but none of them achieve the desired minimum coverage probability.

Wang (2015) provides confidence intervals for the hypergeometric parameter N that have a minimum coverage probability of $1 - \alpha$. Since Wang's procedure is the only exact procedure previously developed, we provide an explanation of his method for estimating N along with a comparison to the minimal span procedure.

3.6.1 A Description of Wang's Method

Wang uses a similar approach for estimating N as he does with estimating M . He first starts with one-sided confidence intervals, which are based on Type O acceptance functions, to obtain two-sided confidence intervals and then improves upon said intervals. The only difference in the case of estimating N is that the acceptance functions are not symmetric, and we must look at acceptance functions of the form $AF(l - \min\{n, M\})$ in addition to acceptance functions of the form $AF(0 - u)$.

Consider Figure 16, which shows the Type O acceptance functions of the form $AF(l - \min\{n, M\})$ that lie above $1 - \alpha = 0.95$ for the case when $n = 7$ and $M = 20$. The

shortest upper one-sided confidence interval for each x will have a lower bound where $AF((x + 1) - 7)$ drops below the desired confidence level. Take for example, $x = 3$ in Figure 16. The function $AF(4 - 7)$ drops below 0.95 at $N = 28$. Thus, the lower bound is 28, and $x = 3$ is contained in the rest of the acceptance sets as we move to the right. Therefore the upper one-sided confidence interval of x is $[28, +\infty)$.

Figure 16. Type O acceptance functions of the form $AF(7 - l)$ for the case when $n = 7$, $M = 20$, and $1 - \alpha = 0.90$.

We find the lower one-sided confidence intervals of x , now, by using acceptance functions of the form $AF(0 - u)$. For optimal viewing, Figure 6 displays the acceptance functions of the form $AF(0 - u)$ only for $20 \leq N \leq 170$. The shortest lower one-sided confidence intervals for each x will have an upper limit where $AF(0 - (x - 1))$ drops below the desired confidence level. For example, the upper limit of the confidence interval for $x = 3$ would be the value of N just before where $AF(0 - 2)$ reaches above 0.95. Figure 17 shows that this occurs when $N = 150$, and $x = 3$ is contained in the rest of the acceptance sets as we move to the left. Thus, the resulting lower one-sided confidence interval is $[\max\{n, M\}, 150] = [20, 150]$.

Figure 17. Acceptance functions of the form $AF(0 - u)$ between $20 \leq N \leq 170$ that lie above $1 - \alpha = 0.95$ for the case $n = 7$ and $M = 20$.

Wang's two-sided confidence intervals are constructed from the previously found one-sided confidence intervals for each x and improved upon using an iterative algorithm. Wang first finds the $1 - \alpha/2$ one-sided confidence intervals to determine the initial upper and lower bounds for his $1 - \alpha$ two-sided confidence intervals, just as he did with estimating M . For example, given the 95% one-sided intervals for $x = 3$ in Figures 16 and 17, his corresponding 90% initial confidence interval would be $x \in [28, 150]$.

Wang then uses an algorithm to shorten the length of these intervals while maintaining coverage probability above $1 - \alpha$, beginning with the confidence interval for $x = 0$ and ending with $x = \min\{n, M\}$. Just as the case of estimating M , Wang's algorithm increases the lower limit of each confidence interval by an integer and decreases the upper limit by another integer, when possible, until no shorter confidence interval with coverage probability above $1 - \alpha$ can be found.

3.7 A Comparison

In contrast to the graphical approach of the minimal span confidence procedure, Wang's method successively modifies confidence intervals based on maintaining coverage probability of at least $1 - \alpha$. Just as the case of estimating M , Wang's method typically leads to the use of minimal span acceptance functions.

Table 5 displays the 90% confidence intervals for the minimal span procedure and Wang's procedure for the case when $n = 25$ and $M = 50$ with coloring to represent when the intervals differ. Intervals written in red represents when Wang's intervals are shorter than those produced by the minimal span procedure. Intervals written in green represents when the minimal span procedure's intervals are shorter than those produced by Wang. Blue represents when the intervals are different, but both procedures produce an interval of the same length.

Although both intervals for $x = 0$ are infinite, they are highlighted in red because we see that Wang's lower limit is 521 while the minimal span's lower limit is 495. The lower limit of Wang's procedure cuts off more values of N , so for the purpose of comparing, we consider Wang's interval as being shorter in this particular case. We notice that, in general, Wang's method produces shorter length intervals for very small values of x , while the minimal span procedure results in shorter intervals more often elsewhere.

x	Wang's Procedure	Minimal Span Procedure
0	[521, ∞)	[495, ∞)
1	[324, 11901]	[312, 11901]
2	[240, 2316]	[228, 2316]
3	[198, 1102]	[193, 1102]
4	[162, 688]	[162, 688]
5	[141, 520]	[140, 494]
6	[127, 422]	[126, 422]
7	[115, 323]	[114, 311]
8	[105, 269]	[105, 269]
9	[97, 239]	[97, 227]
10	[91, 197]	[91, 192]
11	[87, 168]	[87, 168]
12	[82, 155]	[81, 155]
13	[77, 140]	[76, 139]
14	[72, 126]	[72, 125]
15	[70, 114]	[70, 113]
16	[67, 104]	[66, 104]
17	[63, 96]	[63, 96]
18	[61, 86]	[61, 86]
19	[59, 81]	[59, 80]
20	[57, 76]	[57, 75]
21	[55, 71]	[55, 71]
22	[54, 66]	[54, 65]
23	[52, 62]	[52, 62]
24	[51, 58]	[51, 58]
25	[50, 54]	[50, 54]

Table 5. 90% confidence intervals for Wang's and the minimal span procedure when $n = 25$ and $M = 50$. **Red** represents when Wang's intervals are shorter than those produced by the minimal span procedure. **Green** represents when the minimal span procedure's intervals are shorter than those produced by Wang. **Blue** represents when the intervals are different, but both procedures yield an interval of the same length.

Observing the confidence intervals in Table 5 for both procedures, we see that the intervals are extremely wide for small values of x , and narrow for values of x that are close to the recapture size. We find that this pattern emerges in all cases of estimating N and occurs because there is more uncertainty in the estimate of the population size when x is small. Thus, to maintain the desired confidence level, the intervals naturally widen.

Since the length of the intervals for x varies significantly, there becomes an issue of how to properly compare the two methods. In particular, the intervals for $x = 0$ are infinite, so comparing the average length of all the intervals for x as we did in Chapter 2 would not provide any useful information. Therefore we will only compare the intervals for $x = 1$ to $x = \min\{n, M\}$.

Figure 18 shows the ratio of confidence interval length for the minimal span procedure versus Wang's procedure for each $x > 0$, with a reference line for a ratio equal to one. Wang's procedure has shorter intervals than our method produces when the points lie above the line, and longer intervals when the points lie below the line. Wang's method produces shorter intervals for six out of the 25 intervals, while the minimal span method produces shorter intervals for nine out of the 25 intervals. We can see that when the minimal span method produces shorter intervals, it shortens the intervals by a maximum of 8.5%, while Wang's only shortens the intervals by a maximum of 2.7%. In this example, the average ratio of interval length for the minimal span procedure compared to that of Wang's procedure is 98.3%. Hence, the minimal span method produces intervals that are 1.7% shorter on average for this particular case.

Figure 18. The ratio of length of the minimal span procedure versus Wang's procedure for $x > 0$ for the case when $n = 25$, $M = 50$, and 90% confidence level with a reference line shown when the ratio is equal to one.

The same calculation was performed on the average ratio for all $n \leq M$ for each associated $M \in \{1, 2, \dots, 49, 50\}$ at the $1 - \alpha = 0.90$ confidence level. Wang's procedure obtained shorter average length for only four cases, while the minimal span procedure yielded shorter average length for 1,110 (87.1%) of the 1,275 cases that were tested. In the remaining cases, the average lengths of both procedures were the same. The minimal span procedure was shorter than Wang's by a maximum of 16.67%, while Wang's procedure was shorter by a maximum of 1.36%. On average the minimal span confidence procedure shortened the average length of the confidence intervals by 2.3%.

Figure 19 displays the coverage probability for both procedures for the case $n = 25$ and $M = 50$ considered previously, shown in red where Wang's method differs from the minimal span procedure. Figure 19(a) shows the coverage probability function, truncated at $N = 30,000$ while Figure 19(b) shows a condensed portion of 19(a) in order to better view the differences between the two procedures.

Figure 19. (a) – (b) Coverage Probability of the minimal span procedure along with Wang's shown in red where it differs for the case $n = 25$ and $M = 50$.
(a) displays coverage probability for $20 \leq N \leq 30,000$.
(b) shows the coverage probability for a dense area of (a), $20 \leq N \leq 800$.

In addition to the advantage of having shorter average length as discussed earlier, Figure 19 shows that in many places the minimal span procedure has higher coverage probability. This is a typical result that we see when we compare the coverage of Wang's procedure with the minimal span procedure for interval estimation of N . We observe that Wang's procedure uses the same minimal span acceptance functions as our method in this example, but transitions between them at different values of N , resulting in considerable differences between the confidence intervals, as well as the coverage probability function.

3.8 Other Criteria

Just as in the problem of estimating M discussed in Chapter 2, there are certain properties that we would like to see in a viable confidence procedure for estimating N .

Firstly, we would like to see monotonicity in the limits of the confidence interval as x increases with n fixed, when n increases with x fixed, as well as when both n and x increase. We will also investigate whether there are violations in nesting.

It is natural to assume that the number of successes—the number of marked animals in the recapture for the ecological application—has a direct relationship with the limits of the confidence intervals. Since small values of x yield large values of N and vice versa, we expect the lower and upper confidence limits for each x of a given confidence procedure to decrease as x increases. This property was tested for all $n \leq M$ for each associated $M \in \{1, \dots, 49, 50\}$ at the 90%, 95%, and 99% confidence levels and no violations to monotonicity were observed.

We would also like to observe the monotonicity behavior of the limits of the confidence intervals when an additional trial is added. There are two situations: (i) the additional trial results in a failure and (ii) the additional trial results in a success. Again, since small values of x yield large values of N , if the additional trial results in a failure, we would expect both limits of the confidence interval to increase. Note that this condition fails for $x = 0$, where the upper limit is always infinity, and $x = \min\{n, M\}$, where the lower limit is always $\max\{n, M\}$. Thus, for the case of estimating the hypergeometric parameter N , we check that the lower and upper limits of each interval for x are nondecreasing as n increases. This property was testing for all $n \leq M$ for each associated $M \in \{1, \dots, 49, 50\}$ at the 90%, 95%, and 99% confidence levels. We found that violations to this behavior occurred 213 times for some x out of all 3,825 cases (5.57%).

On the other hand, if an additional trial results in a success, we expect both limits of the confidence interval to decrease. This fails to occur when $n < M$ for $x = n$, which has a lower confidence limit of M , since adding an additional trial and success would lead to $x + 1 = n + 1$ also having a lower confidence limit of M . Thus, we tested that the upper and lower limits of the confidence interval for $x + 1$ with $n + 1$ trials are less than or equal to the limits of x with n trials. Testing was performed on the same 3,825 cases discussed above and violations occurred for some x for only 17 cases (0.44%).

Another desirable property that we performed testing on was nesting. Recall that this property describes the behavior that given any two confidence intervals with different confidence levels computed by the same data set, the confidence interval obtained by the higher confidence level contains the interval obtained by the lower confidence level for each value of x . Investigation was performed for all $n \leq M$ for each associated $M \in \{1, 2, \dots, 49, 50\}$ at each confidence level $1 - \alpha \in \{0.900, 0.901, 0.902, \dots, 0.997, 0.998, 0.999\}$. Just as the case of estimating M , violations to nesting only occurred after gaps had been resolved. However every time a gap occurred, there was not always a violation to nesting.

3.9 Applications

As previously discussed, the capture-recapture method is commonly used in ecology to estimate the population of a particular species. The capture-recapture model may also be used in epidemiology to estimate the prevalence of a disease; refer to Gill, Ismail, and Beeching (2001). As an example, we discuss a straightforward application to ecology.

An article published by Pollock et al. (1990) provides an example of capture-recapture data involving bobwhite quail. 148 quail were captured and marked ($M = 148$), while the recapture consisting of 82 ($n = 82$) birds total included 39 ($x = 39$) marked. The article used a formula-based confidence interval to obtain $[250, 366]$, a method that doesn't necessarily obtain the desired confidence level of 95%.

Wang (2015) also performs the calculations for this confidence interval and obtains the exact 95% confidence interval $[263, 394]$. While the length of Wang's interval is 131, which is longer than the length of formula-based interval (116), his procedure achieves the desired confidence level. The minimal span procedure also achieves the confidence level and we find that it produces the interval $[261, 391]$, which has a length of 130, thus shorter than Wang's.

Chapter 4. Conclusion

The idea of first constructing a coverage probability function with at least $1 - \alpha$ coverage and deducing the confidence intervals from this function is a new approach to interval estimation that has found substantial success. See, for example, Choi (2015), Schilling and Doi (2014), Schilling and Holladay (2017).

Blaker (2000) has shown that no confidence procedure simultaneously achieves the three properties of (i) length optimality, (ii) nesting, and (iii) confidence sets that are intervals. The minimal span procedure presented in this thesis provides gapless short confidence intervals for the parameters of the hypergeometric distribution, while maintaining high coverage probability. We observe violations to nesting and monotonicity only infrequently.

Our method outperforms a newly developed interval estimation procedure for the hypergeometric distribution due to Wang (2015). In particular, we find that the minimal span procedure is superior in confidence interval length in both the case of estimating M and the case of estimating N , while also achieving higher coverage probabilities.

The advantages of our confidence procedure for estimating M are particularly evident for small values of x , the number of successes. It is due to this result that we find our procedure applies especially well to the field of medicine, where such situations are common.

References

- Blaker, H. (2000), "Confidence Curves and Improved Exact Confidence Intervals for Discrete Distributions," *Canadian Journal of Statistics*, 28:4, 783-798.
- Blyth, C.R. and Still, H.A. (1983), "Binomial Confidence Intervals," *Journal of the American Statistical Association*, 78, 108-116.
- Buonaccorsi, J. (1987), "A Note on Confidence Intervals for Proportions in Finite Populations," *The American Statistician*, 41:3, 215-218.
- Choi, S.A. (2015), "Confidence Intervals for the Success Parameter of the Negative Binomial Random Variable," Master's Thesis, *California State University, Northridge*.
- Cochran, W.G. (1977), *Sampling Techniques* (3rd ed.), New York: Wiley.
- Crow, E.L. (1956), "Confidence Intervals for a Proportion," *Biometrika*, 43, 423 – 425.
- Engelberger, Aujesky, Calanca, Staeger, Hugli, Mazzolai (2011), "Comparison of the diagnostic performance of the original and modified Wells score in inpatients and outpatients with suspected deep vein thrombosis." *Thrombosis Research*, 127:6, 535-539.
- Fine, A., Nizet, V., & Mandl, K. (2012), "Large-Scale Validation of the Centor and McIsaac Scores to Predict Group A Streptococcal Pharyngitis," *Archives Of Internal Medicine*, 172:11, 847-852.
- Gill, G., Ismail, A., & Beeching, N. (2001), "The use of capture-recapture techniques in determining the prevalence of type 2 diabetes," *Qjm-Monthly Journal Of The Association Of Physicians*, 94:7, 341-346.
- Kabaila, P. and Byrne, J. (2001), "Exact Short Poisson Confidence Intervals," *Canadian Journal of Statistics*, 29:1, 99-106.
- Katz, L. (1953), "Confidence Intervals for the Number Showing a Certain Characteristic in a Population when Sampling is Without Replacement," *Journal of the American Statistical Association*, 48, 256-261.
- Kinney, J. (2014), *Probability: An Introduction with Statistical Applications*, Hoboken: Wiley.
- Konijn, H. S. (1973), *Statistical Theory of Sample Survey Design and Analysis*, Amsterdam: North-Holland.

- Pollock, K. H., Nichols, J. D., Brownie, C., and Hines, J. E. (1990), "Statistical Inference for Capture-Recapture Experiments," *Wildlife Society Monographs*, 107, 1–97.
- Sahai, H. and Khurshid, A. (1995), "A Note on Confidence Intervals for the Hypergeometric Parameter in Analyzing Biomedical Data," *Computers in Biology and Medicine*, 25, 35-38.
- Schilling, M.F., and Doi, J.A. (2014), "A Coverage Probability Approach to Finding an Optimal Binomial Confidence Procedure," *The American Statistician*, 68:3, 133-145.
- Schilling, M.F., and Holladay, B. (2017), "Length Minimization for Poisson Confidence Procedures," *Communications in Statistics – Theory and Methods*, 46:2, 861-873.
- Southwood, T. & Henderson, P. (2009), *Ecological Methods*. (3rd ed.), Hoboken: Wiley.
- Streitz, M., Oliver, J., Hyams, J., Wood, R., Maksimenko, Y., Long, B., Barnwell, R., April, M. (2017), "A retrospective external validation study of the HEART score among patients presenting to the emergency department with chest pain," *Internal and Emergency Medicine*, 11 September 2017.
- Thompson, S.K. (2002), *Sampling* (2nd ed.), New York: Wiley.
- Wang, W. (2015), "Exact Optimal Confidence Intervals for Hypergeometric Parameters," *Journal of the American Statistical Association*, 110:512, 1491 – 1499.
- Xu, Eisenberg, Madden, Secura, Peipert. (2014), "Medical contraindications in women seeking combined hormonal contraception," *American Journal of Obstetrics and Gynecology*, 210:3, 210.e1-210.e5.

Appendix

Maximum and Minimum of Type O Functions

Theorem: Type O curves of the form $AF(0 - u)$ achieve a maximum of 1 at $x = 0$ and a minimum of 0 at $x = N$.

Proof: We first show that all functions of the form $AF(0 - u)$ achieve the previously mentioned values. We then prove that acceptance functions of the form $AF(0 - u)$ are monotonically decreasing. Combining the results completes the proof.

First, note that $P(x \leq u)$ is simply $AF(0 - u)$. By a basic property of probability, we know that $0 \leq P(x \leq u) \leq 1$. Thus, we must show that there exists a value of M such that $AF(0 - u) = 0$ and $AF(0 - u) = 1$ to determine that these are the minimum and maximum, respectively.

To show this, consider expanding the acceptance function:

$$AF(0 - u) = \sum_{x \in \{0, \dots, u\}} \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}} = \frac{\binom{M}{0} \binom{N-M}{n}}{\binom{N}{n}} + \frac{\binom{M}{1} \binom{N-M}{n-1}}{\binom{N}{n}} + \dots + \frac{\binom{M}{u} \binom{N-M}{n-u}}{\binom{N}{n}}.$$

Observing the value of $AF(0 - u)$ at $M = 0$, one finds

$$AF(0 - u)_{M=0} = \frac{\binom{0}{0} \binom{N-0}{n}}{\binom{N}{n}} + \frac{\binom{0}{1} \binom{N-0}{n-1}}{\binom{N}{n}} + \dots + \frac{\binom{0}{u} \binom{N-0}{n-u}}{\binom{N}{n}} = 1.$$

Observing the value of $AF(0 - u)$ at $M = N$, one finds

$$AF(0 - u)_{M=N} = \frac{\binom{N}{0} \binom{0}{n}}{\binom{N}{n}} + \frac{\binom{N}{1} \binom{0}{n-1}}{\binom{N}{n}} + \dots + \frac{\binom{N}{u} \binom{0}{n-u}}{\binom{N}{n}} = 0.$$

Now suppose we have two sets, S_1 and S_2 , of equal size such that S_1 has M successes and $N - M$ failures and S_2 has $M + 1$ successes and $N - M - 1$ failures. Now, suppose we mark one failure in S_1 , call it f , and one success in S_2 , call it s . In doing so,

the remaining components of each set are identical and each set has M unmarked successes and $N - M - 1$ unmarked failures.

Let $P_1(x \leq u)$ denote the probability that $x \leq u$ given that a random sample of size n is taken from S_1 , and $P_2(x \leq u)$ denote the probability that $x \leq u$ given that a random sample of size n is taken from S_2 . Note that we are simply comparing the value of the Type O acceptance function, $AF(0 - u)$, at M and $M + 1$.

$P_1(x \leq u)$ can be broken down into two cases: (i) the sample taken from S_1 does not contain f and (ii) the sample taken does contain f . Thus, we have:

$$P_1(x \leq u) = P_1(x \leq u \ \& \ f \notin \text{sample}) + P_1(x \leq u \ \& \ f \in \text{sample}).$$

And similarly:

$$P_2(x \leq u) = P_2(x \leq u \ \& \ s \notin \text{sample}) + P_2(x \leq u \ \& \ s \in \text{sample}).$$

Here, we notice that $P_1(x \leq u \ \& \ f \notin \text{sample}) = P_2(x \leq u \ \& \ s \notin \text{sample})$ since the unmarked units in both sets are identical. Now all that is left to compare is

$P_1(x \leq u \ \& \ f \in \text{sample})$ and $P_2(x \leq u \ \& \ s \in \text{sample})$. By properties of probability,

$$P_1(x \leq u \ \& \ f \in \text{sample}) = P_1(x \leq u \mid f \in \text{sample})P_1(f \in \text{sample})$$

and

$$P_2(x \leq u \ \& \ s \in \text{sample}) = P_2(x \leq u \mid s \in \text{sample})P_2(s \in \text{sample}).$$

$P_1(f \in \text{sample})$ and $P_2(s \in \text{sample})$ are both simply equal to n / N .

Now, accounting for the fact that the marked failure, f , is included in the sample, $P_1(x \leq u \mid f \in \text{sample}) = P_1(x \leq u \mid \text{sample of size } n - 1, \text{ excluding } f)$. Similarly, accounting for the marked success, s , being in the sample, $P_2(x \leq u \mid s \in \text{sample}) = P_2(x \leq u - 1 \mid \text{sample of size } n - 1, \text{ excluding } s)$. Once we acknowledge the marked success and failure, the remaining components of both sets are identical. Thus, comparing

$P_1(x \leq u \ \& \ f \in \text{sample})$ and $P_2(x \leq u \ \& \ s \in \text{sample})$ breaks down to comparing $P(x \leq u)$ and $P(x \leq u - 1)$ from the same set.

Clearly $P(x \leq u) = P(x \leq u - 1) + P(x = u) \geq P(x \leq u - 1)$. Recall that $P(x \leq u)$ is associated with $AF(0 - u)$ at M and $P(x \leq u - 1)$ is associated with $AF(0 - u)$ at $M + 1$. Thus, we have the result that Type O functions of the form $AF(0 - u)$ are monotonically decreasing for all M . □

A similar approach can be used to show that acceptance functions of the form $AF(l - n)$ are monotonically *increasing*. It is easy to show that $AF(l - n)$ attains the value 1 at $M = N$ and the value 0 at $M = 0$, thus establishing the maximum and minimum of $AF(l - n)$.