Approximate Confidence Intervals for a Parameter of the Negative Hypergeometric Distribution

Lei Zhang[1], William D. Johnson[2]

1. Office of Health Data and Research, Mississippi State Department of Health, 570 East Woodrow Wilson, Jackson, MS 39215-1700
2. Pennington Biomedical Research Center, Louisiana State University System, 6400 Perkins Road, Baton Rouge, LA 70808

ABSTRACT

The negative hypergeometric distribution is of interest in applications of inverse sampling without replacement from a finite population where a binary observation is made on each sampling unit. Thus, sampling is performed by randomly choosing units sequentially one at a time until a specified number of one of the two types is selected for the sample. Assuming the total number of units in the population is known but the number of each type is not, we consider the problem of estimating this unknown parameter. We investigate the maximum likelihood estimator and an unbiased estimator for the parameter. We use the method of Taylor's series to develop five approximations for the variance of the parameter estimators. We then propose five large sample confidence intervals for the parameter. Based on these results, we simulated a large number of samples from various negative hypergeometric distributions to investigate performance of three of these formulas. We evaluate their performance in terms of empirical probability of parameter coverage and confidence interval length. The unbiased estimator is a better point estimator relative to the maximum likelihood estimator as evidenced by empirical estimates of closeness to the true parameter. Confidence intervals based on the unbiased estimator tended to be shorter than two competitors because of its relatively small variance estimator but at a slight cost in terms of coverage probability.

Key Words: Confidence interval, Empirical coverage probability, Inverse sampling, Large sample theory.

1. INTRODUCTION

The negative hypergeometric distribution, also known as the inverse hypergeometric, or hypergeometric waiting-time distribution, has many useful applications in public health research. The probability distribution function is a discrete probability model that was first described by Wilks (1963), discussed by Moran (1968) and Johnson and Kotz (1969), and further developed by Guenther (1975). Expressions for the mean and variance of the negative hypergeometric distribution are well known. Discrete distributions, such as the binomial, geometric, Poisson, and negative binomial, are discussed in most introductory mathematical statistic books, but the negative hypergeometric distribution has not often appeared in such texts or in peer-reviewed literature. Piccolo (2001) recently derived some approximations for the asymptotic variance of the maximum likelihood estimator for the parameter of the negative hypergeometric distribution. Zelterman (2004) presented some variations of the negative hypergeometric distribution.

In this paper, we use the method of Taylor's series to develop approximations for the variance of estimators of a parameter of the negative hypergeometric distribution. We then propose five large sample confidence intervals for the parameter. We simulated a large number of samples from various negative hypergeometric distributions to investigate performance of three confidence intervals based on these results. We evaluated their performance in terms of empirical probability of parameter coverage and interval length for three formulations of confidence intervals. We begin in Section 2 with an overview of the salient characteristics of the distribution.

2. THE NEGATIVE HYPERGEOMETRIC DISTRIBUTION

Consider an urn that contains a total of N balls where R of these balls are red and B are blue. Suppose we wish to select a random sample from the urn and observe the number of balls of each color in the selected sample. Our goal might be, for example, to estimate the number of red balls in the urn where N is known and R (hence, B) is not.

Suppose the balls are well mixed in the urn and a given trial of an "experiment" is as follows: we randomly select a ball from the urn, observe the ball's color, and place it on the side; we then randomly select a second ball, and place it aside; and we continue to randomly draw from the total of N balls, sampling without replacement, until we obtain a fixed number of red balls (successful balls), denoted as r, where $r \in \{1, 2, \ldots, R\}$. Let $X \in \{0, 1, \ldots, B\}$ denote the number of blue balls that must be drawn to get r red balls. Note that we stop selecting balls when the r^{th} red ball is chosen so that some permutation of $r - 1$ red balls and x blue balls will be chosen in the first $r + x - 1$ selections and the last ball drawn will always be red. Let A_1 be the event that $r - 1$ red balls are drawn in $r + x - 1$ trials and let A_2 be the event that the r^{th} red ball is drawn at the $(r + x)^{th}$ trial given that event A_1 has occurred. Now, the probability $X = x$ is

$$P(X = x) = P(A_1) \times P(A_2 \mid A_1)$$

This can be expressed as

$$P(X = x) = \left[\frac{\binom{R}{r-1}\binom{N-R}{x}}{\binom{N}{r+x-1}} \right] \frac{R-r+1}{N-r-x+1} \quad , \quad x \in \{0, 1, \ldots, N-R\}.$$

We refer to this expression as the probability distribution function (pdf) for the random variable X. For given N, R and r, we refer to the non-zero probabilities determined by the pdf for all values in the domain of the random variable, together with the corresponding values of the random variable that occur with these non-zero probabilities, as the negative hypergeometric distribution. Negative hypergeometric distributions are skewed to the left when $R < B$ and to right when $R > B$, but when R and B are approximately equal, the probability distributions are close to being bell-shaped and resemble a normal distribution.

Theorem 2.1 Let X denote a random variable that has a negative hypergeometric distribution as defined earlier. Let X denote the number of unsuccessful draws observed before obtaining r red balls. Then the expected value and variance of X are, respectively,

$$\mu_x = E(X) = \frac{rB}{R+1} \quad \text{and,}$$

$$\sigma_x^2 = V(X) = \frac{rB(R-r+1)(N+1)}{(R+2)(R+1)^2}$$

3. ESTIMATION

We call attention to the estimation problem for two situations:

1. R is a known integer and N is an unknown integer that we wish to estimate.
2. N is a known integer and R is an unknown integer that we wish to estimate.

Both situations are relevant in many applied problems. The first arises in capture-recapture problems [Bailey (1952)]. This paper investigates the second issue.

A heuristic *point estimator* of R is $\hat{R} = N(r/(r+x))$. However, this estimator may yield non-integer estimates. This concern is addressed as follows.

Theorem 3.1: Let the estimator $\hat{R}_m$ be the greatest integer such that

$$\frac{r}{r+x}N \le \hat{R}_m < \frac{r}{r+x}N + 1, \quad \text{then } \hat{R}_m \text{ is the maximum likelihood estimator (MLE) for } R.$$

Guenther (1975) mentioned the MLE, but our result appears to differ from his in the manner of determining the integer for the final estimate. We verified our result numerically by iteratively solving for maximum likelihood estimates for a variety of parameters of the distribution. For example, let $r = 15$, while R takes values from the set $\{0, 1, \ldots, 100\}$ for a specific x. Given that a specific sample yields $x = 0$, the possible values for the likelihood, denoted prob_x, are plotted against corresponding values of R in Figure 3.1. We see that the likelihood has its greatest value when $R = 100$; hence, if a specific sample yields $x = 0$, the MLE is 100. Similarly, as shown in Figure 3.2, if a specific sample yields $x = 5$, the likelihood has its largest value when $R = 75$ so the MLE is 75. Finally, if $x = 25$, the initial calculation yields 37.5 but, as shown in Figure 3.3, the likelihood has its largest value when $R = 38$, so the MLE is 38.

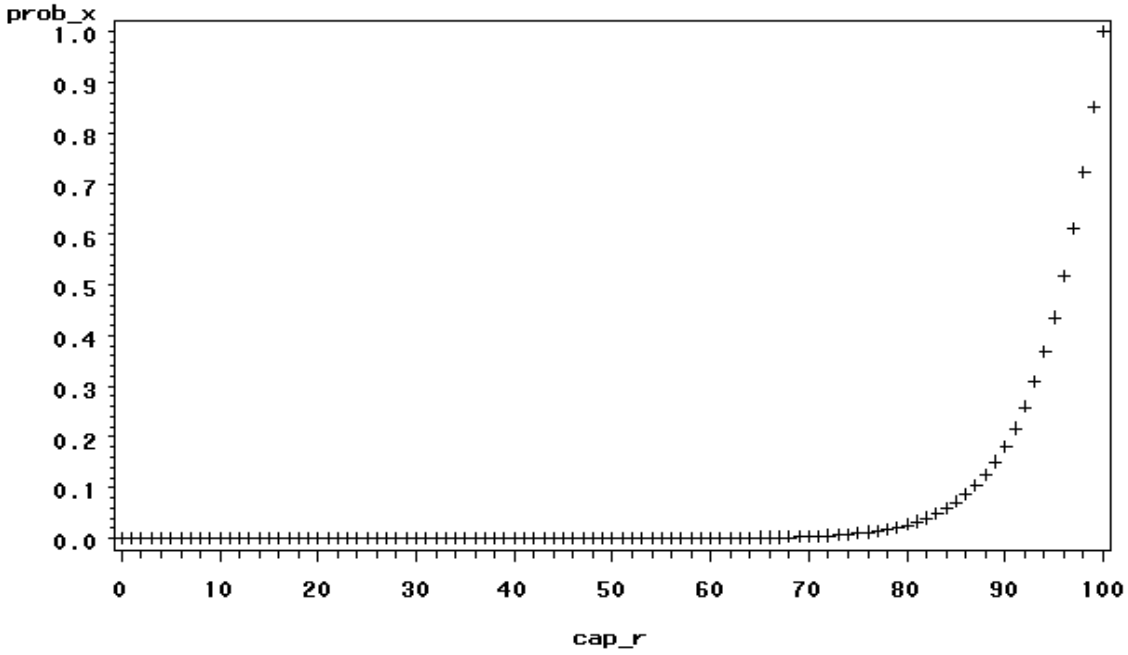

Figure 3.1 MLE for R when n = 100, r = 15, and the sample yields x = 0.

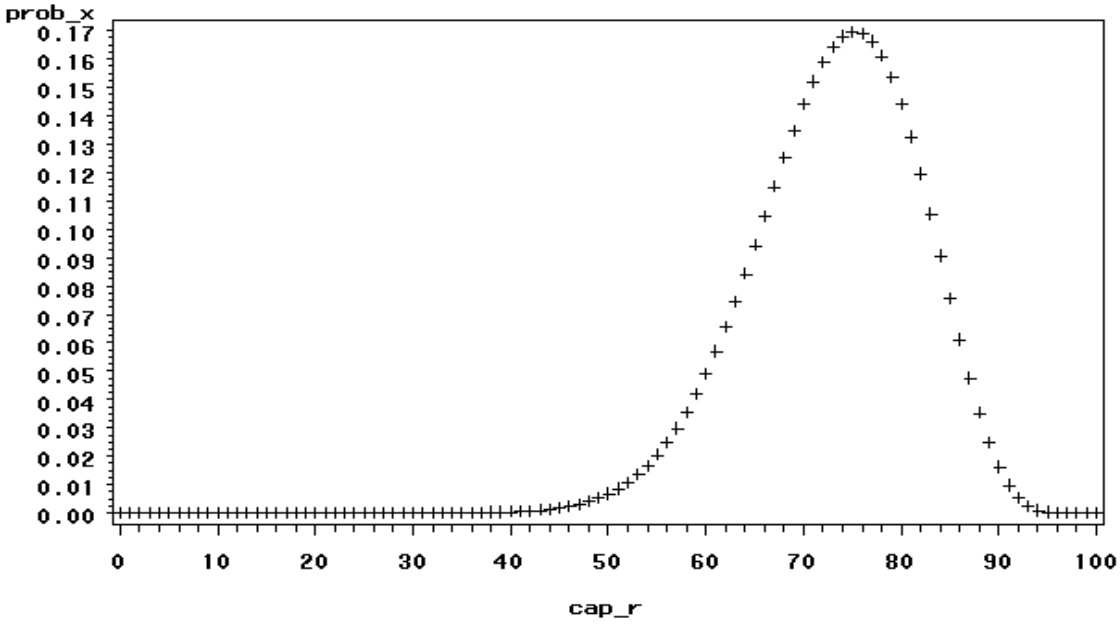

Figure 3.2 MLE for R when n = 100, r = 15, and the sample yields x = 5.

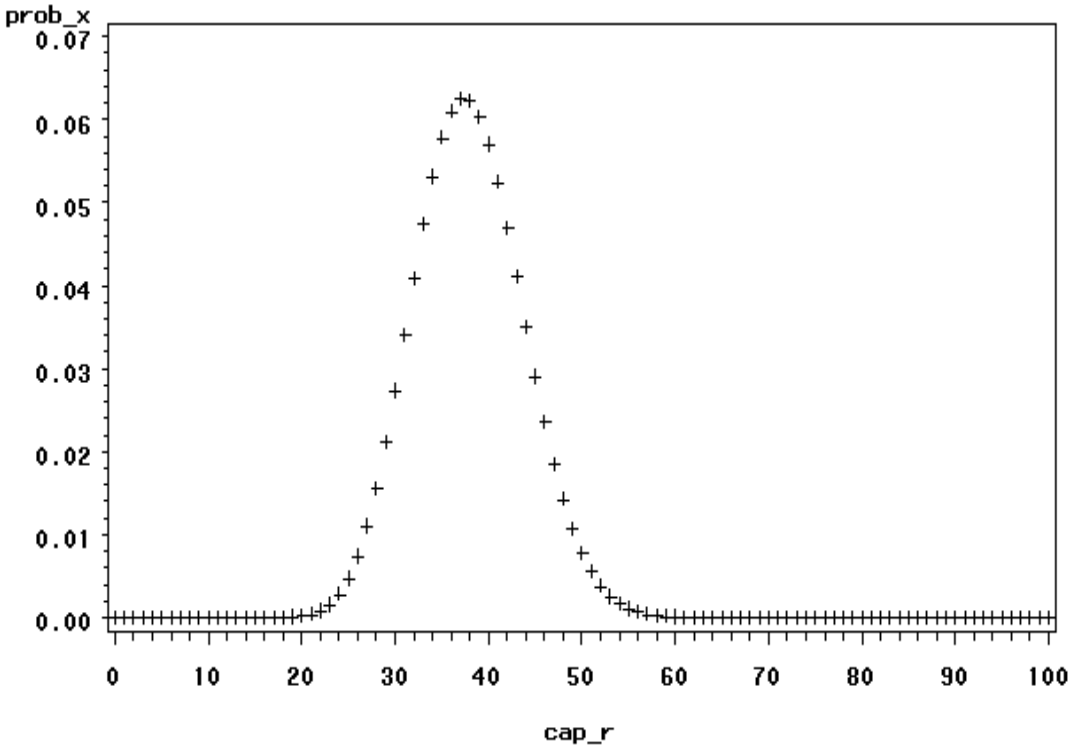

Figure 3.3 MLE for R when n = 100, r = 15, and the sample yields x = 25.

Although MLE's have well known and useful large sample properties, we often prefer unbiased estimators that are functions of MLE's where the functions carry the asymptotic properties. We can easily show that the estimator given in the following theorem is unbiased as claimed by Guenther (1975).

Theorem 3.2: The estimator $\hat{R}_u = \dfrac{r-1}{r+x-1}N$ is an unbiased estimator for *R*.

4. APPROXIMATION FORMULAS FOR VARIANCE OF *ESTIMATORS*

We note that $\hat{R}_u = f(x)$ and use the Taylor series method to find an estimator for the variance of the unbiased estimator given above. Thus,

$$V\left[f(x)\right] \approx \left[f'(x)\right]^2\Big|_{x=E(X)} V(X)$$

or,

$$V\left(\hat{R}_u\right) \approx \frac{(r-1)^2 N^2 (R+1)^2 r(N-R)(N+1)(R-r+1)}{(\mathrm{R}+2)(rN-R+r-1)^4}$$

If we do not know *R*, we can substitute $\hat{R}_u$ to for *R*, in which case we find

$$V\left(\hat{R}_u\right) \approx \frac{(r-1)^2 N^2 (\hat{R}_u+1)^2 r\left(N-\hat{R}_u\right)(N+1)(\hat{R}_u-r+1)}{(\hat{R}_u+2)(rN-\hat{R}_u+r-1)^4}$$

For large samples, both the MLE and unbiased point estimators for R have approximately normal sampling distributions. So a $100\times(1-\alpha)\%$ confidence interval (CI) based on the unbiased estimator is:

$$\hat{R}_u \pm Z_{\alpha/2}\sqrt{V(\hat{R}_u)}$$

$$= \hat{R}_u \pm Z_{\alpha/2}\frac{N(r-1)\left(\hat{R}_u+1\right)}{\left(rN-\hat{R}_u+r-1\right)^2}\sqrt{\frac{r\left(N-\hat{R}_u\right)\left(\hat{R}_u-r+1\right)(N+1)}{\hat{R}_u+2}} \qquad (4.1)$$

When N and R are very large, we have $N+1\approx N$ and $R+1\approx R+2\approx R$, so an approximation to the above CI is

$$\hat{R}_u \pm Z_{\alpha/2}\frac{N(r-1)}{\left(rN-\hat{R}_u+r\right)^2}\sqrt{N\hat{R}_u r\left(N-\hat{R}_u\right)\left(\hat{R}_u-r\right)}$$

To obtain an interval estimate, we need to have $r \le R$. If $r > R$, we always have to draw all the balls (N) because it is impossible to observe the specified number of red balls. In this case, we observe the exact value of R, so an interval estimate is not required. Further, when an estimate of R results in $\hat{R}=N$, the CI reduces to a point estimate. This occurs when $x = 0$ and the resulting point estimate may be undesirable because such an estimate may occur when $R \neq N$ as is implied in this circumstance. For example, we may observe $x = 0$ by choosing r red balls on the first r selections, giving $\hat{R}_u = N$ even when there is at least one blue ball in the urn. To circumvent our dilemma with this happening, we arbitrarily substituted $x + 0.1$ in computing $\hat{R}_u$ for use in the formula for $\hat{\sigma}\left(\hat{R}_u\right)$. Our simulation results support our use of this modification because we obtained excellent empirical coverage when $r = 3, 5, 7$ despite having found numerous samples with $x = 0$.

Following an approach similar to that used above leads to a CI based on the MLE of R. That is

$$\hat{R}_m \pm z_{\alpha/2}\frac{N\left(\hat{R}_m+1\right)}{N+1}\sqrt{\frac{\left(N-\hat{R}_m\right)\left(\hat{R}_m-r+1\right)}{r(N+1)\left(\hat{R}_m+2\right)}} \qquad (4.2)$$

or the simplified approximation

$$\hat{R}_m \pm z_{\alpha/2}\sqrt{\frac{\hat{R}_m\left(N-\hat{R}_m\right)\left(\hat{R}_m-r\right)}{rN}}$$

To avoid producing point estimates for CI's using these two formulas when we find $r + x = N$, which may occur by choosing the r^{th} red ball on the N^{th} selection so that $\hat{R}_m = r$, we again arbitrarily substituted $x + 0.1$ to ensure obtaining an interval estimate.

Let Y = r + X denote the total number of balls that must be drawn to get r red balls and further let $\theta = R/N$ where *R* and *N* are both large so that $R+1 \approx R,\ R+2 \approx R,$ and $N+1 \approx N.$ If, in addition, r is small relative to *R*, then

$$E(Y) \approx \frac{rN}{R} = \frac{r}{R/N} = \frac{r}{\theta}$$

and

$$V(Y) \approx \frac{rB(R)(N)}{R^3} = \frac{r(N-R)/N}{(R/N)^2} = \frac{r(1-\theta)}{\theta^2}$$

That is, under these conditions, the mean and variance of the negative hypergeometric distribution, respectively, are approximately equal to the mean and variance of the negative binomial. Here, an approximate confidence interval is

$$\frac{Nr}{y} \pm z_{\alpha/2}\frac{Nr}{y}\sqrt{\frac{1}{r}-\frac{1}{y}} \tag{4.3}$$

If x = 0 so that y = r, we again substitute x + 0.1 for x as in the above.

5. NUMERICAL EXAMPLE

The negative hypergeometric distribution is relevant in planning sample surveys that use the method of random digit dialing. For a complex sample design, the way the sampling is conducted determines the primary sampling unit (PSU). Consider a sampling frame comprised of a list of telephone numbers that is a mixture of residential and non-residential telephone numbers. Researchers often randomly sample "one at a time" a sequence of telephone numbers (PSU's) from a "bank" of 100 numbers (the sampling frame) and calls these numbers until a specified quantity of residential households is contacted. Researchers may need to estimate the expected or average number of calls required before reaching the specified quantity of residential numbers. The requirements may specify a point estimate or an interval estimate. It is easy to see the analogy between this problem and the model that uses this inverse sampling method to select balls from an urn as described earlier. The negative hypergeometric distribution provides a useful framework for developing a theory for estimation in both applications.

Suppose *N* = 100 and *r* = 15 are known, but *R* is unknown and we want to estimate *R*. Further, suppose in a given 100 bank, we find *y* = 21 total calls are required to reach *r* = 15 residential numbers (i.e., we observe *x* = 6 nonresidential numbers before finally observing the 15th residential number). Using the unbiased estimator for *R*, we get $\hat{R}_u = 70$. An estimate of the standard error of the unbiased point estimator $\hat{R}_u$ is $\hat{\sigma}(\hat{R}_u) = 8.96$. On constructing a 95% CI, we find $\hat{R}_u \pm z_{1-\alpha/2}\ \hat{\sigma}(\hat{R}_u) = 70 \pm 18$. In view of our simulation results presented in Section 7, we know the true confidence level is not exactly 95%, but very likely exceeds 90%.

6. DESIGN OF SIMULATION STUDY

To further study point estimators and CI's for *R*, we used SAS 9.1 to simulate random samples from a negative hypergeometric distribution and compute the mean of the estimates based on the unbiased and MLE estimators. We also obtained the empirical estimates of the coverage probabilities and expected lengths for the confidence interval formulas shown in Eq. 4.1-4.3. We used a "population" of size $N = 100$ with parameter R taking one of the values in the set $R_S = \{90, 80, 70, 60, 50, 40, 30, 20\}$ as the number of red balls and $B = 100 - R$ the number of blue balls. For each combination of values in the set of R_S with a value of r ranging from 3 to 25, we generated 10,000 samples. For each sample, we computed three point estimates and three CI's for *R*. In this known environment for the combinations of *R* in the set of R_S and for every sample, we determined whether or not each of the three CI's included the "known" parameter *R*. Finally, we computed the percentage of samples in which the CI included or "covered" the parameter *R*. The result provided empirical estimates of coverage probabilities for CI's and empirical estimates of expected lengths of CI's.

7. SIMULATION RESULTS

7.1 Point Estimator

We judged the quality of point estimators in terms of empirical estimates of the expected differences between the estimators and the true *R*. The point estimator with the smaller empirical estimate of the expected difference was preferred.

1. $R = 90$. The unbiased estimator is a better point estimator compared to the MLE because the majority of the estimates are closer to the reference line $R = 90$. The MLE tended to over estimate *R*, especially when *r* is between 5 and 20 but appeared to begin converging to *R* when $r > 17$.
2. $R = 80, 70, 60$. The estimates based on $\hat{R}_u$ are very close to the reference line $R = 80, 70, 60$, respectively, whereas the estimates based on $\hat{R}_m$ converged to the reference lines as *r* increased (See, for example Figure 7.1).
3. $R = 50$. Figure 7.2 shows that estimates based on $\hat{R}_u$ are very close to the reference line $R = 50$. The estimates based on $\hat{R}_m$ subsequently converged to the reference line as *r* increased.
4. $R = 40, 30$. The estimates based on $\hat{R}_u$ are very close to the reference line, $R = 40, 30$, respectively, regardless of value *r*. The estimates based on $\hat{R}_m$ converged rapidly to the reference lines as *r* increased (see, for example, Figure 7.3).

In conclusion, the unbiased estimator is uniformly closer to *R* compared to the MLE, as expected.

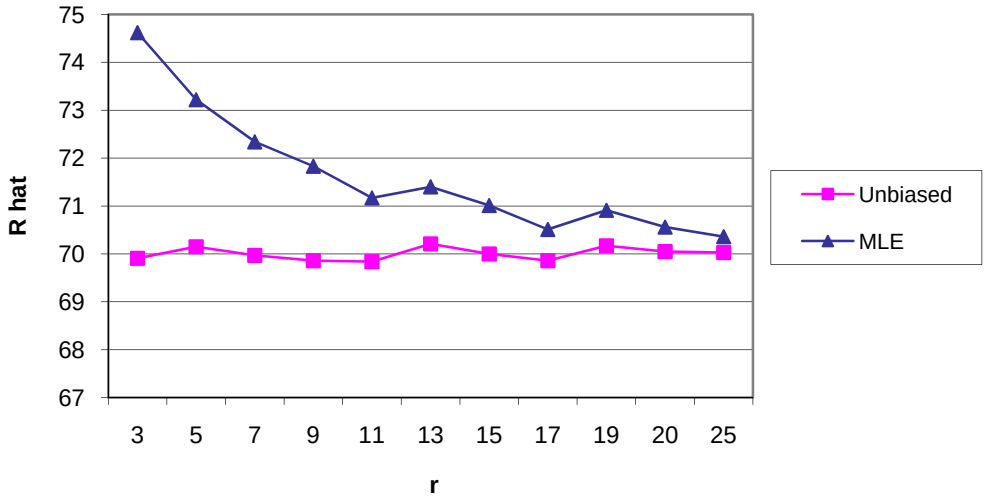

Figure 7.1 Mean value of point estimates for R (n = 100, R = 70, number of replicates = 10,000)

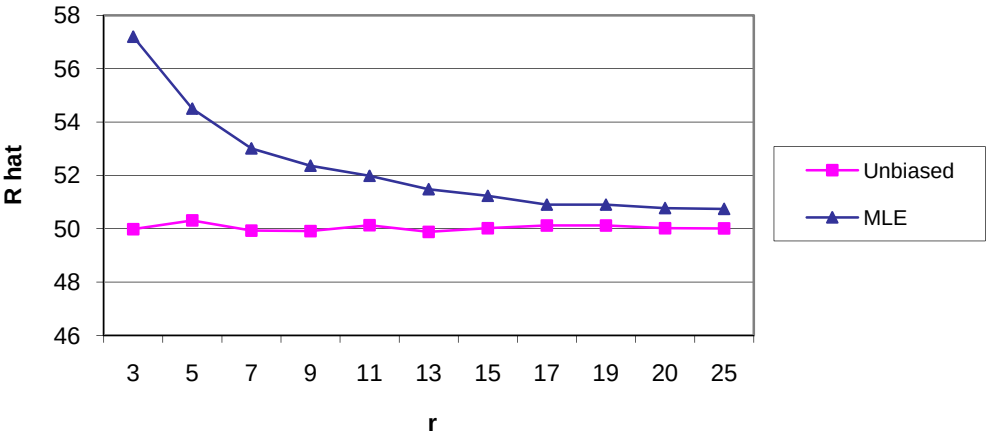

Figure 7.2 Mean value of point estimates for R (n = 100, R = 50, number of replicates = 10,000)

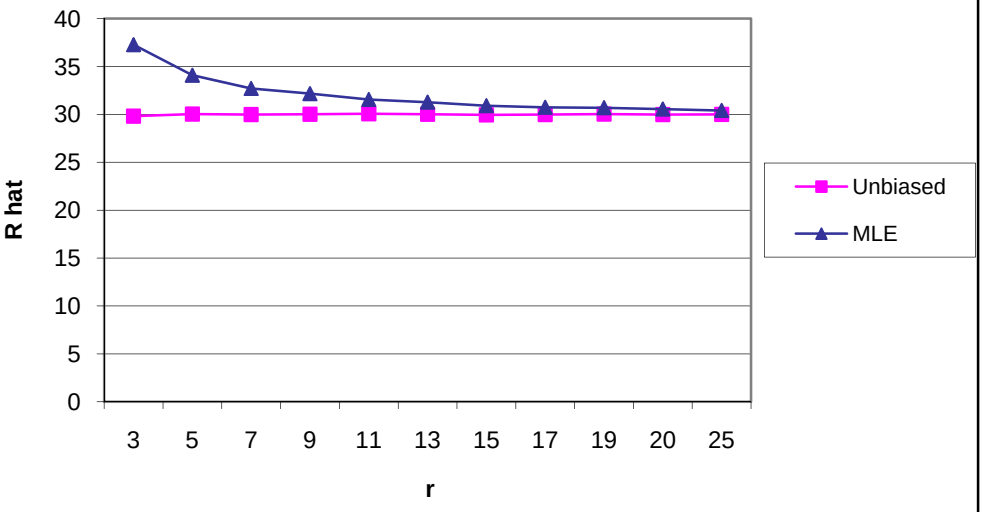

Figure 7.3 Mean value of point estimates for R (n = 100, R = 30, number of replicates = 10,000)

7.2 Empirical Coverage Probabilities for CI's

To construct a CI, we would like the actual coverage probability to be close to the nominal level (i.e., 95% in this discussion). CI's based on large sample theory do not always provide coverage that is exactly equal to the nominal level but, typically, the actual coverage converges to the nominal level as the sample size becomes very large although the rate of convergence varies as the parameters change. Thus, it is desirable to compare the empirical coverage with the specified nominal level for different values of the parameters to determine whether the coverage is sufficiently close to the nominal level for sample sizes that are small enough to be of practical use.

We arbitrarily considered empirical coverage probability between 93% and 97% to be reasonably good performance. We regarded any empirical coverage probability less than 93% to be anti-conservative and any greater than 97% to be conservative. We found:

1. None of the CI's provided adequate coverage when r is very small.
2. None of the CI's performed uniformly best over different values of *r*.
3. In most cases, the estimates of CI coverage based on $\hat{R}_u$ and $\hat{R}_m$ appeared to converge to 95% as *r* increased. The empirical estimates using the unbiased estimator tended to be more anti-conservative while the empirical estimates using the negative binomial approximation tended to be more conservative (See, for example, Figure 7.4).
4. The empirical estimates of CI coverage of *R* tended to be more anti-conservative when *r* is small (e.g., $r < 5$) regardless of type of the estimators and the magnitude of *R*.
5. In most of cases, when *r* is not too small (e.g., $r > 5$) and *R* is less than half the "population" size *N* (e.g., $N = 100$, $R = 30$), the empirical estimates of CI coverage using the MLE tended to have better coverage.

6. In most of cases, when r is not too small (e.g., $r > 5$) and R is about half of the "population" size N (e.g., $N = 100$, $R = 50$), the empirical estimates of CI coverage appeared to be good regardless of the estimator.
7. In most of cases, when r is not too small (e.g., $r > 5$) and R is more than half of the "population" size N (e.g., $N = 100$, $R = 70$), the empirical estimates of CI coverage appeared to be poor regardless of the estimator. Also, the empirical estimates of CI coverage fluctuated as r changed.
8. For a fixed r, especially when r is equal or greater than 7, the empirical estimates of CI coverage decreased as R increased regardless of the estimator (Figure 7.5).

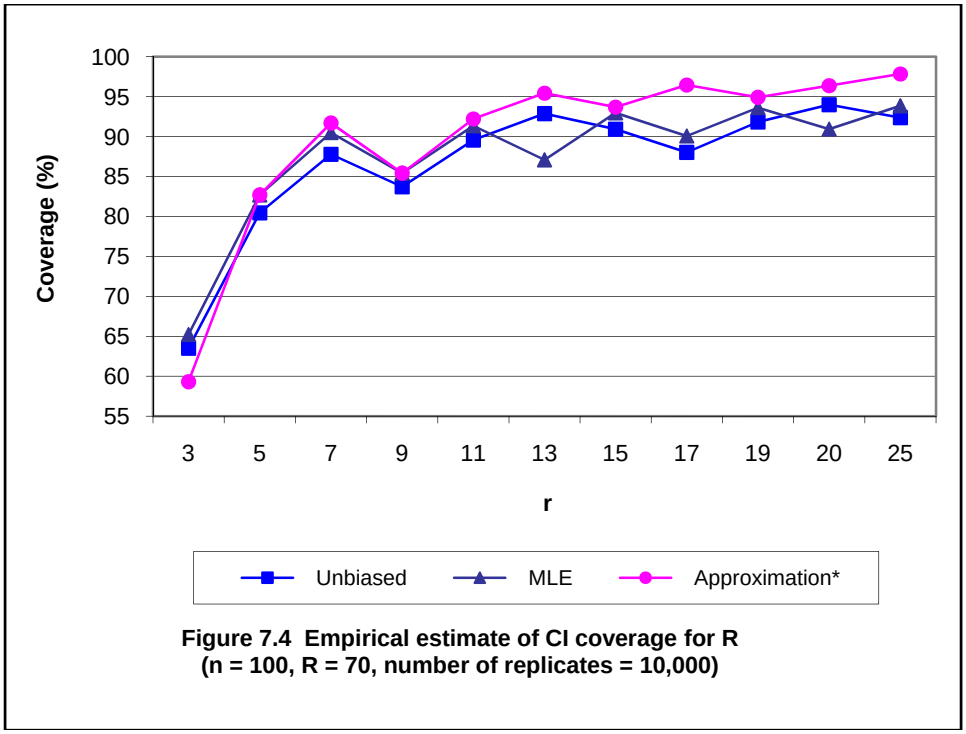

Figure 7.4 Empirical estimate of CI coverage for R (n = 100, R = 70, number of replicates = 10,000)

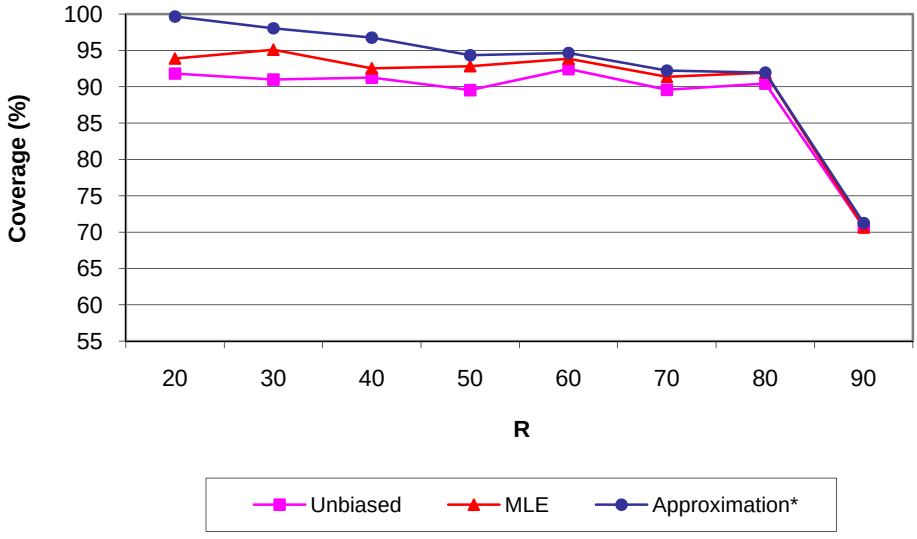

Figure 7.5 Empirical estimate of CI coverage (n = 100, r = 11, number of replicates = 10,000)

7.3 Empirical Estimates of Expected CI Length

The expected length of a CI is the expected difference between the upper bound and the lower bound. It is another important criterion used to evaluate CI's besides coverage. For similar coverage, the smaller the expected lengths, the better the performance of CI's.

1. For a fixed R, empirical estimates of expected lengths decreased as r increased (Figure 7 .6, 7.7. In addition, the empirical estimates of expected CI length using the unbiased estimator tended to be shorter lengths for fixed small r (e.g., N = 100, $r \leq 5$). However, estimates of expected lengths using the MLE converged to those using the unbiased estimator as r increased.
2. When R is large enough (e.g., N = 100, R = 70), the expected lengths using the MLE converged to those using the unbiased estimator regardless of the magnitude of r.
3. For a fixed r, the expected lengths increased as R increased from 20 to 60, and it reached peak at R = 60, then decreased as R increased (Figure 7.8).
4. We define the parameter θ as $\theta = \frac{R}{N}$ (where θ = 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9). For a given θ with a fixed r, the expected lengths increased with similar magnitude as the "population" size N increased regardless of the estimator (Figure 7.9).

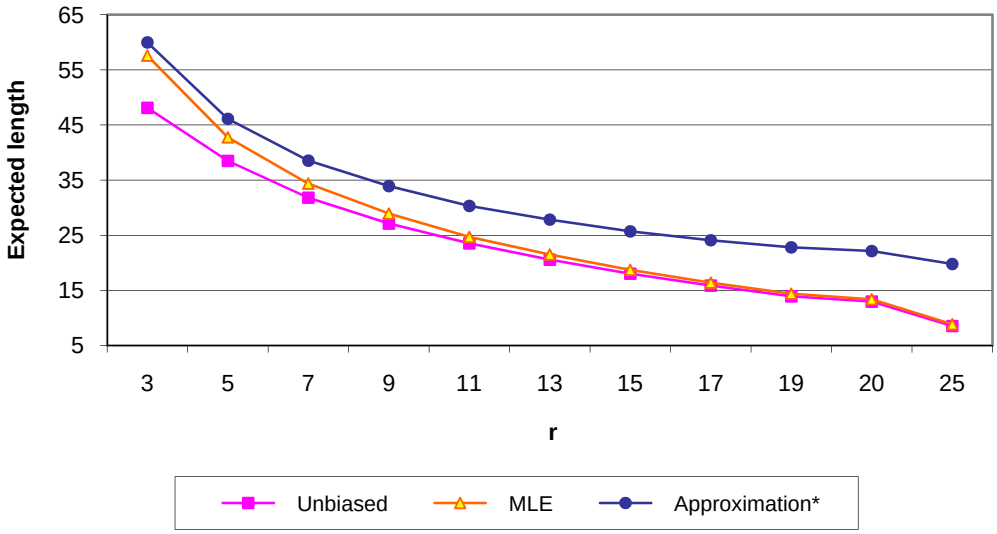

Figure 7.6 Empirical estimates of expected lengths of CIs (n = 100, R = 30, number of replicates = 10,000)

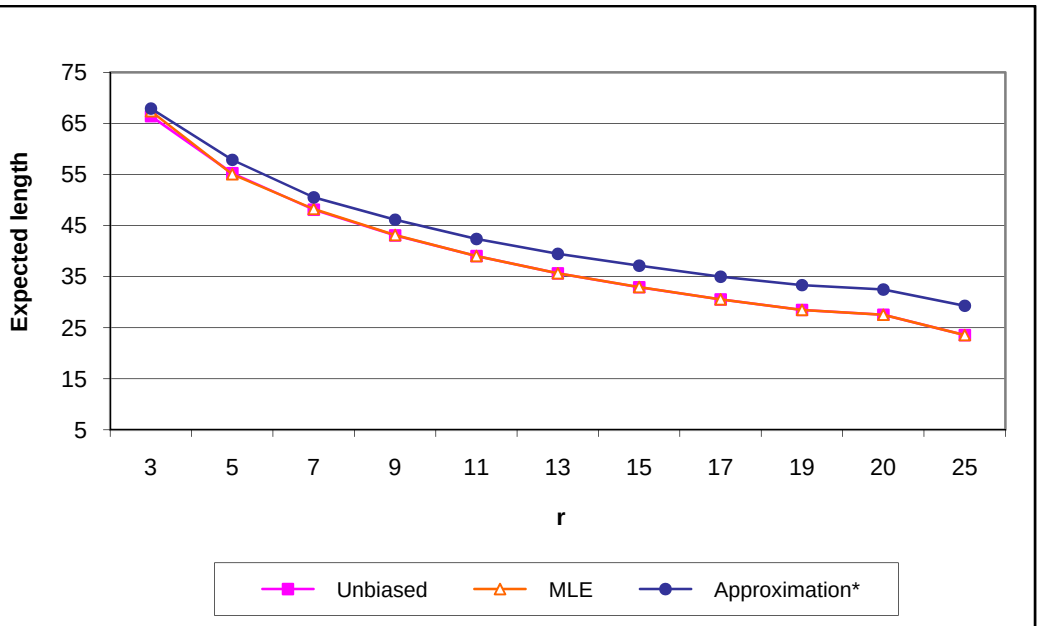

Figure 7.7 Empirical estimate of expected lengths of CIs (n = 100, R = 70, number of replicates = 10,000)

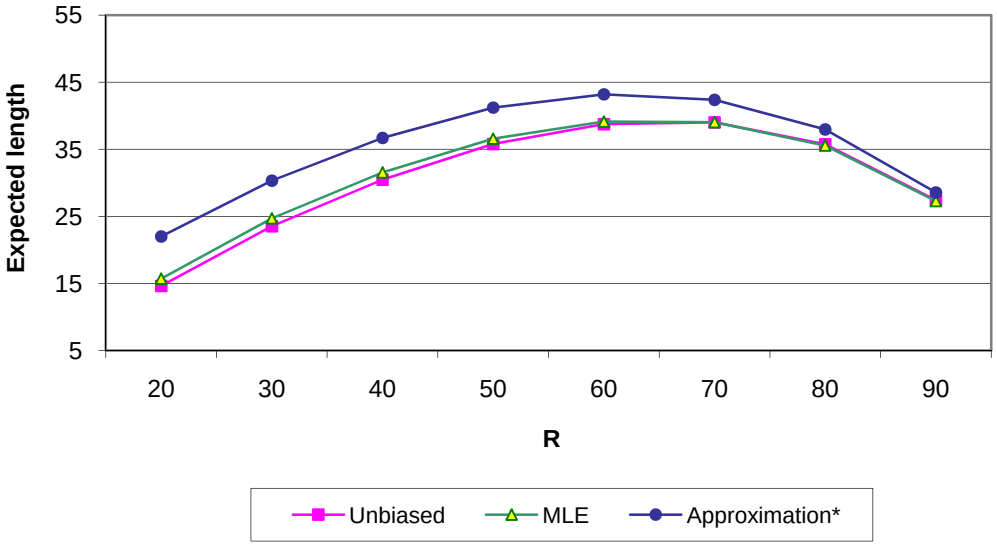

Figure 7.8 Empirical estimate of expected lengths of CIs (n = 100, r = 11, number of replicates = 10,000)

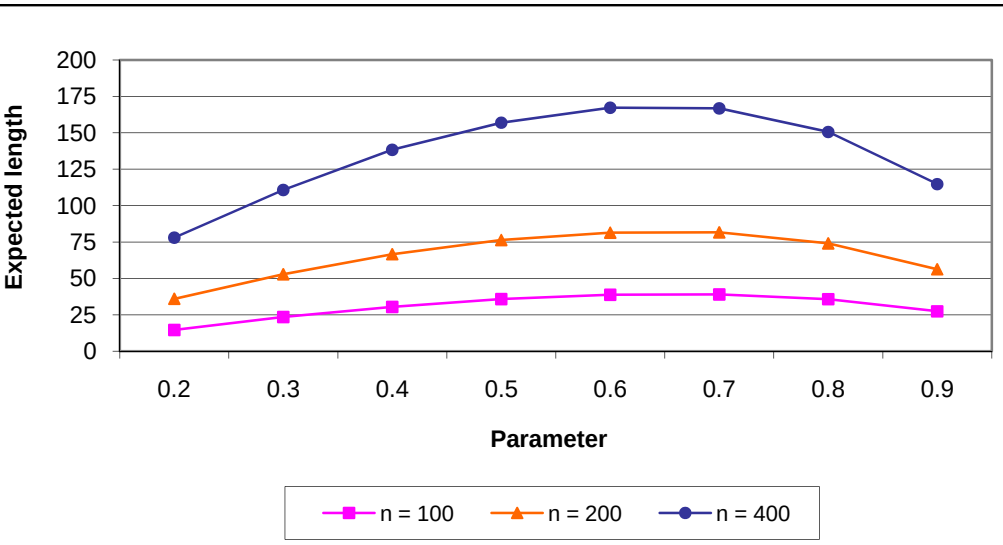

Figure 7.9 Empirical estimates of expected lengths of CIs using the unbiased estimator (r = 11, with 10,000 replicates)

In summary, CI's based on the negative binomial approximation do not provide adequate coverage properties to be recommended for general use. With respect to CI's

based on the unbiased estimator and the MLE, we conclude that either a smaller r (e.g., $r = 3$) or a bigger R (e.g., $R = 90$) will cause poor performances. In order to construct CI's with good properties, we must have reason to believe the range of R is 20 to 80, and r must be specified in the range of 10 to 20. Although the unbiased estimator is the point estimator of choice, CI's based on the MLE frequently out performed those based on the unbiased estimator in terms of coverage but the latter tended to be shorter in length. None of the CI types held coverage consistently at the 95% level.

REFERENCES

Bailey, N.T. (1951). Estimating the Size of Mobile Population from Recapture Data." *Biometrik*a, Vol. 38, No. ¾, 293-306.

Guenther, W.C. (1975). "The Inverse Hypergeometric – A Useful Model." *Statistica Neerlandica*, 29, 129-144.

Johnson, N. L. and Kotz, S. (1969). *Distributions in Statistics: Discrete Distributions.* Houghton Mifflin.

Moran, P.A.P. (1968). *An Introduction to Probability Theory*. Oxford, Great Britain.

Piccolo, D. (2001). "Some Approximation for the Asymptotic Variance of the Maximum Likelihood Estimator of the Parameter in the Inverse Hypergeometric Random Variable." *Quaderni di Statisca*, Vol. 3, 215-229.

SAS. (2005). *SAS* 9.1.3 *Language Reference*: *Dictionary*. SAS, Inc., Cary, NC.

SAS. (2005). Base *SAS* 9.1.3 *Procedures Guide*. SAS, Inc., Cary, NC.

Wilks, S. (1963). *Mathematical Statistics*, John Wiley & Sons, New York.

Zelterman, D. (2004). *Discrete Distributions*. John Wiley & Sons, New York.

Submitting author
Lei Zhang, PhD, Office of Health Data and Research, Mississippi State Department of Health, 570 East Woodrow Wilson, Jackson, MS 39215, USA. Phone (601) 576-8165, E-mail: lei.zhang@msdh.state.ms.us