8-2007

# Confidence Intervals for Population Size in a Capture-Recapture Problem.

Xiao Zhang
*East Tennessee State University*

## Recommended Citation

Zhang, Xiao, "Confidence Intervals for Population Size in a Capture-Recapture Problem." (2007). *Electronic Theses and Dissertations.* Paper 2022. https://dc.etsu.edu/etd/2022

# Confidence Intervals for Population Size in a Capture-Recapture Problem

A thesis

presented to

the faculty of the Department of Mathematics

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Master of Science in Mathematical Sciences

by

Xiao Zhang

August, 2007

Robert Price, Ph.D., Chair

Robert Gardner, Ph.D.

Yali Liu, Ph.D.

Keywords: Capture-recapture, Multinomial model, Hypergeometric model, Chapman interval, Wilson interval, Sample size, Population size, Mean coverage

# ABSTRACT

Confidence Intervals for Population Size in a Capture-Recapture Problem

by

Xiao Zhang

In a single capture-recapture problem, two new Wilson methods for interval estimation of population size are derived. Classical Chapman interval, Wilson and Wilson-cc intervals are examined and compared in terms of their expected interval width and exact coverage properties in two models. The new approach performs better than the Chapman in each model. Bayesian analysis also gives a different way to estimate population size.

# DEDICATION

I dedicate this thesis to my parents, Zhunan Zhang and Zhiling Dong, for their fosterage and cultivation, and my fiance, Liang Guo, for his love.

# ACKNOWLEDGMENTS

I would like to thank Dr. Price, my supervisor, for all his patience and invaluable help during my study at ETSU. I owe him a deep gratitude for his guidance and suggestion on my preparation for this thesis. I would also like to thank Dr. Gardner, our graduate coordinator: he has been ready to help me since my entrance to this program and he is a really kind and respectable friend of mine. Special thanks to Dr. Liu: she always encouraged and supported me like an elder sister and I truly learned a lot from her. Lastly, I express my gratitude and appreciation to all the people in the Department of Mathematics who have taught and helped me in the past two years.

# CONTENTS

## LIST OF TABLES

# 1 INTRODUCTION

## 1.1 The Problem of the Capture-Recapture

In capture-recapture sampling to estimate the total number of individuals in a population, an initial sample is obtained and the individuals in that sample are marked or otherwise identified. A second sample is independently marked. If the second sample is representative of the population as a whole, then the sample proportion of marked individuals should be about the same as the population proportion of marked individuals. From this relationship, the total number of individuals in the population can be estimated.

Capture-recapture methods were originally developed in the wildlife biology to monitor the census of bird, fish, and insect populations [7]. They have been used to estimate the abundance of animal populations, to estimate the detectability of animals for other survey methods, and to estimate survival and other population parameters. Recently, these methods have been utilized to estimate the abundance of elusive human populations such as the homeless, to adjust for census undercounts of minority groups, and to estimate the number of vital events such as accidents in a population [7].

The animals or other individuals need not literally be captured or marked or recaptured. If it is possible to identify individual animals by natural markings, then two independent sighting surveys may be carried out, and the number of individuals sighted in both surveys is the number of "recaptures". Similarly, if a number of animals in a population has been fitted with radio transmitters and hence has known locations, then in a survey in which observers detect animals by some means independently of the transmitters, the number of transmitter-fitted animals detected is the number of recaptures. For other species, however, it may be necessary to capture the animals by such means as traps or nets, and to mark them with bands, tags, coded wire implants, paint, or streamers.

For human populations, the two samplers often consist of two lists. For instance, the first list may be from the census data and the second list may be data from a follow-up survey. Or the first list may be health department records of accidents, and the second list, insurance company records.

In more complex capture-recapture animal studies, animals may be captured and released on several different occasions, with the capture history of any animal in the sample identifiable from the previous marks. Complicating factors include capture probabilities that vary from animal to animal or from sample to sample, mortality caused by tagging, mortality between sample times, births, immigration and emigration from the study area, and animals becoming “trap happy” or “trap shy” through the handling procedure [7].

## 1.2 2 × 2 Contingency Table

In the following summary of simple capture-recapture methods, a $2 \times 2$ contingency table is used to facilitate consideration of sampling design aspects (Table 1). The total number of marked individuals in the population, which is also the number of individuals in the initial sample, is $n_1$. The number of individuals in the second sample is $n_2$, of which $n_{11}$ are detected. The total number $N$ of animals in the population may then be estimated.

Table 1: $2 \times 2$ Contingency Table

| | First Sample | | Total | |
|---|---|---|---|---|
| | $n_{11}$ | $n_{21}$ | $n_2$ | Second Sample |
| | $n_{12}$ | $n_{22}$ | | |
| Total | $n_1$ | | $N$ | |

In the general $2 \times 2$ contingency table for a single capture and recapture of a closed population, the capture or detection history of any animal in the population can be categorized into exactly one of four categories: detected on both first and second occasions, detected on first but not on second, detected on second but not on first, detected on neither occasion. Here we restrict our model and assume that the detection probability is the same for each individual in the population during a sampling occasion. Independence between the two sampling occasions is also assumed. If the second sample is representative of the population as a whole, the proportion of marked animals in the sample will be about the same as the proportion of the whole population in the sample. The total number $N$ of animals in the population may be estimated by assuming that the proportion of marked animals in the second sample is representative of the proportion of marked animals in the population, i.e., by setting

$$\frac{n_{11}}{n_2} = \frac{n_1}{N} \tag{1}$$

and solving for the unknown population size $N$. Equivalently, the proportion of marked animals in the population that is captured in the second sample should approximately equal the proportion of the population as a whole captured in the second sample, that is,

$$\frac{n_{11}}{n_1} = \frac{n_2}{N}. \tag{2}$$

Solving either equation for the unknown population total $N$ gives the Petersen estimator

$$\widehat{N} = \frac{n_2 n_1}{n_{11}}. \tag{3}$$

An estimator of the variance of $N$ [3] is

$$\widehat{var(\hat{N})} = \frac{n_1 n_2 n_{12} n_{21}}{n_{11}^3}. \tag{4}$$

The maximum likelihood estimator of the probability $p_1$ of capture in the first sample is $\hat{p_1} = n_1/N$. The MLE of capture probability for the second sample is $\hat{p_2} = n_2/N$.

# 2 ESTIMATION AND INFERENCE IN SIMPLE CAPTURE-RECAPTURE MODELS

## 2.1 Multinomial and Hypergeometric Models

In a multinomial model, $N$ individuals are regarded as being multinomially distributed into a number of capture histories, the observable ones with probability $p_{ij}, i, j = 1, 2$, and the unobservable category with the probability $1-p^*$, where $p^* = \sum p_{ij} \leq 1$. Thus, a multinomial model, with the four probabilities for the four cells adding to one, applies to the capture history of each animal. If, in addition, on each sampling occasion the detection outcomes for different individuals are independent, then the model for the numbers of individuals with each capture history will be a product of multinomials. In the general model, the probability of detection may be different for different sampling occasions. The general models contain too many parameters in relation to the number of observations, so that further restrictions are needed for effective estimation of $N$ or of detection probabilities.

One such restricted model assumes that detection probability is the same for each individual in the population during a sampling occasion, but may differ for the two samples. Independence between the two sampling occasions is also assumed. With this model, the maximum likelihood estimator of population total is the integer part of the Petersen estimator $\hat{N} = (n_2 n_1)/n_{11}$. Even if capture probabilities are different for different individuals at the first sample but equal at the second sample, the estimator is still the Petersen estimator. If a single capture probability $p$ applies to both samples and to all individuals, with independence between samples, the maximum likelihood estimators of $p$ and $N$ are $\hat{p} = 2n_{11}/(n_1 + n_2)$ and $\hat{N} = (n_1 + n_2)/2\hat{p}$.

If the numbers $n_1$ and $n_2$ of individuals in the two samples are fixed and the second sample is a simple random sample of the individuals in the population, we now try to find $P(\{n_{ij}\}_{i,j=1,2}|n_1, n_2)$, the conditional density of $n_{ij}$ given the number of samples. It is easily deduced from the multinomial model that $n_1$, $n_2$ are independent binomial variables $B(N, p_i)$, $i = i, 2$. Therefore

$$P(n_1, n_2) = \prod_{i=1}^{2} \binom{N}{n_i} p_i^{n_i} (1 - p_i)^{N-n_i}. \tag{5}$$

We notice that

$$\begin{aligned} P(\{n_{ij}\}_{i,j=1,2}) &= \binom{N}{n_{11}\ n_{12}\ n_{21}\ n_{22}} p_{11}^{n_{11}} p_{12}^{n_{12}} p_{21}^{n_{21}} p_{22}^{n_{22}} \qquad (6) \\ &= \binom{N}{n_{11}\ n_{12}\ n_{21}\ n_{22}} p_1^{n_1} p_2^{n_2} (1 - p_1)^{N-n_1} (1 - p_2)^{N-n_2} \end{aligned}$$

and

$$\begin{aligned} P(\{n_{ij}\}_{i,j=1,2}|n_1, n_2) &= \frac{\binom{N}{n_{11}\ n_{12}\ n_{21}\ n_{22}}}{\binom{N}{n_1}\binom{N}{n_2}} \qquad (7) \\ &= \frac{n_1!\, n_2!\, (n_{21} + n_{22})!\, (n_{12} + n_{22})!}{N!\, n_{11}!\, n_{22}!\, n_{12}!\, n_{21}!} = \frac{\binom{n_1}{n_{11}}\binom{N-n_1}{n_{21}}}{\binom{N}{n_{21}}}, \end{aligned}$$

where $0 \leq n_{11} \leq n_1$, $0 \leq n_{12}, n_{21} \leq min\{n_1, n_2\}$, with the linear constraints $n_{11} + n_{12} = n_1$, $n_{11} + n_{21} = n_2$, (7) is the generalized hypergeometric density. Then the number $n_{11}$ of marked animals in the second sample has a hypergeometric distribution. With equal capture probabilities among individuals, this is the conditional distribution under the multinomial model of $n_{11}$ given $n_1$ and $n_2$. Under this model, the maximum likelihood estimator of $N$ is again the integer part of the Petersen estimator.

In a multinomial model the sample size $n_i, i = 1, 2$, are random variables while $p_i, i = 1, 2$, are parameters. This model is therefore applicable when the effort put into the catching of every sample is fixed before the experiment begins since the $p_i$ are then fixed, though unknown. The hypergeometric model, on the other hand, involves the $n_i$ as parameters and should be used only when the experimenter is determined to catch no more and no less than

$n_i$ individuals at the $i$th sample; and he or she will only be able to do this when animals are fairly easily caught. In fact, if we had to generalize, we could say that the hypergeometric model is likely to be appropriate when the main limiting factor on sample size is the trouble involved and the multinomial model is more appropriate when the limiting factor is the source of difficulty in catching them [8].

## 2.2 Chapman Interval and Wilson Interval for the Paired Data

### 2.2.1 Wald and Wilson Interval for the Binomial Parameters

Assume that $n_{ij}$ follows a multinomial distribution with parameters $N$ and $p_{ij}$. The corresponding probabilities for the $2 \times 2$ contingency table are shown in Table 2 where $p_1 = p_{11} + p_{12}$ and $p_2 = p_{11} + p_{21}$ are the marginal probabilities of interest.

Table 2: $2 \times 2$ Contingency Table

| | First Sample | | Total | |
|---|---|---|---|---|
| | $p_{11}$ | $p_{21}$ | $p_2$ | Second Sample |
| | $p_{12}$ | $p_{22}$ | | |
| Total | $p_1$ | | 1 | |

A Wald confidence interval for a single proportion is defined as

$$\hat{p}_i \pm z_{\alpha/2}\sqrt{\hat{p}_i(1-\hat{p}_i)/N} \tag{8}$$

where $P(Z > z_{\alpha/2}) = \alpha/2, i = 1, 2$.

A binomial method, noted for its computational simplicity, was recently proposed by Agresti and Coull [1]. The Agresti-Coull interval is defined as

$$\tilde{p}_i \pm z_{\alpha/2}\sqrt{\tilde{p}_i(1-\tilde{p}_i)/(N+4)} \tag{9}$$

where $\tilde{p}_i = (n_i + 2)/(N + 4)$.

Wilson [5] gave a general approach, sometimes referred to as a score interval, that was derived by inverting an approximate normal test using a standard error estimate under the constraint of the null hypothesis. It has the form

$$\tilde{p}_i \pm \frac{z_{\alpha/2}}{\tilde{N}}\sqrt{N\tilde{p}_i(1-\tilde{p}_i) + \frac{z^2_{\alpha/2}}{4}} \tag{10}$$

where $\tilde{N} = N + z^2_{\alpha/2}$.

An approximate $100(1-\alpha)\%$ Wald confidence interval for $p_1 - p_2$ is

$$\hat{p_1} - \hat{p_2} \pm z_{\alpha/2}\sqrt{(\hat{p_2} + \hat{p_1} - (\hat{p_1} - \hat{p_2})^2)/N}. \tag{11}$$

An alternative for $p_1 - p_2$ is

$$\tilde{p_1} - \tilde{p_2} \pm z_{\alpha/2}\sqrt{(\tilde{p_2} + \tilde{p_1} - (\tilde{p_1} - \tilde{p_2})^2)/(N+2)} \tag{12}$$

where $\tilde{p_i} = (n_i + 1)/(N+2)$.

In some applications, a ratio of marginal proportions $p_1/p_2$ may be more interesting or meaningful than a difference. A $100(1-\alpha)\%$ transformed Wald interval for $\theta = p_1/p_2$ has the simple form

$$exp[\ln(n_1/n_2) \pm z_{\alpha/2}\{(n_{12} + n_{21})/n_1 n_2\}^{1/2}]. \tag{13}$$

Newcombe [10] describes Wilson intervals for a single proportion with and without a continuity correction and Newcombe [9] combines two Wilson confidence intervals to obtain a confidence interval for the difference in proportions using independent samples. Bonett and Price [4] proposed an alternative to the Wald interval that combines two Wilson confidence intervals. The width of the interval could only depend on $n_{11}, n_{12}$ and $n_{21}$, so if we let $n' = n_{11} + n_{21} + n_{12}$, $\hat{p'}_1 = n_1/n'$ and $\hat{p'}_2 = n_2/n'$, and a proposed $100(1-\alpha)\%$ Wilson confidence interval for $p_1/p_2$ may be expressed as

$$[\exp\{\ln(L_1) - \ln(U_2)\}, exp\{\ln(U_1) - \ln(L_2)\}] = (L_1/U_2, U_1/L_2), \tag{14}$$

where $\ln(L_1) = \ln(\hat{p'}_1) - kz_{1-\alpha/2}se\{\ln(\hat{p'}_1)\}$, $\ln(U_1) = \ln(\hat{p'}_1) + kz_{1-\alpha/2}se\{\ln(\hat{p'}_1)\}$, $\ln(L_2) = \ln(\hat{p'}_2) - kz_{1-\alpha/2}se\{\ln(\hat{p'}_2)\}$, $\ln(U_2) = \ln(\hat{p'}_1) + kz_{1-\alpha/2}se\{\ln(\hat{p'}_2)\}$, and $k = se\{\ln(\hat{p'}_1) -$

$\ln(\hat{p'}_2)\}/[se\{\ln(\hat{p'}_1)\} + se\{\ln(\hat{p'}_2)\}]$.

The Wilson interval for $\hat{p'}_i$ without a continuity correction is

$$[2n_i + z^2 \pm z\{z^2 + 4n_i(1 - \hat{p'}_i)\}^{1/2}]/b \tag{15}$$

where $b = 2(n' + z^2)$ and $z = kz_{1-\alpha/2}$. The lower and upper endpoint of Wilson interval for $\hat{p'}_i$ with a continuity corretion are

$$[2n_i + z^2 - 1 - z\{z^2 - 2 - 1/n' + 4\hat{p'}_i(n' - n_i + 1)\}^{1/2}]/b \tag{16}$$

$$[2n_i + z^2 + 1 + z\{z^2 + 2 - 1/n' + 4\hat{p'}_i(n' - n_i - 1)\}^{1/2}]/b.$$

### 2.2.2 Chapman and Wilson Interval for Population Size

In a capture-recapture problem, a simple, approxiamate $100(1-\alpha)\%$ confidence interval from (3) and (4) is the standard

$$\frac{n_1 n_2}{n_{11}} \pm z_{1-\alpha/2}\sqrt{\frac{n_1 n_2 n_{12} n_{21}}{n_{11}^3}}. \tag{17}$$

Because the number $n_{11}$ of marked animals in the second sample may be zero, the estimator does not have a finite variance. Therefore the following estimator $\hat{N}$ modified (3) was proposed by Chapman [6]:

$$\hat{N} = \frac{(n_1+1)(n_2+1)}{n_{11}+1} - 1. \tag{18}$$

An approximate unbiased estimator of the variance of the modified estimator is

$$\widehat{var(\hat{N})} = \frac{(n_1+1)(n_2+1)n_{12}n_{21}}{(n_{11}+1)^2(n_{11}+2)}. \tag{19}$$

An approximate $100(1-\alpha)\%$ confidence interval for the Chapman estimator is

$$\hat{N} \pm z_{1-\alpha/2}\sqrt{\widehat{var(\hat{N})}}. \tag{20}$$

As pointed out by Bonett and Price [4], the new Wilson methods for the ratio of the proportions are easy to compute and perform as well or better than the traditional method. To achieve comparable results of an interval estimate for population size $N$, a Wilson interval is appealing here. Notice the width of the interval depends only on $n_{11}$, $n_{12}$, $n_{21}$. Thus (20) may be expressed as

$$\exp[\ln(\frac{n'\hat{p'}_1\hat{p'}_2}{\hat{p'}_{11}}) \pm z_{1-\alpha/2}\ln(se\{\frac{n'\hat{p'}_1\hat{p'}_2}{\hat{p'}_{11}}\})]. \tag{21}$$

Assuming the value of $k$ is

$$k = se\{\ln(\hat{p_1}') + \ln(\hat{p_2}') - \ln(\hat{p_{11}}')\}/[se\{\ln(\hat{p_1}')\} + se\{\ln(\hat{p_2}') + ln(\hat{p_{11}}')\}] \tag{22}$$

The upper and lower bounds of (21) would be:

$$\exp[\ln(n') + U_1 + U_2 - L_{11}] \tag{23}$$

$$\exp[\ln(n') + L_1 + L_2 - U_{11}] \tag{24}$$

where $U_1$ is the upper bound for $\ln(\hat{p_1}')$, $U_2$ is the upper bound for $\ln(\hat{p_2}')$, $L_{11}$ is the lower bound for $\ln(\hat{p_{11}}')$, $L_1$ is the lower bound for $\ln(\hat{p_1}')$, $L_2$ is the lower for $\ln(\hat{p_2}')$, $U_{11}$ is upper bound for $\ln(\hat{p_{11}}')$. These are all using the adjusted $z$. The proposed $100(1-\alpha)\%$ confidence interval for $N$ replaces the standard Wald interval estimates with Wilson interval estimates.

An estimate of $k$ is needed to compute for Wilson interval here. To avoid problems with sampling zeros, we use $[(1-p_i^*)/\{(n'+2)p_i^*\}]^{1/2}$ to estimate $se\{\ln(\hat{p_i}')\}$ where $p_i^* = (n_i+1)/(n'+2)$ are Laplace estimates. We propose estimating $se\{\ln(\hat{p_1}') + \ln(\hat{p_2}') - \ln(\hat{p_{11}}')\}$ by using delta method as $\{1/n_{11} - 1/(n'+2) - (n_{12}+n_{21}+2)/(n_1+1)(n_2+1)\}^{1/2}$, where the lower limit is set to 0. The confidence interval used by (15) and (16) will be referred to as Wilson and Wilson-cc methods respectively. Matlab code is given in the Appendix for the three methods that are compared in this thesis.

# 3 APPLYING REFERENCE ANALYSIS

## 3.1 The Comparison of the Three Intervals

In practice, we add to the actual Wilson and Wilson-cc confidence interval a correction, i.e., replace all $n'$ with $n'+3$, all $n_i$ and $n_{ij}$ with $n_i+2$ and $n_{ij}+1$ respectively, $i, j = 1, 2$. This adjustment was proposed to avoid sampling zeros and nonpositive numbers in the square root for Wilson-cc interval. Results of the computation suggests the mean exact coverage probabilities are close, but the adjusted Wilson CIs are truely better than unadjusted CIs in terms of interval width, especially for larger sampling.

We examined 3000 different $2 \times 2$ contingency tables for different sample probabilities in multinomial models (Tables 3-5). In this model, the two marginal probabilities were regarded as fixed, though unknown. The value of $p_1$, $p_2$ was randomly generated from a Gamma $(1/2,\ 1)$ distribution subjected to $p_{ij} > .0001$. The minimum exact probability, the mean coverage probability, and the median expected interval width are computed for all $2 \times 2$ contingency tables with respect to Chapman, Wilson and Wison-cc intervals for $N = 5, 10, 30, 50, 100, 150, 200$ and $1 - \alpha = .9, .95, .99$. A small adjustment $c = .25$ was performed on Chapman CI since $n_{12}$ and $n_{21}$ may be zeros.

When applying the hypergeometric model with fixed sample size $n_1$ and $n_2$, for convenience we assume $n_1 = n_2$. Since the hypergeometric model is applicable mainly when individuals are available to capture as desired, it is practical to simplify this model by equal sample size. We notice this model eliminates parameter $p_i$ and leaves only $N$ to be estimated, which allow us to get the exact coverage probability and width for each fixed sample and computation can be facilitated for larger sampling. Population size $N = 10, 50, 100, 200, 500$

and $1-\alpha = .9, .95$ were chosen with different fixed values of $n_1$ and $n_2$ in Tables 6 and 7. Some very small or very large samples were excluded by our study: a small value may result in a substantial deviation and yield an extremely low or even zero coverage probability while a large value has no more actual meaning.

The results in Tables 3-5 suggest that all of the three intervals can have a true coverage probability of zero. The Chapman interval is clearly the worst one in that its mean coverage is always no more than .80 even for $1-\alpha = .99$ and $N = 200$ although its mean coverage probability increases as population size increases. The Wilson and Wilson-cc intervals have similar performance. They may have a mean coverage close to $1-\alpha$, although sometimes they still fall below $1-\alpha$ a bit. The Wilson-cc interval is slightly better than Wilson interval in terms of the actual coverage probability in small populations, say $N \leq 30$ and the Wilson interval is slightly better than Wilson-cc interval in terms of the actual coverage probability in larger population. Moreover, the Wilson-cc interval exhibits a better characteristic in narrowing the median width which is another primary consideration for our study.

From Tables 6 and 7 it can be seen that all of the three intervals have satisfactory coverage probabilities except for very small samples. In these hypergeometric tables it appears that they all perform better than in the multinomial table although we will interpret later that the two models are in fact equivalent. The Wilson interval still tends to be wider than the Wilson-cc interval and Chapman still has the worst performance among the three methods. Hence here we would say the Wilson and Wilson-cc intervals can be recommended for general use.

In [8], Darroch mentioned that the hypergeometric model may be regarded as a very useful device for eliminating the nuisance parameters $p_i$ when $n_i$ are variables. One feels intuitively that to estimate $N$ as if the $n_i$ are constants, when in fact they are not, is not a serious misrepresentation, and this is supported by the discovery that the two models lead to the same estimate $\hat{N}$ of $N$ and to the same asymptotic estimate of $var(\hat{N})$. Apart from demonstrating this, it may be wondered why there is any need to consider the multinomial model at all. The main reason is that it is capable of generalization which the hypergeometric model is unable to accommodate. Since the multinomial model in our study naturally includes each possible value of $n_1$, $n_2$ through $p_1$, $p_2$ and the hypergeometric model only covers appropriate chosen values of $n_1$, $n_2$, the mean coverage probability in Tables 3-5 will be more likely to have lower value than that in the Tables 6 and 7, and the same for the minimum coverage probability.

Table 3: Coverage Property Summary of Three Intervals for $1 - \alpha = .90$

| | Min | Cov | | Mean | Cov | | Med | Width | |
|---|---|---|---|---|---|---|---|---|---|
| N | Chap | Wil-cc | Wil | Chap | Wil-cc | Wil | Chap | Wil-cc | Wil |
| 10 | 0.0000 | 0.9635 | 0.5640 | 0.3805 | 0.9940 | 0.9427 | 8.1218 | 34.4033 | 42.9913 |
| 30 | 0.0000 | 0.4342 | 0.1292 | 0.4826 | 0.9405 | 0.9101 | 24.0188 | 69.2462 | 96.4618 |
| 50 | 0.0000 | 0.0086 | 0.0715 | 0.5218 | 0.8834 | 0.8921 | 38.8833 | 96.6887 | 133.6375 |
| 100 | 0.0000 | 0.0002 | 0.0406 | 0.5891 | 0.8644 | 0.8767 | 72.2277 | 151.3313 | 185.3231 |
| 150 | 0.0000 | 0.0000 | 0.0326 | 0.6303 | 0.8659 | 0.8788 | 106.8450 | 191.8241 | 243.7200 |
| 200 | 0.0000 | 0.0000 | 0.0272 | 0.6446 | 0.8768 | 0.8774 | 129.7571 | 248.9656 | 274.8937 |

Table 4: Coverage Property Summary of Intervals for $1 - \alpha = .95$

| | Min | Cov | | Mean | Cov | | Med | Width | |
|---|---|---|---|---|---|---|---|---|---|
| N | Chap | Wil-cc | Wil | Chap | Wil-cc | Wil | Chap | Wil-cc | Wil |
| 10 | 0.0007 | 0.9770 | 0.7139 | 0.4673 | 0.9947 | 0.9610 | 9.4993 | 40.6493 | 57.4718 |
| 30 | 0.0000 | 0.6211 | 0.1552 | 0.5248 | 0.9640 | 0.9281 | 28.5148 | 81.5489 | 126.6523 |
| 50 | 0.0000 | 0.0158 | 0.1891 | 0.5671 | 0.9128 | 0.9300 | 47.1717 | 112.0127 | 173.8949 |
| 100 | 0.0000 | 0.0012 | 0.0857 | 0.6464 | 0.8991 | 0.9239 | 88.0090 | 177.9665 | 253.6435 |
| 150 | 0.0000 | 0.0000 | 0.0795 | 0.6713 | 0.9015 | 0.9185 | 127.3137 | 243.3423 | 282.5242 |
| 200 | 0.0000 | 0.0001 | 0.1437 | 0.7004 | 0.9084 | 0.9219 | 161.1529 | 286.5510 | 327.1040 |

Table 5: Coverage Property Summary of Three Intervals for $1 - \alpha = .99$

| | Min | Cov | | Mean | Cov | | Med | Width | |
|---|---|---|---|---|---|---|---|---|---|
| N | Chap | Wil-cc | Wil | Chap | Wil-cc | Wil | Chap | Wil-cc | Wil |
| 10 | 0.0007 | 0.9813 | 0.8074 | 0.4907 | 0.9947 | 0.9718 | 12.7407 | 50.9118 | 94.7730 |
| 30 | 0.0020 | 0.9553 | 0.4930 | 0.5857 | 0.9905 | 0.9641 | 37.6153 | 102.3571 | 205.5328 |
| 50 | 0.0000 | 0.3022 | 0.2653 | 0.6166 | 0.9548 | 0.9559 | 60.1164 | 144.5744 | 278.1554 |
| 100 | 0.0000 | 0.0134 | 0.2905 | 0.6988 | 0.9368 | 0.9618 | 117.0141 | 236.6653 | 409.7357 |
| 150 | 0.0000 | 0.0030 | 0.3259 | 0.7217 | 0.9364 | 0.9609 | 163.4965 | 302.0030 | 481.9743 |
| 200 | 0.0000 | 0.0019 | 0.3552 | 0.7600 | 0.9413 | 0.9645 | 213.7892 | 358.0413 | 494.1940 |

Table 6: Coverage Property of Three Intervals with Fixed Sample Size for $1 - \alpha = .90$

| | | Exact | Cov | | Width | | |
|---|---|---|---|---|---|---|---|
| N | $n_1, n_2$ | Chap | Wil-cc | Wil | Chap | Wil-cc | Wil |
| 10 | 1 | −− | 0.9000 | 0.9000 | −− | 14.5628 | 39.1880 |
| 10 | 2 | 0.6222 | 0.9778 | 0.9778 | 11.0453 | 22.4357 | 56.9836 |
| 10 | 3 | 0.8167 | 0.9917 | 0.9917 | 13.7900 | 29.4843 | 51.8550 |
| 10 | 4 | 0.8810 | 0.9952 | 0.8810 | 12.3175 | 30.7255 | 31.4707 |
| 50 | 5 | 0.5766 | 0.9282 | 0.9282 | 51.5070 | 61.0883 | 182.2700 |
| 50 | 10 | 0.6856 | 0.9819 | 0.9034 | 72.1728 | 109.9600 | 151.0685 |
| 50 | 15 | 0.7524 | 0.9767 | 0.9080 | 50.2647 | 81.7312 | 66.7883 |
| 50 | 20 | 0.8117 | 0.9772 | 0.8602 | 33.2816 | 47.9082 | 35.8838 |
| 100 | 5 | −− | −− | 0.7696 | −− | −− | 229.3699 |
| 100 | 10 | 0.7385 | 0.9400 | 0.7385 | 137.7927 | 148.9500 | 361.7927 |
| 100 | 20 | 0.8273 | 0.9363 | 0.9363 | 121.0957 | 180.7800 | 173.1040 |
| 100 | 40 | 0.8482 | 0.9399 | 0.8557 | 48.1634 | 58.1052 | 46.5609 |
| 200 | 10 | 0.5915 | 0.3268 | 0.9182 | 190.5761 | 158.5700 | 550.6330 |
| 200 | 20 | 0.6787 | 0.8778 | 0.8778 | 310.7522 | 362.1900 | 602.9639 |
| 200 | 40 | 0.8646 | 0.9315 | 0.9179 | 179.4076 | 236.0957 | 203.7034 |
| 200 | 60 | 0.8754 | 0.9213 | 0.9039 | 106.4456 | 123.2678 | 105.5214 |
| 500 | 10 | −− | −− | 0.8196 | −− | −− | 702.4753 |
| 500 | 25 | 0.6405 | 0.8776 | 0.8776 | 735.1968 | 649.9512 | 1570.0000 |
| 500 | 50 | 0.7793 | 0.8894 | 0.8855 | 627.8726 | 819.9072 | 822.0001 |
| 500 | 100 | 0.8877 | 0.9204 | 0.9044 | 289.9297 | 327.0182 | 293.2510 |

Table 7: Coverage Property of Three Intervals with Fixed Sample Size for $1-\alpha=.95$

| | | Exact | Cov | | Width | | |
|---|---|---|---|---|---|---|---|
| N | $n_1, n_2$ | Chap | Wil-cc | Wil | Chap | Wil-cc | Wil |
| 10 | 1 | $--$ | 0.9000 | 0.9000 | $--$ | 14.7436 | 52.9729 |
| 10 | 2 | 0.6222 | 0.9778 | 0.9778 | 13.1613 | 23.7910 | 77.6478 |
| 10 | 3 | 0.8167 | 0.9917 | 0.9917 | 16.4318 | 32.4121 | 69.7475 |
| 10 | 4 | 0.8810 | 0.9952 | 0.8810 | 14.6772 | 34.4425 | 40.7838 |
| 50 | 5 | 0.5766 | 0.9282 | 0.9282 | 61.3744 | 69.6557 | 248.8556 |
| 50 | 10 | 0.6856 | 0.9819 | 0.9034 | 85.9992 | 129.1636 | 196.4072 |
| 50 | 15 | 0.9094 | 0.9767 | 0.9094 | 59.8940 | 94.6461 | 81.6053 |
| 50 | 20 | 0.9295 | 0.9801 | 0.9262 | 39.6574 | 55.1449 | 43.0164 |
| 100 | 5 | $--$ | $--$ | 0.7696 | $--$ | $--$ | 314.9744 |
| 100 | 10 | 0.7385 | 0.9400 | 0.9400 | 164.1902 | 182.9894 | 485.6957 |
| 100 | 20 | 0.8273 | 0.9363 | 0.9363 | 144.2944 | 210.6808 | 214.9689 |
| 100 | 40 | 0.9272 | 0.9665 | 0.8979 | 59.3903 | 67.2107 | 55.6146 |
| 200 | 10 | 0.5915 | 0.9182 | 0.9182 | 227.0854 | 204.2069 | 748.2942 |
| 200 | 20 | 0.6787 | 0.9655 | 0.8778 | 370.2841 | 431.6111 | 785.5403 |
| 200 | 40 | 0.8646 | 0.9726 | 0.9315 | 213.7773 | 276.8388 | 246.1879 |
| 200 | 60 | 0.9339 | 0.9620 | 0.9213 | 126.8378 | 143.9254 | 126.1652 |
| 500 | 10 | $--$ | $--$ | 0.8156 | $--$ | $--$ | 960.5970 |
| 500 | 25 | 0.6405 | 0.8776 | 0.8776 | 876.0411 | 786.3512 | 2077.4000 |
| 500 | 50 | 0.8894 | 0.9517 | 0.9517 | 748.1564 | 957.8782 | 1016.5000 |
| 500 | 100 | 0.9354 | 0.9562 | 0.9204 | 345.4725 | 384.5816 | 351.0182 |

## 3.2 Two Examples

*Example* 1 : In a field study, $x = 300$ mice are caught in traps, tagged, and released. A few days later the researchers go to the study area and independently capture $y = 200$ mice, of which they find that $x = 50$ have tags. The Chapman estimate for equation (18) is $N = 1185.3$, the estimated variance is 16774.5. The approximate 95% Chapman confidence interval is $(931.5, 1439.1)$. The approximate 95% Wilson and Wilson-cc intervals are $(973.2, 1467.3)$ and $(961.2, 1486.9)$ respectively. Since the sample size is large enough, the three intervals give similar results. Using the Wilson-cc interval we conclude with 95% confidence that in this field the population of mice is between 961 to 1487.

*Example* 2 : In a wildlife survey in which the samples are selected by canvassing a study region from a helicopter landing to mark 27 red deer detected on the first sampling occasion and later noting 3 of 37 observed on the second occasion are marked. The typical multinomial model may apply reasonably to this problem since neither sample size is fixed. The animals are assumed to be distributed evenly in this study region. The approximate 95% Chapman, Wilson and Wilson-cc confidence intervals for population size are $(60.8, 495.2)$, $(122.6, 882.3)$ and $(101.2, 684.3)$ respectively. We can see some difference among them and all of them may be too wide to provide useful information for this small population. Either relocation for this single recapture problem or further multiple-recapture procedures may be required for future survey studies.

# 4 BAYESIAN STATISTICAL METHOD

## 4.1 Introduction to Bayesian Inference

Unlike methods of traditional statistical inference that are primarily based on a retrospective evaluation of the distribution of possible $y$ values conditional on the true unknown parameter $\theta$, Bayesian methods distinguish themselves explicitly by conditioning on the observed data to quantify uncertainty in statistical data analysis.In order to obtain such a probability statement about $\theta$ given $y$, we must begin with a model providing a joint probability distribution for $\theta$ and $y$. From probability theory, the joint probability density function can be written as a product of two densities, that are often referred to as the prior distribution $p(\theta)$ and the sampling distribution $p(y|\theta)$ respectively:

$$p(\theta, y) = p(\theta)p(y|\theta). \tag{25}$$

Simple conditioning on the known value of the data $y$, using the basic property of conditional probability known as Bayes' rule, yields the posterior density:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}, \tag{26}$$

where $p(y) = \sum_\theta p(\theta)p(y|\theta)$, and the sum is over all possible values of $\theta$ (or $p(\theta) = \int p(\theta)p(y|\theta)d\theta$ in the case of continuous $\theta$). An equivalent form of (26) omits the factor $p(y)$, which does not depend on $\theta$ and, with fixed $y$, can thus be considered a constant, yielding the unnormalized posterior density, which is the right side of:

$$p(\theta|y) \propto p(\theta)p(y|\theta). \tag{27}$$

These simple expressions encapsulate the technical core of Bayesian inference: the primary task of any specific application is to develop the model $p(\theta, y)$ and perform the necessary computations to summarize $p(\theta|y)$ in appropriate ways.

Using Bayes' rule with a chosen probability model means that the data $y$ affect the posterior inference only through the function $p(y|\theta)$, which, when regarded as a function of $\theta$, for fixed $y$, is called the likelihood function. In this way Bayesian inference obeys what is sometimes called the likelihood principle, which states that for a given sample of data, any two probability models $p(y|\theta)$ that have the same likelihood function yield the same inference for $\theta$ [2].

## 4.2 Multiparameter Models

Virtually every practical problem in statistics involves more than one unknown or unobservable quantity. It is in dealing with such problems that the simple conceptual framework of the Bayesian approach reveals its principal advantages over other methods of inference. Although a problem can include several parameters of interest, conclusions will often be drawn about one, or only a few, parameters at a time. In this case, the ultimate aim of a Bayesian analysis is to obtain the marginal posterior distribution of the particular parameters of interest. In principal, the route to achieving this aim is clear: we first require the joint posterior distribution of all unknowns that are not of immediate interest to obtain the desired marginal distribution. Or equivalently, using simulation, we draw samples from the joint posterior distribution and then look at the parameters of interest and ignore the values of the other unknowns. Parameters of this kind are often called nuisance parameters.

To express the idea of joint and marginal posterior distributions mathematically, suppose $\theta$ has two parts, each of which can be a vector, $\theta = (\theta_1, \theta_2)$, and further suppose that we are only interested in inference for $\theta_1$, so $\theta_2$ may be considered a 'nuisance' parameter. We seek conditional distribution of the parameter of interest given the observed data; in this case, $p(\theta_1|y)$. This is derived from the joint posterior density,

$$p(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2)p(\theta_1, \theta_2), \tag{28}$$

by averaging over $\theta_2$:

$$p(\theta_1|y) = \int p(\theta_1, \theta_2|y)d\theta_2. \tag{29}$$

Alternatively, the joint posterior density can be factored to yield

$$p(\theta_1|y) = \int p(\theta_1|\theta_2, y)p(\theta_2|y)d\theta_2, \tag{30}$$

which shows that the posterior distribution of interest, $p(\theta_1|y)$, is a mixture of the conditional posterior distributions given the nuisance parameter, $\theta_2$. The weights depend on the posterior density of $\theta_2$ and thus on a combination of evidence from data and prior modeling. We rarely evaluate the integral (30) explicitly, but it suggests an important practical strategy for both constructing and computing with multiparameter models. Posterior distributions can be computed by marginal and conditional simulation, first drawing $\theta_2$ from its marginal posterior distribution and then $\theta_1$ from its conditional posterior distribution, given the drawing of $\theta_2$.

We will perform Bayesian analysis on this topic. Let's begin with multinomial model. Assume $n = (n_{11}, n_{12}, n_{21})$, $p = (p_{11}, p_{12}, p_{21})$, where $p_{ij},\ i, j = i, 2$, are the multinomial success probabilities in each cell as mentioned above. Thus the likelihood function is

$$P(n|p, N) = \binom{N}{n_{11}\ n_{12}\ n_{21}\ N - \sum n_{ij}} p_{11}^{n_{11}}\ p_{12}^{n_{12}}\ p_{21}^{n_{21}}\ (1 - \sum p_{ij})^{N - \sum n_{ij}}. \tag{31}$$

The distribution is typically thought of as implicitly conditioning on the number of observations. The conjugate prior distribution is a multivariate generalization of the beta distribution known as Dirichlet,

$$P(p|\alpha) \propto p_{11}^{\alpha_1 - 1}\ p_{12}^{\alpha_2 - 1}\ p_{21}^{\alpha_3 - 1}\ (1 - \sum p_{ij})^{\alpha_4 - 1}. \tag{32}$$

A uniform density is obtained by setting $\alpha_i = 1$ for all $i$; this distribution assigns equal density to any vector $p$. Setting $\alpha_j = 0$ for all $j$ results in an improper prior distribution that is uniform in the $\log(p_{ij})$'s. The resulting posterior distribution is proper if there is at least one observation in each of the three categories, so that each component of $n$ is positive. We continue to use a noninformative prior distribution for $N$, $P(N) \propto N^{-2}$.

Therefore, the joint posterior distribution for $p$ and $N$ is

$$P(N, p|n) \propto P(N, p)P(n|N, p) \propto p(N)P(p)P(n|N, p)$$

$$= \binom{N}{n_{11}\ n_{12}\ n_{21}\ N-\sum n_{ij}} p_{11}^{n_{11}+\alpha_1-1}\ p_{12}^{n_{12}+\alpha_2-1}\ p_{21}^{n_{21}+\alpha_3-1}\ (1-\sum p_{ij})^{N-\sum n_{ij}+\alpha_4-1} N^{-2}. \tag{33}$$

Given $N$ and $n$, the components of $p$ have independent posterior densities that are of the form $p_{11}^{A}\ p_{12}^{B}\ p_{21}^{C}\ (1-\sum p_{ij})^{D}$, that is, Dirichlet densities since $P(N|n) = P(N,P|n)/P(p|N,n)$ should be free of $p$. Thus the joint density is

$$P(p|N,n) = \frac{\Gamma(N+\alpha_1+\alpha_2+\alpha_3+\alpha_4)}{\Gamma(n_{11}+\alpha_1)\Gamma(n_{12}+\alpha_2)\Gamma(n_{21}+\alpha_3)\Gamma(N-\sum n_{ij}+\alpha_4)}$$
$$\times p_{11}^{n_{11}+\alpha_1-1}\ p_{12}^{n_{12}+\alpha_2-1}\ p_{21}^{n_{21}+\alpha_3-1}\ F(1-\sum p_{ij})^{N-\sum n_{ij}+\alpha_4-1}. \tag{34}$$

Choosing $\alpha_j = 1/2$ for all $j$, the posterior distribution for $N$ is obtained,

$$P(N|n) = \frac{\Gamma(N+1)\ \Gamma(n_{11}+1/2)\ \Gamma(n_{12}+1/2)\ \Gamma(n_{21}+1/2)\ \Gamma(N-\sum n_{ij}+1/2)}{\Gamma(N+2)\ \Gamma(n_{11}+1)\ \Gamma(n_{12}+1)\ \Gamma(n_{21}+1)\ \Gamma(N-\sum n_{ij}+1)}. \tag{35}$$

Hence the distribution for population size could be simulated whenever the number of detected and undetected individuals in the frist and second sample are given. The following strategy is used here:

- Sample the cell probabilities $p_{ij}$ from prior distribution, $p_{ij} \sim$ Gamma(1/2,1) which form a joint distribution $P(p) \sim$ Dirichlet(1/2,1/2,1/2,1/2).

- Sample the $n_{ij}$ from multinomial distribution with success probability $p_{ij}$ .

- For each $n = (n_{11}, n_{12}, n_{21})$, sample $N$ from its marginal posterior distribution $P(N|n)$ and so the confidence interval for $N$ can be obtained through the previous procedure. Approximate minimum coverage, mean coverage probabilities, and the median width for confidence intervals of $N$ are listed in Table 8.

- The fixed samples in multinomial distribution would allow us to apply Bayesian analysis for the hypergeometric model in a convenient way. The approximate coverage probability and width are shown in Tables 9 and 10.

Table 8: Coverage Property Summary of Intervals in Bayesian Analysis

| | N | Appro Min Cov | Appro Mean Cov | Appro Med Width |
|---|---|---|---|---|
| $1-\alpha=.90$ | 50 | 0.0000 | 0.7126 | 40.7900 |
| | 100 | 0.0000 | 0.7378 | 78.1050 |
| | 150 | 0.0000 | 0.7422 | 114.6330 |
| | 200 | 0.0000 | 0.7441 | 149.8555 |
| | 300 | 0.0000 | 0.7470 | 222.1775 |
| $1-\alpha=.95$ | 50 | 0.0000 | 0.8062 | 94.6447 |
| | 100 | 0.0000 | 0.8431 | 180.1073 |
| | 150 | 0.0000 | 0.8676 | 266.0961 |
| | 200 | 0.0002 | 0.8656 | 352.4288 |
| | 300 | 0.0000 | 0.8738 | 510.3923 |
| $1-\alpha=.99$ | 50 | 0.0002 | 0.8250 | 137.0191 |
| | 100 | 0.0003 | 0.8620 | 265.3556 |
| | 150 | 0.0003 | 0.8834 | 385.6087 |
| | 200 | 0.0002 | 0.8912 | 513.6499 |
| | 300 | 0.0001 | 0.9015 | 770.7565 |

Compared with the Frequentist methods, the Bayesian analysis used here does not show better performance than the new Wilson and Wilson-cc intervals proposed in this thesis. However it tends to grow steadily as $N$ increases; Wilson and Wilson-cc intervals appear more oscillating among the small $N$s. In the hypergeometric model, the Bayesian interval guarantees a coverage of probability which is close to $1-\alpha$ even for very low proportion of $n_1$, $n_2$ to $N$ whereas the approximate width in fact is too large to make sense. The Wilson-cc interval does not present the same characteristic. As a whole, the Bayesian interval provides a different and stable approach on the estimation of population size for capture-recapture problems. Whether the informative prior and hierarchical models should be used to improve

Table 9: Coverage Property in Bayesian Analysis with Fixed Sample Size for $1 - \alpha = .90$

| N | $n_1, n_2$ | Appro Cov | Appro Width |
|---|---|---|---|
| 50 | 5 | 0.9221 | 334.5250 |
| 50 | 10 | 0.8758 | 99.3000 |
| 50 | 15 | 0.8804 | 59.5750 |
| 50 | 20 | 0.8948 | 37.1508 |
| 100 | 5 | 0.8004 | 441.4519 |
| 100 | 10 | 0.9198 | 276.6543 |
| 100 | 20 | 0.8960 | 138.1500 |
| 100 | 40 | 0.8928 | 50.2471 |
| 200 | 10 | 0.9152 | 1071.0750 |
| 200 | 20 | 0.9100 | 410.8256 |
| 200 | 40 | 0.8926 | 191.2755 |
| 200 | 60 | 0.8916 | 108.2107 |
| 500 | 10 | 0.8296 | 1307.5758 |
| 500 | 25 | 0.8864 | 1319.7254 |
| 500 | 50 | 0.8926 | 656.5759 |
| 500 | 100 | 0.8816 | 291.4511 |

the level of coverage probabilities or not will remain an issue of debate in future study.

Table 10: Coverage Property in Bayesian Analysis with Fixed Sample Size for $1-\alpha = .95$

| N | $n_1, n_2$ | Appro Cov | Appro Width |
|---|---|---|---|
| 50 | 5 | 0.9292 | 548.6125 |
| 50 | 10 | 0.9430 | 135.2625 |
| 50 | 15 | 0.9444 | 75.6256 |
| 50 | 20 | 0.9338 | 45.2000 |
| 100 | 5 | 0.9024 | 750.8755 |
| 100 | 10 | 0.9382 | 404.9625 |
| 100 | 20 | 0.9300 | 178.6749 |
| 100 | 40 | 0.9388 | 61.0443 |
| 200 | 10 | 0.9218 | 1537.7123 |
| 200 | 20 | 0.9533 | 559.9873 |
| 200 | 40 | 0.9333 | 231.9748 |
| 200 | 60 | 0.9356 | 130.7376 |
| 500 | 10 | 0.8604 | 1953.1769 |
| 500 | 25 | 0.9422 | 1824.5500 |
| 500 | 50 | 0.9406 | 826.7660 |
| 500 | 100 | 0.9348 | 345.6231 |

# 5 CONCLUSION

Different CIs in a generalized multinomial model could show any tendency of difference in coverage probabilities; more specific performance on conditional restriction is given by the hypergeometric model. Although all of the intervals could have a substantially low minimum coverage and an average coverage below $1 - \alpha$, the new Wilson method should certainly be preferred over the Chapman CI based on their coverage criteria and width of intervals. The Bayesian interval provides an alternative estimation. Further studies including prior analyzing was recommended for improving Bayesian inference. For strengthening our conclusion, all of the above intervals need finer partitions on population size, as well as more coverage computation on larger $N$.

# BIBLIOGRAPHY

[1] A. Agresti and B. A. Coull (1998), Approximate is better than exact for interval estimation of Binomial Proportions. *Amer. Statist.*, **45**, 119-126.

[2] A. Gelman, J. Carlin, H. Stern and D. Rubin (2004), *Bayesian Data Analysis*, second edition. London: Chapman & Hall.

[3] C. C. Sekar and W. E. Deming (1949), On a method of estimating birth and death rates and the extent of registration. *J. Amer. Statist. Assoc.* , **44**, 101-115.

[4] Douglas G. Bonett and Robert M. Price (2006), Confidence intervals for a ratio of binomial proportions based on paired data. *Statistics in Medicine*, **25**, 3039-3047.

[5] E. B. Wilson (1927), Probable inference, the law of succession, and statistical inference. *J. Amer. Statist. Assoc.*, **22**, 209-212.

[6] G. A. F. Seber (1982), *The estimate of Animal Abundance*, second edition. London: Griffin.

[7] Steven K. Thompson (1992), *Sampling (Wiley Series in Probability and Statistics)*. New York: Wiley Interscience.

[8] J. N. Darroch (1958), The multiple-recapture census: I. Estimation of a closed population. *Biometrika*, **45**, 343-359.

[9] RG. Newcombe (1998), Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statistics in Medicine*, **17**, 873-890.

[10] RG. Newcombe (1998), Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine*, **17**, 857-872.

# APPENDICES

## Appendix A: Matlab Code for Multinomial Model

```
% population size estimation for multinomial model

clear

%format compact

warning('off')

tic;

for alpha = [ .05 ]

cc = 1 - alpha

z0 = icdf('norm',1 - alpha/2,0,1);


for samp = [50]

n = samp

parms = [];

p11s = [];

p12s = [];

p21s = [];

p22s = [];

%rand('state',4)

%rand('state',sum(100*clock));

y1 = random('gam',.4,1,3000,1);

y2 = random('gam',.4,1,3000,1);

y3 = random('gam',.4,1,3000,1);
```

```
y4 = random('gam',.5,1,3000,1);

sumy = y1 + y2 + y3 +y4 ;

p1s = y1./sumy;

p2s = y2./sumy;

p11s = p1s.*p2s;

p21s = p1s - p11s;

p12s = p2s - p11s;

p22s = ones(3000,1)-p11s-p12s-p21s;

n11= zeros(1);

n12 = zeros(1);

n21 = zeros(1);

k = 1;

while k < n +1

x1 = (0:k)';

x2 = (0:k)';

m = k+1;

n110 = kron(x1,ones(m,1));

n120 = kron(ones(m,1),x2);

F = [n110,n120];

t = find(sum(F') <= k);

F = F(t,:);

n210=(k- sum(F'))';

n11 = [n11', F(:,1)']';
```

```
n12 = [n12’,F(:,2)’]’;
n21 = [n21’,n210’]’;
k=k+1;
n22=n-n11-n12-n21;
nk = gammaln(n+1) - gammaln(n11+1) - gammaln(n12+1)
- gammaln(n21+1) -gammaln(n22+1);

% Chapman estimator
for c=0.25
n0=(n11+n21+1).*(n11+n12+1)./(n11+1)-1;
var=((n11+n21+1).*(n11+n12+1).*(n21+c).*(n12+c))./((n11+1).
*(n11+1).*(n11+2));

% Chapman interval
lbchap = n0-z0*sqrt(var);
ubchap = n0+z0*sqrt(var);
widthchap = ubchap - lbchap;
covchap = [];
totwidthchap= [];

% Wilson estimator ( to estimate var I use delta method)
p11p=(n11+c)./(n11+n12+n21+2);
p1p=(n11+n21+1)./(n11+n12+n21+2);
```

```
p2p=(n11+n12+1)./(n11+n12+n21+2);
p1pp=(n11+n21+2*c)./(n11+n12+n21+3*c);
p2pp=(n11+n12+2*c)./(n11+n12+n21+3*c);
p11pp=(n11+c)./(n11+n12+n21+3*c);
selnp11=sqrt((1-p11p)./((n11+n12+n21+2).*p11p));
selnp1=sqrt((1-p1p)./((n11+n12+n21+2).*p1p
));
selnp2=sqrt((1-p2p)./((n11+n12+n21+2).*p2p));
var1=1./(n11+c)-1./(n11+n12+n21+2)-(n12+n21+2)./((n11+n12+1).
*(n11+n21+1));
end
end
end
m=length(var1);
var3=[];
for i= 1:m
if var1(i)< 0
var3(i) = 0;
else var3(i)= var1(i);
end
end
var2 = var3';
k1=sqrt((var2))./(selnp11+selnp1+selnp2);
```

```
z=z0.*k1;


%lp1=[2.*(n11+n21+2)+z.*z-1-z.*sqrt(z.*z-2-1./(n11+n12+n21+3)+
4.*p1pp.*(n12+1+1))]./(2.*(n11+n12+n21+3+z.*z));
%lp2=[2.*(n11+n12+2)+z.*z-1-z.*sqrt(z.*z-2-1./(n11+n12+n21+3)+
4.*p2pp.*(n21+1+1))]./(2.*(n11+n12+n21+3+z.*z));
%up1=[2.*(n11+n21+2)+z.*z+1+z.*sqrt(z.*z+2-1./(n11+n12+n21+3)+
4.*p1pp.*(n12+1-1))]./(2.*(n11+n12+n21+3+z.*z));
%up2=[2.*(n11+n12+2)+z.*z+1+z.*sqrt(z.*z+2-1./(n11+n12+n21+3)+
4.*p2pp.*(n21+1-1))]./(2.*(n11+n12+n21+3+z.*z));
%lp11=[2.*(n11+1)+z.*z-1-z.*sqrt(z.*z-2-1./(n11+n12+n21+3)+
4.*p11pp.*(n12+n21+2+1))]./(2.*(n11+n12+n21+3+z.*z));
%up11=[2.*(n11+1)+z.*z+1+z.*sqrt(z.*z+2-1./(n11+n12+n21+3)+
4.*p11pp.*(n12+n21+2-1))]./(2.*(n11+n12+n21+3+z.*z));


lp1=[2.*(n11+n21+2)+z.*z-z.*sqrt(z.*z+4.*(1-p1pp).*(n21+n11+2))].
/(2.*(n11+n12+n21+3+z.*z));
lp2=[2.*(n11+n12+2)+z.*z-z.*sqrt(z.*z+4.*(1-p2pp).*(n12+n11+2))].
/(2.*(n11+n12+n21+3+z.*z));
up1=[2.*(n11+n21+2)+z.*z+z.*sqrt(z.*z+4.*(1-p1pp).*(n21+n11+2))].
/(2.*(n11+n12+n21+3+z.*z));
up2=[2.*(n11+n12+2)+z.*z+z.*sqrt(z.*z+4.*(1-p2pp).*(n12+n11+2))].
/(2.*(n11+n12+n21+3+z.*z));
```

```
lp11=[2.*(n11+1)+z.*z-z.*sqrt(z.*z+4.*(1-p11pp).*(n11+1))].

/(2.*(n11+n12+n21+3+z.*z));

up11=[2.*(n11+1)+z.*z+z.*sqrt(z.*z+4.*(1-p11pp).*(n11+1))].

/(2.*(n11+n12+n21+3+z.*z));

end


%Wilson interval

ubwil=(n11+n12+n21+3).*up1.*up2./lp11;

lbwil=(n11+n12+n21+3).*lp1.*lp2./up11;

widthwil=ubwil-lbwil;

covwil = [];

totwidthwil = [];


for i = 1:3000

if p11s(i)>= .0001 & p12s(i)>= .0001 & p21s(i)>= .0001& p22s(i)>= .0001

lnkp = nk + n11.*log(p11s(i)) + n12.*log(p12s(i)) + n21.*log(p21s(i))+

 n22.*log(p22s(i));

prob = exp(lnkp);

ind1 = (lbchap <= n & n <= ubchap);

totchap = sum(prob.*ind1);

covchap = [covchap; totchap];

width1 = widthchap'*prob;

totwidthchap = [totwidthchap; width1];
```

```
ind2 = (lbwil <= n & n <= ubwil);
totwil = sum(prob.*ind2);
covwil = [covwil; totwil];
width2 = widthwil'*prob;
totwidthwil = [totwidthwil; width2];
else
end
end
Chap = [min(covchap),mean(covchap), median(totwidthchap)];
Wil = [min(covwil),mean(covwil), median(totwidthwil)];
toc;
```

## Appendix B: Matlab Code for Hypergeometric Model

```
% This is hypergeometric model using both Chapman interval and

Wilson interval

clear

%format compact

warning('off')

tic;

for alpha = [ .1]

cc = 1 - alpha

z0 = icdf('norm',1 - alpha/2,0,1);


%generate counts and we assume the numbers of first and second sample

 are fixed.

%both of them are 300

for samp = [30]

n = samp

n11= zeros(1);

n12 = zeros(1);

n21 = zeros(1);

k =6;

x1 = (0:k)';

x2 = (0:k)';

m = k+1;
```

```
n11 = kron(x1,ones(m,1));

n12 = kron(ones(m,1),x2);

F = [n11,n12];

t = find(sum(F’) == k);

F = F(t,:);

n11 = F(:,1);

n12 = F(:,2);

n21=(k- n11);

n22=n-n11-n12-n21;


%Chapman estimator

for c=0.25

n0=(n11+n21+1).*(n11+n12+1)./(n11+1)-1;

var=((n11+n21+1).*(n11+n12+1).*(n21+c).*(n12+c))./((n11+1).

*(n11+1).*(n11+1).*(n11+2));


% Wilson estimator ( to estimate var I use delta method)

p11p=(n11+c)./(n11+n12+n21+2);

p1p=(n11+n21+1)./(n11+n12+n21+2);

p2p=(n11+n12+1)./(n11+n12+n21+2);

p1pp=(n11+n21)./(n11+n12+n21);

p2pp=(n11+n12)./(n11+n12+n21);

p11pp=n11./(n11+n12+n21);
```

```
selnp11=sqrt((1-p11p)./((n11+n12+n21+2).*p11p));

selnp1=sqrt((1-p1p)./((n11+n12+n21+2).*p1p));

selnp2=sqrt((1-p2p)./((n11+n12+n21+2).*p2p));


var1=1./(n11+c)-1./(n11+n12+n21+2)-(n12+n21+2)./((n11+n12+1).

*(n11+n21+1));


m=length(var1);

var3=[];

for i= 1:m

if var1(i)< 0

var3(i) = 0

else var3(i)= var1(i);

end

end

end

var2 = var3'

k1=sqrt((var2))./(selnp11+selnp1+selnp2);

z=z0.*k1;


lp1=[2.*(n11+n21+2)+z.*z-1-z.*sqrt(z.*z-2-1./(n11+n12+n21+3)+

4.*p1pp.*(n12+1+1))]./(2.*(n11+n12+n21+3+z.*z));

lp2=[2.*(n11+n12+2)+z.*z-1-z.*sqrt(z.*z-2-1./(n11+n12+n21+3)+
4.*p2pp.*(n21+1+1))]./(2.*(n11+n12+n21+3+z.*z));
up1=[2.*(n11+n21+2)+z.*z+1+z.*sqrt(z.*z+2-1./(n11+n12+n21+3)+
4.*p1pp.*(n12+1-1))]./(2.*(n11+n12+n21+3+z.*z));
up2=[2.*(n11+n12+2)+z.*z+1+z.*sqrt(z.*z+2-1./(n11+n12+n21+3)+
4.*p2pp.*(n21+1-1))]./(2.*(n11+n12+n21+3+z.*z));
lp11=[2.*(n11+1)+z.*z-1-z.*sqrt(z.*z-2-1./(n11+n12+n21+3)+
4.*p11pp.*(n12+n21+2+1))]./(2.*(n11+n12+n21+3+z.*z));
up11=[2.*(n11+1)+z.*z+1+z.*sqrt(z.*z+2-1./(n11+n12+n21+3)+
4.*p11pp.*(n12+n21+2-1))]./(2.*(n11+n12+n21+3+z.*z));

%lp1=[2.*(n11+n21+2)+z.*z-1-z.*sqrt(z.*z-2-1./(n11+n12+n21+3)+
4.*p1pp.*(n12+1+1))]./(2.*(n11+n12+n21+3+z.*z));
%lp2=[2.*(n11+n12+2)+z.*z-1-z.*sqrt(z.*z-2-1./(n11+n12+n21+3)+
4.*p2pp.*(n21+1+1))]./(2.*(n11+n12+n21+3+z.*z));
%up1=[2.*(n11+n21+2)+z.*z+1+z.*sqrt(z.*z+2-1./(n11+n12+n21+3)+
4.*p1pp.*(n12+1-1))]./(2.*(n11+n12+n21+3+z.*z));
%up2=[2.*(n11+n12+2)+z.*z+1+z.*sqrt(z.*z+2-1./(n11+n12+n21+3)+
4.*p2pp.*(n21+1-1))]./(2.*(n11+n12+n21+3+z.*z));
%lp11=[2.*(1+n11)+z.*z-1-z.*sqrt(z.*z-2-1./(n11+n12+n21+3)+
4.*p11pp.*(n12+n21+2+1))]./(2.*(n11+n12+n21+3+z.*z));
%up11=[2.*(1+n11)+z.*z+1+z.*sqrt(z.*z+2-1./(n11+n12+n21+3)+
4.*p11pp.*(n12+n21+2-1))]./(2.*(n11+n12+n21+3+z.*z));
end
```

```
%wilson interval

ubwil=exp(log(n11+n12+n21+3)+log(up1.*up2./lp11));

lbwil=exp(log(n11+n12+n21+3)+log(lp1.*lp2./up11));


%chapman interval

lbchap = n0-z0.*sqrt(var);

ubchap = n0+z0.*sqrt(var);

widthchap = ubchap - lbchap;

widthwil=ubwil-lbwil;


% hypergeometric probability

prob= hygepdf(n11,30,6,6);

ind1 = (lbchap <= n & n <= ubchap);

ind2=(lbwil <= n & n <= ubwil);

totchap = sum(prob.*ind1);

totwil=sum(prob.*ind2);

width1 = widthwil'*prob

width2=widthchap'*prob

totchap

totwil

end

toc;
```

## Appendix C: R code for Multinomial Model in Bayesian Statistics

```
y1<-rgamma(1000,.5,1)

y2<-rgamma(1000,.5,1)

y3<-rgamma(1000,.5,1)

y4<-rgamma(1000,.5,1)

sum<-y1+y2+y3+y4

p11<-y1/sum

p12<-y2/sum

p21<-y3/sum

p22<-y4/sum

for (i in 1:1000){

if (p11[i]<=.0001|p21[i]<=.0001|p12[i]<=.0001|p22[i]<=.0001)

p11[i]<-rgamma(1,.5,1)

p12[i]<-rgamma(1,.5,1)

p21[i]<-rgamma(1,.5,1)

p22[i]<-rgamma(1,.5,1)

p11[i]<-p11[i]/sum(p11[i]+p12[i]+p21[i]+p22[i])

p12[i]<-p12[i]/sum(p11[i]+p12[i]+p21[i]+p22[i])

p21[i]<-p21[i]/sum(p11[i]+p12[i]+p21[i]+p22[i])

p22[i]<-p22[i]/sum(p11[i]+p12[i]+p21[i]+p22[i])}


samplen<-rmultinom(1000,50,c(p11[1],p12[1],p21[1],p22[1]))

n11=samplen[1,]
```

```
n12=samplen[2,]

n21=samplen[3,]


for (i in 2:1000){

samplen<-rmultinom(1000,50,c(p11[i],p12[i],p21[i],p22[i]))

n11<-rbind(n11,samplen[1,])

n12<-rbind(n12,samplen[2,])

n21<-rbind(n21,samplen[3,])}

mm<-dim(samplen)


pp<-function(N,n1,n2,n3){

p1<-lgamma(N+1)+lgamma(n1+1/2)+lgamma(n2+1/2)+lgamma(n3+1/2)+

lgamma(N-n1-n2-n3+1/2)-2*log(N)+700

p2<-lgamma(N+2)+lgamma(n1+1)+lgamma(n2+1)+lgamma(n3+1)+

lgamma(N-n1-n2-n3+1)

p<-p1-p2

return(exp(p))}


s99<-NA

s95<-NA

s90<-NA

mean1<-NA

t=n11+n12+n21+1
```

```
for (j in 1:1000){

sample1<-sample(t[j,1]:5000,1000,prob=pp(t[j,1]:

5000,n11[j,1],n12[j,1],n21[j,1]),replace=T)

for (i in 2:1000){

sample1<-rbind(sample1,sample(t[j,i]:5000,1000,prob=pp(t[j,i]:

5000,n11[j,i],n12[j,i],n21[j,i]), replace=T))

cred.intervals <- apply(sample1,1,quantile,

c(0.01,0.025,.05,.9,0.975,.99))}

ci99<-NA

ci95<-NA

ci90<-NA

for(i in 1:1000){

if(cred.intervals[1,i]<= 50 &  50 <=cred.intervals[6,i])

ci99[i]=1

else

ci99[i]=0

if(cred.intervals[2,i]<= 50 &  50 <=cred.intervals[5,i])

ci95[i]=1

else

ci95[i]=0

if(cred.intervals[3,i]<= 50 &  50 <=cred.intervals[4,i])

ci90[i]=1

else
ci90[i]=0}


s99<-cbind(s99,sum(ci99))

s95<-cbind(s95,sum(ci95))

s90<-cbind(s90,sum(ci90))

mean1<-cbind(mean1,apply(cred.intervals,1,mean))}


mean(s99[2:1000])/99

mean(s95[2:1000])/99

mean(s90[2:1000])/99


apply(mean1[,2:1000],1,quantile,c(.5))
```

## Appendix D: R code for Hypergeometric Model in Bayesian Statistics

```
n1<-100

n2<-100

n11<-rhyper(5000,n1,500-n1,n2)

pp<-function(N,n1,n2,n11){

p1<-lgamma(N-n1+1)-lgamma(N-n1-n2+n11+1)-lgamma(N+1)+

lgamma(N-n2+1)-2*log(N)+700

return(exp(p1))}


t=n1+n2-n11+1

sample1<-sample(t[1]:5000,100,prob=pp(t[1]:

5000,n1,n2,n11[1]),replace=T)


for (i in 2:5000){

sample1<-rbind(sample1,sample(t[i]:

5000,100,prob=pp(t[i]:5000,n1,n2,n11[i]), replace=T))}


cred.intervals <- apply(sample1,1,quantile, c(0.025,.5,0.975))

dim(cred.intervals)


ci<-NA

for(i in 1:5000){

if(cred.intervals[1,i]<= 500 &  500 <=cred.intervals[3,i])
ci[i]=1

else

ci[i]=0}

sum(ci)/5000

median(cred.intervals[3,]-cred.intervals[1,])
```

# VITA

## XIAO ZHANG

| | |
|---|---|
| Personal Data: | Date of Birth: April 4, 1980 |
| | Place of Birth: Dalian, P. R. China |
| | Marital Status: Single |
| Education: | B.S. Mathematics ( Applied Mathematics), Jilin University, Changchun, P. R. China, 2002 |
| | M.S. Mathematics (Pure Mathematics), Jilin University, Changchun, P. R. China, 2004 |
| | M.S. Mathematics, East Tennessee State Univeristy, Johnson city, Tennessee 2007 |
| Professional Experience: | Graduate Assistant, Jilin University, Mathematics College, 2003–2004 |
| | Data Analyst, Associate, Dalian East Market Research Co. LTD, Dalian, P. R. China, 2004–2005 |
| | Graduate Assistant, East Tennessee State University, Mathematics Department, 2005–2007 |