AN ABSTRACT OF THE THESIS OF

KENNETH PAUL BURNHAM for the DOCTOR OF PHILOSOPHY
(Name) (Degree)

in STATISTICS presented on March 9, 1972
(Major) (Date)

Title: ESTIMATION OF POPULATION SIZE IN MULTIPLE CAPTURE-RECAPTURE STUDIES WHEN CAPTURE PROBABILITIES VARY AMONG ANIMALS

Abstract approved: Redacted for Privacy
W. Scott Overton

A robust estimator of population size (N) is developed based on a model for livetrapping, the main features of which are a) population closure, and b) capture probabilities remain constant over trapping occasions, but vary among animals. Moreover, the set of capture probabilities is modelled as a random sample from a probability distribution $F(p)$, $p \in (0, 1]$. Given this model the capture frequencies are a sufficient statistic for N and F, and they have a multinomial distribution.

First a parametric approach is investigated by assuming F is a Beta distribution, $B(\alpha, \beta)$. Some general results are developed for Maximum Likelihood estimation with the multinomial distribution when sample size, N, is an unknown parameter of interest in addition to the cell probabilities being functions of an unknown

parameter. These results are then applied to Maximum Likelihood estimation of N, α and β. It is found that $\hat{N}_{MLE}$ will generally be unsatisfactory for values of N, α and β likely to apply to real livetrapping data.

A nonparametric estimator of population size is developed by restricting attention to linear combinations of the capture frequencies, that is, $\hat{N} = \sum_{i=1}^{t} a_i f_{it}$ for some constants $a_1, \ldots, a_t$. Because the capture frequencies are multinomial random variables $\hat{N}$ has approximately a Normal distribution, and the variance of $\hat{N}$ can easily be estimated.

An extension of the jackknife method of bias reduction is developed and used to generate some specific linear combinations which have good properties as estimators of N. A procedure based on the data is then suggested for choosing one of these estimators for use in any given study. This estimator, $\hat{N}_J$, can be validly used whenever the capture frequencies have a multinomial distribution and the expected number of animals seen, $E(S_t)$, has approximately an expansion as $N + \frac{b_1}{t} + \frac{b_2}{t^2} + \ldots$, where $b_1, b_2, \ldots$ are constants. Simulation evaluation of the properties of $\hat{N}_J$ show it to be quite robust.

Estimation of Population Size in Multiple Capture-Recapture Studies When Capture Probabilities Vary Among Animals

by

Kenneth Paul Burnham

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

June 1972

APPROVED:

Redacted for Privacy

Professor of Statistics

in charge of major

Redacted for Privacy

Acting Chairman of Department of Statistics

Redacted for Privacy

Dean of Graduate School

Date thesis is presented March 9, 1972

Typed by Clover Redfern for Kenneth Paul Burnham

ACKNOWLEDGMENT

The author has benefited greatly from the encouragement and support of his major professor, Scott Overton. Especially appreciated are the opportunities to attend and speak at professional meetings, and the financial aid secured for the author by Dr. Overton.

Special thanks are also given to Drs. Calvin and Guthrie for their contributions to the author's graduate studies.

Support for this research came from an N.I.H. pre-doctoral fellowship and from the Coniferous Forest Biome of the International Biological Program. A grant from the Oregon State University Computer Center supported the computer time necessary to this research.

To all those who have contributed to and encouraged his education, and especially to his wife, the author extends deeply felt thanks.

TABLE OF CONTENTS

LIST OF TABLES

ESTIMATION OF POPULATION SIZE IN MULTIPLE CAPTURE-RECAPTURE STUDIES WHEN CAPTURE PROBABILITIES VARY AMONG ANIMALS

1. INTRODUCTION

There is a large amount of work in the statistical literature on capture-recapture methods (see Cormack, 1968 for a recent review), but most of it has been developed around the idea that capture probabilities are equal for all animals in the population. The purpose of this thesis is to investigate the statistical problem of estimation of population size given a model for livetrapping in which capture probabilities vary among animals. The main feature of this model is the assumption that individual capture probabilities are a random sample from a distribution on the unit interval.

This model was originally used by the author as the basis for a simulation study of livetrapping and estimation of population size (Burnham and Overton, 1969). For a population of size 100, the simulation study investigated a variety of capture probability distributions, mostly Beta distributions, and the degenerate case of constant capture probabilities. The range of average capture probabilities was .01 to .24. Thirty days of trapping were simulated.

This simulation study showed the model was capable of generating data with a very wide range of properties, including those properties commonly found in real livetrapping experiments. Also, it

showed that some conventional estimators based on equal capture probabilities performed poorly when capture probabilities varied among individuals. Because of these results it appeared worthwhile to pursue the problem of estimation of population size given this model.

The first chapter of this thesis gives the model and shows that, in general (i.e., without specifying the distribution of capture probabilities), the frequencies of capture are a sufficient statistic for estimation of population size. Furthermore, these capture frequencies have a multinomial distribution.

After Chapter 1, most of the thesis is oriented to examining two approaches to estimation of population size given the model. First, the Maximum Likelihood estimator is derived assuming that capture probabilities have a Beta distribution. This Maximum Likelihood estimator turns out to be unsatisfactory. A nonparametric approach is then examined and found to be more rewarding.

The nonparametric approach to estimation of population size is based on linear combinations of the capture frequencies. An extension of Quenouille's (1956) jackknife method of bias reduction in estimation is used to generate a finite sequence of estimators of population size. These estimators are also linear combinations of the capture frequencies. A procedure is then proposed for selecting one estimator of this sequence as the estimator of population size to be used.

Advantages of this approach are several. A simple to compute, robust estimator is achieved, based on the sufficient statistics. Because the capture frequencies have a multinomial distribution, the variance of any linear combination of these frequencies can be estimated from the frequencies themselves. Finally, approximate confidence intervals for the population size can be found based on the asymptotic normality of linear combinations of multinomial variables.

The final chapter of the thesis again deals with the model. A test is given for the hypothesis that individual capture probabilities do not change during the livetrapping study. Second, an extension of the model is examined.

2. THE MODEL AND SOME BASIC RESULTS DERIVED FROM IT

2.1 A Model for Multiple Capture-Recapture Studies

Assume that there is a defined population of animals on which livetrapping is to be conducted for a specified number of occasions. A general structural model for this situation may be given as follows:

N_i = the number of individuals in the population on the ith trapping occasion, $i = 1, 2, \ldots, t$.

p_{ji} = probability of capturing the jth individual in the population on the ith trapping occasion.

$$X_{ji} = \begin{cases} 1 \text{ if the jth individual is captured on the ith trapping occasion,} \\ 0 \text{ otherwise.} \end{cases}$$

By introducing some simplifications and relationships a model can be generated that will be useful for making statistical inferences. A key assumption which will be maintained throughout this thesis is population closure: no births, deaths, immigration or emigration affect the population. This implies the population size is a constant (N). Moreover, the same individuals compose this population on each trapping occasion, thus they can be considered as uniquely identified and numbered 1 through N.

The capture probabilities are the crucial part of this model. They will be regarded as constant for any given individual, but variable among individuals. In order to introduce a relationship among these capture probabilities it is assumed they are a random sample from some probability distribution on the unit interval.

Finally, the basic random variables, the X_{ji}'s, will be assumed mutually independent. An assumption like this is necessary because the specification of capture probabilities only gives the marginal distribution of each X_{ji}.

The model may now be given as follows

(2.1.1a) Population closure is assumed,

N = population size.

(2.1.1b) p_j = the probability of capturing the jth individual on any given trapping occasion, $j = 1, \ldots, N$, and $p_1, \ldots, p_N$ are a random sample from a probability distribution $F(p)$, $p \in (0, 1]$.

(2.1.1c) The random variables X_{ji}, $j = 1, \ldots, N$, $i = 1, \ldots, t$, are mutually independent for given $p_1, \ldots, p_N$.

When this model was first conceived it was intended that the capture probabilities have a Beta distribution and estimation of population size would involve estimating the parameters of this

distribution. But many aspects of this model (e.g., the sufficient statistic) do not depend upon the distribution of capture probabilities.

It is not necessary in the model to specify what the trapping occasions are. It is anticipated they will be equal, short periods of time such as days. For convenience it will be assumed throughout that trapping occasions are consecutive days, but days can be translated into more general occasions and the consecutive restriction is not necessary.

2.2 The Sample Space and Some Basic Capture-Recapture Variables

The basic data are the trapping histories of the individuals in the population. These data can be expressed in matrix form as

$$[X_{ji}] = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1t} \\ X_{21} & X_{22} & \cdots & X_{2t} \\ \vdots & \vdots & & \vdots \\ X_{N1} & X_{N2} & \cdots & X_{Nt} \end{bmatrix} .$$

Row j gives the trapping results on individual j, while column i gives the results for the ith day of trapping.

The sample space is the set of all possible 2^{Nt} matrices $[X_{ji}]$ where each element is either zero or one. Let an element of this space be denoted by ω, and let the space be denoted by Ω.

Some basic capture-recapture variables will now be defined as

functions of the matrix $[X_{ji}]$. In what follows, $\delta_{r,s}$ is the Kronecker delta.

$y_{jt} = \sum_{i=1}^{t} X_{ji}$ = the number of times individual j has been captured by the end of t days of trapping.

$n_i = \sum_{j=1}^{N} X_{ji}$ = the total number of captures on day i.

$C_t = \sum_{i=1}^{t} n_i$ = the total number of captures on all t days of trapping.

$S_t = N - \sum_{j=1}^{N} \delta_{0, y_{jt}}$ = the number of individuals seen at least once during the t days of trapping.

$f_{it} = \sum_{j=1}^{N} \delta_{i, y_{jt}}$ = the number of individuals captured exactly i times in t days of trapping. For $i = 1, \ldots, t$, these are the capture frequencies, while f_{0t} is the number of individuals never captured.

It is useful to note the alternative expressions for S_t and C_t as linear functions of the capture frequencies:

$$S_t = \sum_{i=1}^{t} f_{it} ,$$

$$C_t = \sum_{i=1}^{t} i f_{it} .$$

Another useful relationship is $N = f_{0t} + S_t = \sum_{i=0}^{t} f_{it}$, in which S_t is observable, but f_{0t} can not be observed.

Finally, note that the matrix $[X_{ji}]$ can not be observed in its entirety. Only S_t rows will be observed, but this is sufficient for computing the above capture-recapture data because the f_{0t} unobserved rows are composed entirely of zeros.

2.3 The Sufficient Statistic and Its Distribution

Let $\underline{P} = (p_1, \ldots, p_N)'$, be the vector of capture probabilities for the population. For p_j given, $X_{ji} \sim$ Bernoulli (p_j), thus

$$P\{X_{ji} | p_j\} = p_j^{X_{ji}} (1-p_j)^{1-X_{ji}} \qquad X_{ji} = 0, 1.$$

For $\underline{P}$ given, the X_{ji} are mutually independent, therefore their probability distribution is given by

$$P\{[X_{ji}] | \underline{P}\} = \prod_{j=1}^{N} \prod_{i=1}^{t} p_j^{X_{ji}} (1-p_j)^{1-X_{ji}} ,$$

$$= \prod_{j=1}^{N} p_j^{y_{jt}} (1-p_j)^{t-y_{jt}} .$$

This probability distribution is not useful for estimation of N. A useful distribution is obtained by treating $\underline{P}$ as a random sample and averaging over it to obtain the compound distribution of the random variables X_{ji}. Define

$$P\{[X_{ji}]|F\} = \int_0^1 \cdots \int_0^1 P\{[X_{ji}]|\underline{P}\}dF(p_1)\cdots dF(p_N),$$

which reduces to

$$P\{[X_{ji}]|F\} = \prod_{j=1}^{N}\left[\int_0^1 p^{y_{jt}}(1-p)^{t-y_{jt}}dF(p)\right]. \tag{2.3.1}$$

Further simplification is now possible because the y_{jt} take on only the values $0, 1, \ldots, t$:

$$P\{[X_{ji}]|F\} = \prod_{i=0}^{t}\left[\int_0^1 p^i(1-p)^{t-i}dF(p)\right]^{f_{it}},$$

$$= \left[\int_0^1 (1-p)^t dF(p)\right]^{N-S_t}\prod_{i=1}^{t}\left[\int_0^1 p^i(1-p)^{t-i}dF(p)\right]^{f_{it}},$$

$$= \left[\int_0^1 (1-p)^t dF(p)\right]^{N}\prod_{i=1}^{t}\left[\frac{\int_0^1 p^i(1-p)^{t-i}dF(p)}{\int_0^1 (1-p)^t dF(p)}\right]^{f_{it}}.$$

The last form clearly shows that for this compound distribution of the

X_{ji}, the sufficient statistic is the set of capture frequencies $f_{1t}, \ldots, f_{tt}$.

Given this sufficient statistic what is its distribution? This problem can be approached by first finding the distribution of the variables $f_{1t}, \ldots, f_{tt}$ for $\underline{P}$ given, then taking the expectation of this distribution with respect to the distribution of $\underline{P}$. Note that because $N = f_{0t} + \sum_{i=1}^{t} f_{it}$, the distribution of $f_{0t}, f_{1t}, \ldots, f_{tt}$ is the same as that of $f_{1t}, \ldots, f_{tt}$.

Let $\underline{f} = (f_{1t}, \ldots, f_{tt})'$ be a fixed, arbitrary set of capture frequencies. Let $A \subset \Omega$ be the subset of all $\omega \in \Omega$ such that $\underline{f}(\omega) = (f_{1t}, \ldots, f_{tt})'$. By definition

$$P\{\underline{f}|\underline{P}\} = P\{A|\underline{P}\} = \sum_{\omega \in A} P\{\omega|\underline{P}\} .$$

The unconditional distribution of the capture frequencies is defined as

$$P\{\underline{f}|F\} = E_{\underline{P}} P\{\underline{f}|\underline{P}\} = E_{\underline{P}}\{A|\underline{P}\} ,$$

$$= \int_0^1 \cdots \int_0^1 \left[\sum_{\omega \in A} P\{\omega|\underline{P}\} \right] dF(p_1) \cdots dF(p_N) ,$$

$$= \sum_{\omega \in A} \left[\int_0^1 \cdots \int_0^1 P\{\omega|\underline{P}\} dF(p_1) \cdots dF(p_N) \right] ,$$

$$= \sum_{\omega \in A} E_{\underline{P}} P\{\omega | \underline{P}\} = \sum_{\omega \in A} P\{\omega | F\} .$$

From (2.3.1) it is known that

$$P\{\omega | F\} = \prod_{j=1}^{N} \left[\int_0^1 p^{y_{jt}(\omega)} (1-p)^{t-y_{jt}(\omega)} dF(p) \right].$$

The individual values $y_{jt}(\omega)$, $j = 1, \ldots, N$ are not known but for all $\omega \in A$ the capture frequencies are the same, hence it follows that

$$P\{\omega | F\} = \prod_{i=0}^{t} \left[\int_0^1 p^i (1-p)^{t-i} dF(p) \right]^{f_{it}} \quad \forall\ \omega \in A.$$

Combining this result and the formula $P\{\underline{f} | F\} = \sum_{\omega \in A} P\{\omega | F\}$ gives

$$(2.3.2) \qquad P\{\underline{f} | F\} = \mathcal{C}(A) \prod_{i=0}^{t} \left[\int_0^1 p^i (1-p)^{t-i} dF(p) \right]^{f_{it}},$$

where $\mathcal{C}(A)$ denotes the cardinality of A. This cardinality will now be determined.

The element ω, or equivalently, the matrix $[X_{ji}]$ is in A if and only if the corresponding frequencies are $\underline{f}$. First consider the number of ways to assign N individuals to the $t+1$ capture categories such that the given frequencies result. This can be done in

$$\binom{N}{f_{0t} \cdots f_{tt}} = \frac{N!}{(f_{0t}!)\cdots(f_{tt}!)}$$

different ways, as can be shown by elementary arguments.

But this does not give $\mathcal{C}(A)$ because given the number of times each individual is captured there is still considerable freedom to assign the actual days on which captures occurred. In general, if individual j is captured i times there are $\binom{t}{i}$ different combinations of days giving exactly i captures. For the f_{it} individuals captured i times there are exactly $\binom{t}{i}^{f_{it}}$ ways this can occur.

Consequently, for each assignment of individuals to capture classes there are

$$\prod_{i=0}^{t} \binom{t}{i}^{f_{it}}$$

different elements in A. It follows that

$$\mathcal{C}(A) = \binom{N}{f_{0t} \cdots f_{tt}} \prod_{i=0}^{t} \binom{t}{i}^{f_{it}} .$$

Substituting this formula for $\mathcal{C}(A)$ into (2.3.2) gives the final result:

$$P\{\underline{f} \mid F\} = \binom{N}{f_{0t} \cdots f_{tt}} \prod_{i=0}^{t} \left[\int_0^1 \binom{t}{i} p^i (1-p)^{t-i} dF(p) \right]^{f_{it}} . \tag{2.3.3}$$

This is a multinomial distribution with cell probabilities being known functions of the distribution of capture probabilities.

For convenience, define these cell probabilities to be

$$(2.3.4) \qquad \pi_i = \pi_i(F) = \int_0^1 \binom{t}{i} p^i (1-p)^{t-i} dF(p) \qquad i = 0, 1, \ldots, t,$$

then (2.3.3) can be written simply as

$$(2.3.5) \qquad P\{f_{1t}, \ldots, f_{tt} \mid F\} = \binom{N}{f_{0t} \cdots f_{tt}} \prod_{i=0}^{t} (\pi_i)^{f_{it}} .$$

Many results concerning the capture frequencies and linear combinations of them can be derived from (2.3.4) and (2.3.5), for example the unconditional expectations of S_t and C_t. Because $S_t = N - f_{0t}$

$$(2.3.6) \qquad E(S_t) = N(1-\pi_0).$$

Using

$$C_t = \sum_{i=1}^{t} i f_{it} = \sum_{i=0}^{t} i f_{it}$$

it follows that

$$(2.3.7) \qquad E(C_t) = NtE(p) .$$

2.4 Some Basic Results for the Case When Capture Probabilities Have a Beta Distribution

The Beta distributions on the unit interval are a class of two parameter, absolutely continuous distributions with the density function given by (Johnson and Kotz, 1970):

$$f(p;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, p^{\alpha-1}(1-p)^{\beta-1} \qquad p \in (0,1),\ \alpha > 0,\ \beta > 0 .$$

The mean and variance of this distribution are given by

$$E(p) = \frac{\alpha}{\alpha+\beta},$$

$$V(p) = \frac{E(p)(1-E(p))}{\alpha+\beta+1} .$$

This is a very rich class of distributions in the sense that the density function can take many different shapes depending upon the values of the parameters α and β.

If $E(p)$ is fixed, then letting α and $\beta \uparrow +\infty$ gives a limit distribution that is degenerate at $E(p)$. Thus the case of constant capture probabilities $(p_j = E(p),\ \text{all } j)$ can be thought of as included in the Beta distributions.

Because the Beta distributions are indexed by the parameters α and β, it is convenient to rewrite (2.3.4) as

$$\pi_i(a,\beta) = \int_0^1 \binom{t}{i} p^i (1-p)^{t-i} f(p;a,\beta)\,dp \qquad i = 0, 1, \ldots, t.$$

Carrying out the above integration yields

$$(2.4.1) \qquad \pi_i(a,\beta) = \binom{t}{i} \frac{\Gamma(a+\beta)}{\Gamma(a)\Gamma(\beta)} \frac{\Gamma(a+i)\Gamma(\beta+t-i)}{\Gamma(a+\beta+t)} .$$

In order to compute the value of π_i a recursive relationship will be developed between π_i and π_{i+1}:

$$\pi_{i+1} = \binom{t}{i+1} \frac{\Gamma(a+\beta)}{\Gamma(a)\Gamma(\beta)} \frac{\Gamma(a+i+1)\Gamma(\beta+t-(i+1))}{\Gamma(a+\beta+t)} ,$$

$$\pi_{i+1} = \left(\frac{t-i}{i+1}\right)\left(\frac{a+i}{\beta+t-i-1}\right)\binom{t}{i} \frac{\Gamma(a+\beta)}{\Gamma(a)\Gamma(\beta)} \frac{\Gamma(a+i)\Gamma(\beta+t-i)}{\Gamma(a+\beta+t)} ,$$

$$(2.4.2) \qquad \pi_{i+1} = \pi_i \left[\frac{(t-i)(a+i)}{(i+1)(\beta+t-i-1)}\right] \qquad i = 0, 1, \ldots, t-1.$$

Given π_0 formula (2.4.2) can be used to compute $\pi_1, \ldots, \pi_t$.

Setting $i = 0$ in formula (2.4.1) yields

$$\pi_0 = \frac{\Gamma(a+\beta)}{\Gamma(a+\beta+t)} \frac{\Gamma(\beta+t)}{\Gamma(\beta)} .$$

Repeated application of the basic property of Gamma functions, $\Gamma(x) = (x-1)\Gamma(x-1)$, to $\Gamma(\beta+t)$ and $\Gamma(a+\beta+t)$ in the above formula gives

$$(2.4.3) \qquad \pi_0 = \prod_{i=1}^{t} \left(\frac{\beta+i-1}{a+\beta+i-1}\right) .$$

For any values of t, α and β, formulae (2.4.3) and (2.4.2) are all that are needed to compute $\pi_i(\alpha, \beta)$, $i = 0, 1, \ldots, t$.

Maximum Likelihood estimation of N, α and β will require the partial derivatives of the cell probabilities. For π_0, repeated application of the rule for taking the derivative of products gives

$$\frac{\partial \pi_0}{\partial(\alpha \text{ or } \beta)} = \pi_0 \left[\sum_{i=1}^{t} \frac{\alpha+\beta+i-1}{\beta+i-1} \frac{\partial\left(\frac{\beta+i-1}{\alpha+\beta+i-1}\right)}{\partial(\alpha \text{ or } \beta)} \right].$$

It follows that

$$\frac{\partial \pi_0}{\partial \alpha} = -\pi_0 \left[\sum_{i=1}^{t} \frac{1}{\alpha+\beta+i-1} \right], \tag{2.4.4}$$

and

$$\frac{\partial \pi_0}{\partial \beta} = \alpha \pi_0 \left[\sum_{i=1}^{t} \frac{1}{(\beta+i-1)(\alpha+\beta+i-1)} \right]. \tag{2.4.5}$$

Formulae for recursive computation of the partial derivatives of π_i for $i > 0$ are developed from (2.4.2):

$$\frac{\partial \pi_{i+1}}{\partial \alpha} = \frac{\pi_{i+1}}{\pi_i} \left(\frac{\partial \pi_i}{\partial \alpha}\right) + \frac{\pi_{i+1}}{\alpha+i}, \tag{2.4.6}$$

and

$$\frac{\partial \pi_{i+1}}{\partial \beta} = \frac{\pi_{i+1}}{\pi_i} \left[\frac{\partial \pi_i}{\partial \beta} - \frac{\pi_i}{\beta+t-i-1} \right], \tag{2.4.7}$$

$i = 0, 1, \ldots, t-1$.

2.5 Estimation of Population Size Given the Model

The purpose of this thesis is to investigate the statistical problem of population size estimation given the model of Section 2.1. A natural approach is to assume capture probabilities have a Beta distribution and then use a standard parametric technique such as Maximum Likelihood (ML) estimation with the multinomial distribution given by (2.3.5). Maximum Likelihood estimation of N, α and β with this model is seen to be a special case of a general problem: ML estimation with truncated multinomial data where the cell probabilities are known functions of unknown parameters. Let truncated multinomial data be defined as data having a multinomial sampling distribution with one or more categories which can not be observed and with sample size, N, not known.

Maximum Likelihood estimation for the multinomial distribution when N is known is a standard problem and has been discussed in the statistical literature (Rao, 1958). The problem of ML estimation for the multinomial distribution when N is a parameter of interest does not appear to have received any attention in the literature. Because of this, a decision was made to investigate this general problem in Chapter 3 and then apply those results in Chapter 4 to population size estimation when the capture probabilities have a Beta distribution.

As discussed in Section 4.2 this ML estimator of N was found to be unsatisfactory. However, prior to this discovery a decision had been made to investigate a nonparametric approach to estimation of population size.

Chapters 5 and 6 are devoted to developing and investigating a nonparametric estimation procedure. In Chapter 5 an extension of the technique called jackknifing is developed. In Chapter 6 this extension is applied to the problem of population size estimation to derive a nonparametric estimation procedure which is easy to apply and reasonably robust.

3. MAXIMUM LIKELIHOOD ESTIMATION WITH TRUNCATED MULTINOMIAL DATA

3.1 The Problem and Some General Considerations

Let a random sample of N entities be drawn from a real or conceptual population. Let there be a finite number of mutually exclusive, exhaustive categories into which these entities can be classified. Finally, let the observed frequency of entities belonging to category i be f_i, $i = 1, \ldots, \ell$, $\ell \geq 2$. Then $f_1, \ldots, f_\ell$ are multinomial random variables (Johnson and Kotz, 1969) and the probability of any given sample of frequencies is specified by

$$P\{f_1, \ldots, f_\ell | N, p_1, \ldots, p_\ell\} = \binom{N}{f_1 \cdots f_\ell} \prod_{i=1}^{\ell} (p_i)^{f_i},$$

where

$$N = \sum_{i=1}^{\ell} f_i, \quad 0 < p_i, \; i = 1, \ldots, \ell, \text{ and } \sum_{i=1}^{\ell} p_i = 1.$$

A common generalization of this situation is to let the cell (i.e., category) probabilities be known functions of r unknown parameters $\theta_1, \ldots, \theta_r$ for $r \leq \ell - 1$. Let $\pi_i = \pi_i(\underline{\theta})$ denote the ith cell probability with $\underline{\theta} = (\theta_1, \ldots, \theta_r)'$ the vector of unknown parameters. It will be assumed that the parameter space Θ is

an open set in r dimensional Euclidean space.

This multinomial probability distribution with N known is a very common model; but, it can, and does, arise when N as well as $\underline{\theta}$ are unknown parameters. The remainder of this section will be devoted to examining some of the consequences which arise when N is unknown and to looking at some heuristic considerations regarding estimation of N and $\underline{\theta}$.

Because $N = \sum_{i=1}^{\ell} f_i$, the assumption that N is not known after sampling is equivalent to the assumption that N is not known prior to sampling from this multinomial distribution and at least one of the categories cannot be observed. Let $\ell = t+m$, $t \geq 1$, $m \geq 1$ and assume without loss of generality that $f_1, \ldots, f_t$ are observable but $f_{t+1}, \ldots, f_{t+m}$ are not observable. In this situation the sample can be called truncated since not only are certain outcomes not observable, but the number of such outcomes is not known.

Since unobservable entities can not be distinguished, a new category may be defined as the union of the unobservable categories. Correspondingly f_0, and π_0 may be defined as $f_0 = \sum_{i=t+1}^{t+m} f_i$, and $\pi_0 = \sum_{i=t+1}^{t+m} \pi_i$. Now the model for truncated multinomial sampling can be reduced without loss of generality to the following canonical form: $f_0, f_1, \ldots, f_t$ for $t \geq 1$, are distributed as multinomial

random variables. The probability of any sample can be given as

$$(3.1.1) \qquad P\{f_1, \ldots, f_t \mid N, \underline{\theta}\} = \binom{N}{f_0 \cdots f_t} \prod_{i=0}^{t} (\pi_i(\underline{\theta}))^{f_i},$$

for $N = \sum_{i=0}^{t} f_i$, and for all $\underline{\theta} \in \Theta$, $\pi_i(\underline{\theta}) > 0$, $i = 0, \ldots, \ell$ with $\sum_{i=0}^{t} \pi_i(\underline{\theta}) = 1$ and neither N nor f_0 are known because N is not known prior to sampling and f_0 can not be observed.

Let the observed frequencies $f_1, \ldots, f_t$ be referred to simply as the frequencies. Let S be the sum of the frequencies, i.e., $S = \sum_{i=1}^{t} f_i$. The marginal distribution of S is $b(N, 1-\pi_0)$, hence $P\{S > 0 \mid N, \underline{\theta}\} = 1 - \pi_0^N$. It will be assumed that this probability is virtually one, so that it is not of practical importance to make inferences conditional on $S > 0$.

To find the exact ML estimates of N and $\underline{\theta}$ it appears necessary to carry out a two stage maximization procedure to find $\hat{N}_{MLE}$ and $\hat{\underline{\theta}}_{MLE}$ which satisfy

$$(3.1.2) \qquad P\{f_1, \ldots, f_t \mid \hat{N}_{MLE}, \hat{\underline{\theta}}_{MLE}\} = \max_{N \geq S} \left[\max_{\underline{\theta} \in \Theta \text{ given } N = n} P\{f_1, \ldots, f_t \mid n, \underline{\theta}\} \right],$$

where N assumes only integer values, and $\overline{\Theta}$ is the closure of Θ. This approach has the advantage that $P\{f_1, \ldots, f_t \mid N, \theta\}$ is a

usual multinomial distribution for any fixed value of N and there is a great deal known about the problem of finding the ML estimator of $\underline{\theta}$ in this situation (e.g., Rao, 1958). This method of finding the ML estimates may require considerable computation, but this is not a serious disadvantage given the ubiquity of high speed computers. A real problem will occur when closed form solutions to (3.1.2) do not exist: there will be no way to derive estimates of the variance-covariance matrix of the ML estimators. In order to derive approximate ML estimators and approximate formulae for the variance-covariance matrix of these estimators two alternative approaches to (3.1.2) will be investigated in this chapter.

One alternative is to treat N as a continuous variable by writing

$$P\{f_1, \ldots, f_t | N, \underline{\theta}\} = \frac{\Gamma(N+1)\pi_0^{N-S}}{\Gamma(N-S+1)(f_1!)\ldots(f_t!)} \prod_{i=1}^{t} (\pi_i)^{f_i} .$$

This gives a continuous function in both $\underline{\theta}$ and N, for $N \geq S_t$, which corresponds exactly with the true likelihood function at all integer values of N in $\{S_t, S_t+1, \ldots\}$. It is reasonable to call this continuous function of $\underline{\theta}$ and N the pseudo-likelihood function; let it be symbolized by $L^*(N, \underline{\theta})$. Let $\hat{N}$ and $\hat{\underline{\theta}}$ give the mode of $L^*(N, \underline{\theta})$; it is reasonable to expect $\hat{N} \doteq \hat{N}_{MLE}$ and $\hat{\underline{\theta}} \doteq \hat{\underline{\theta}}_{MLE}$.

As an example of this sort of approach consider the case $t = 1$ and π_0 known. Then $f_1 = S \sim b(N, 1-\pi_0)$ and Feldman and Fox (1968) have shown that the exact ML estimator of N in this case is the greatest integer part of $S/(1-\pi_0)$, denoted by $[S/(1-\pi_0)]$. Treating the likelihood function as a continuous function of N in this case leads to $\hat{N} \doteq S/(1-\pi_0)$; and, it can be shown that even for the exact value of $\hat{N}$, $P\{|\hat{N}-[S/(1-\pi_0)]| < 1\} = 1$.

Working with $L^*(N, \underline{\theta})$ allows partial derivatives to be computed with respect to N as well as $\theta_1, \ldots, \theta_r$, thus the usual Fisher information matrix (Rao, 1965) for N and $\underline{\theta}$ can be computed and its inverse used as an approximation to the dispersion matrix of $(\hat{N}, \hat{\underline{\theta}})$. This procedure is not rigorously justified, but it has been successfully used before with integer valued parameters (e.g., Lewontin and Prout, 1956).

The other alternative to (3.1.2) is based on the following partition of (3.1.1) into two parts:

$$P\{f_1, \ldots, f_t | N, \underline{\theta}\} = P\{S | N, \underline{\theta}\} P\{f_1, \ldots, f_t | S, \underline{\theta}\} \tag{3.1.3}$$

or explicitly

$$P\{f_1, \ldots, f_t | N, \underline{\theta}\} = \left[\binom{N}{S}(1-\pi_0)^S \pi_0^{N-S}\right]\left[\binom{S}{f_1 \cdots f_t} \prod_{i=1}^{t} \left(\frac{\pi_i}{1-\pi_0}\right)^{f_i}\right]. \tag{3.1.4}$$

As indicated in (3.1.3) and clearly shown in (3.1.4), the conditional distribution of the frequencies given their sum is completely

free of N. Thus if N is only a nuisance parameter this conditional distribution provides a basis for inference about $\underline{\theta}$ which is not contingent on any estimated value of N. In case N is a parameter of interest, then an estimator of N can be defined as $\tilde{N} = S/(1-\tilde{\pi}_0)$, where $\tilde{\pi}_0 = \pi_0(\tilde{\underline{\theta}})$ and $\tilde{\underline{\theta}}$ is the ML estimator of $\underline{\theta}$ obtained from $P\{f_1, \ldots, f_t | S, \underline{\theta}\}$. There are a number of heuristic arguments in support of this procedure.

The relationship $N = S + f_0$ shows that point estimation of N is equivalent to estimating f_0, and because S is known after sampling it would appear reasonable to estimate f_0 conditional on the observed value of S. These considerations imply using the conditional distribution of the frequencies to estimate N and $\underline{\theta}$. Furthermore, the relationships $E(f_0) = N\pi_0$ and $N = S + f_0$ suggest it is reasonable to expect estimators of N, π_0 and f_0 to satisfy the relationships $\tilde{f}_0 = \tilde{N}\tilde{\pi}_0$ and $\tilde{N} = S + \tilde{f}_0$ whether or not these estimators are conditional on S. These two relationships imply $\tilde{N} = S/(1-\tilde{\pi}_0)$.

While $\tilde{N}$ and $\tilde{\underline{\theta}}$ may be thought of as reasonable estimators on their own merits, it is possible to think of them as approximations to $\hat{N}$ and $\hat{\underline{\theta}}$. The latter estimators maximize the pseudo-likelihood function simultaneously in N and $\underline{\theta}$. The estimators $\tilde{N}$ and $\tilde{\underline{\theta}}$ have the property that $\tilde{\underline{\theta}}$ gives $\max_{\underline{\theta} \in \tilde{\Theta}} P\{f_1, \ldots, f_t | S, \underline{\theta}\}$

and then given $\tilde{\underline{\theta}}$, $\tilde{N}$ approximately determines $\max_{N \geq S} P\{S|N, \tilde{\underline{\theta}}\}$. It is apparent that $P\{S|N, \underline{\theta}\}$ supplies very little information about $\underline{\theta}$ so it is to be expected that

$$P\{S|\tilde{N}, \tilde{\underline{\theta}}\}P\{f_1, \ldots, f_t|S, \tilde{\underline{\theta}}\} \doteq P\{f_1, \ldots, f_t|\hat{N}, \hat{\underline{\theta}}\}.$$

The above two approximations to the exact ML estimators will be investigated and compared in Section 3.2. Simple conditions will be given under which both procedures yield nearly the same estimates. In Section 3.3 formulae for the corresponding variance-covariance matrices will be developed and compared.

It has been implied that when closed form solutions to (3.1.2) exist, it would be unnecessary to consider alternative procedures. The existence of closed form solutions would seem to depend primarily upon the simplicity of the functions π_i, $i = 0, 1, \ldots, t$. But, in order that meaningful statistical inferences can be made about all the parameters it is clear that $r < t$ is required, for otherwise the number of functionally independent parameters would exceed the number of observations. This practical requirement means that the cell probabilities are not going to be trivial functions of $\underline{\theta}$ and it should be expected that often they will be nonlinear functions. Thus it may be expected that closed form solutions to (3.1.2) will not generally exist.

An example in which unique closed form ML estimators of N

and $\underline{\theta}$ exist yet are quite unsatisfactory because the number of parameters is $t+1$ is given by the multinomial distribution with unconstrained cell probabilities. Let $\underline{\theta} = (p_1, \ldots, p_t)'$, $\pi_i(\underline{\theta}) = p_i$, $i = 1, \ldots, t$, and $\pi_0(\underline{\theta}) = 1 - p_1 - \cdots - p_t = p_0$; finally, assume that $0 < p_i$, $i = 1, \ldots, t$. When these functions are substituted into (3.1.4) it appears that p_0 is confounded with every other parameter.

The exact ML estimators in this example are found by the sequential maximization procedure indicated in (3.1.2). For a fixed integer value of $N \geq S$, with $f_0 = N - S$ it follows

$$\max_{\substack{\underline{\theta} \in \overline{\Theta} \\ N \text{ fixed}}} P\{f_1, \ldots, f_t | N, \underline{\theta}\} = \binom{N}{f_0 \cdots f_t} \prod_{i=0}^{t} \left(\frac{f_i}{N}\right)^{f_i}.$$

The estimators corresponding to this maximum are $\hat{p}_i(N) = f_i/N$, $i = 1, \ldots, t$ and $\hat{p}_0(N) = (N-S)/N$.

Next, the right hand side of the above formula can be rewritten in the form of (3.1.4) to get

$$\max_{\substack{\underline{\theta} \in \overline{\Theta} \\ N \text{ fixed}}} P\{f_1, \ldots, f_t | N, \underline{\theta}\} = \left[\binom{N}{S}\left(\frac{S}{N}\right)^S\left(1-\frac{S}{N}\right)^{N-S}\right]\left[\binom{S}{f_1 \cdots f_t} \prod_{i=1}^{t} \left(\frac{f_i}{S}\right)^{f_i}\right].$$

The term $\binom{N}{S}(S/N)^S(1-S/N)^{N-S}$ is a binomial probability and it is easily seen that this term equals one iff $N = S$. Consequently in this

example

$$\max_{N \in \{S,S+1,\ldots\}} \left[\max_{\substack{\underline{\theta} \in \overline{\Theta} \\ N \text{ fixed}}} P\{f_1, \ldots, f_t \mid N, \underline{\theta}\} \right] = \binom{S}{f_1 \cdots f_t} \prod_{i=1}^{t} \left(\frac{f_i}{S}\right)^{f_i},$$

and this absolute maximum is achieved at the unique ML estimators $\hat{N}_{MLE} = S$, $\hat{p}_{i,MLE} = \frac{f_i}{S}$, $i = 1, \ldots, t$ and $\hat{p}_{0,MLE} = 0$. It is clear that these estimators are not satisfactory.

This example is of particular interest because it represents the trapping experiment under an arbitrary distribution of capture probabilities. The above solution, $\hat{N}_{MLE} = S$ under arbitrary F, will be employed in the construction of the jackknife estimator of N. Also the above result stimulates the investigation of means of restricting the $\pi_i(\underline{\theta})$, notably by specifying a Beta distribution for capture probabilities.

3.2 The Approximate Maximum Likelihood Estimators

Let $f_0, f_1, \ldots, f_t$ for $t \geq 2$ be multinomial random variables with the probability distribution specified by (3.1.1). It is assumed that these are truncated, that is, N is not known and f_0 can not be observed. Furthermore, it is assumed that $1 \leq r \leq t-1$, and that the parameter space Θ is an open set in r dimensional Euclidean space. It will also be assumed that the cell probability functions, $\pi_0, \ldots, \pi_t$, have second partial derivatives with respect

to all components of $\underline{\theta}$. These assumptions and two more made below will be maintained throughout Sections 3.2 and 3.3.

This section, and the following one, are devoted to examination of the two approximations defined in Section 3.1 to the exact ML estimators of N and $\underline{\theta}$. Under the following two assumptions these two estimators have approximately the same distribution:
a) with probability approaching 1, large sample theory can be used with the conditional probability distribution of $f_1, \ldots, f_t$ given S, and b) $P\{S = N | N, \underline{\theta}\} \doteq 0$.

Because $P\{f_1, \ldots, f_t | S, \underline{\theta}\}$ is a multinomial distribution, large sample theory can be used if S is large enough. Assumption (a) thus implies N must be fairly large, say $N \geq 100$, and also implies $P\{S = 0 | N, \underline{\theta}\} \doteq 0$. As a consequence then of assumption (a), it is not necessary to make inferences conditional on $S > 0$. It is seen that except for the necessity of large N, assumptions (a) and (b) are really assumptions about the marginal distribution of S. These assumptions are admitedly vague, but it should be possible to make a judgement as to their applicability in any given problem.

Corresponding to the factorization

$$P\{f_1, \ldots, f_t | N, \underline{\theta}\} = P\{S | N, \underline{\theta}\} P\{f_1, \ldots, f_t | S, \underline{\theta}\},$$

let the pseudo-likelihood function be written as

$$L^*(N, \underline{\theta}) = \left[\frac{\Gamma(N+1)}{S!\,\Gamma(N-S+1)}(1-\pi_0)^S \pi_0^{N-S}\right] L(\underline{\theta}|S),$$

where

$$L(\underline{\theta}|S) = \binom{S}{f_1 \cdots f_t} \prod_{i=1}^{t} \left(\frac{\pi_i}{1-\pi_0}\right)^{f_i} .$$

Let $\hat{N}$ and $\hat{\underline{\theta}}$ satisfy

$$L^*(\hat{N}, \hat{\underline{\theta}}) = \max_{\substack{N \geq S \\ \underline{\theta} \in \Theta}} L^*(N, \underline{\theta}) ,$$

and define the conditional ML estimators $\tilde{N}$ and $\tilde{\underline{\theta}}$ by

$$L(\tilde{\underline{\theta}}|S) = \max_{\underline{\theta} \in \Theta} L(\underline{\theta}|S) ,$$

$$\tilde{N} = \frac{S}{1-\pi_0(\tilde{\underline{\theta}})} .$$

By the assumptions made about the cell probabilities and the parameter space, $\tilde{\underline{\theta}}$ can be determined as the solution (assuming uniqueness) to the r equations

$$(3.2.1) \qquad \frac{\partial \ln L(\underline{\theta}|S)}{\partial \theta_j} = \sum_{i=1}^{t} f_i \frac{1}{\pi_i} \frac{\partial \pi_i}{\partial \theta_j} + \frac{S}{1-\pi_0} \frac{\partial \pi_0}{\partial \theta_j} = 0 \qquad j = 1, \ldots, r.$$

The estimators $\hat{N}$ and $\hat{\underline{\theta}}$ may be found as the solution (assuming uniqueness) to the set of equations

$$(3.2.2a)\qquad \frac{\partial \ln L^{*}(N,\underline{\theta})}{\partial N} = \begin{cases} \frac{1}{N} + \frac{1}{N-1} + \ldots + \frac{1}{N-S+1} + \ln \pi_0 = 0 & S > 0, \\ \ln \pi_0 & S = 0, \end{cases}$$

$$(3.2.2b)\qquad \frac{\partial \ln L^{*}(N,\underline{\theta})}{\partial \theta_j} = \sum_{i=1}^{t} f_i \frac{1}{\pi_i} \frac{\partial \pi_i}{\partial \theta_j} + \frac{N-S}{\pi_0} \frac{\partial \pi_0}{\partial \theta_j} = 0 \qquad j = 1,\ldots,r.$$

An approximation to (3.2.2a) for $S > 0$ will be developed. From a standard formula found in many reference texts it follows that

$$\frac{1}{2} \ln\left(\frac{x+1}{x-1}\right) = \frac{1}{x} + O\left(\frac{1}{x^3}\right) \qquad x > 1.$$

This suggests $\frac{1}{x} \doteq \frac{1}{2} \ln\left(\frac{x+1}{x-1}\right)$, which is a good approximation for $x \geq 10$. Now let $x = N-i$ and sum over $i = 0, 1, \ldots, S-1$ to get

$$\frac{1}{2} \sum_{i=0}^{S-1} \ln\left(\frac{N-i+1}{N-i-1}\right) = \sum_{i=0}^{S-1} \frac{1}{N-i} + O\left(\frac{1}{(N-S)^3}\right),$$

which is equivalent to

$$(3.2.3)\qquad \frac{1}{2} \ln\left[\frac{N(N+1)}{(N-S)(N-S+1)}\right] = \sum_{i=0}^{S-1} \frac{1}{N-i} + O\left(\frac{1}{(N-S)^3}\right).$$

The term $(1/2) \ln [N(N+1)/(N-S)(N-S+1)]$ can be rewritten:

$$\frac{1}{2}\ln\left[\frac{N(N+1)}{(N-S)(N-S+1)}\right] = \frac{1}{2}\ln\left[\left(\frac{N}{N-S}\right)^2\left(\frac{N+1}{N}\right)\left(\frac{N-S}{N-S+1}\right)\right],$$

$$= \ln\left(\frac{N}{N-S}\right) + \frac{1}{2}\ln\left(1+\frac{1}{N}\right) + \frac{1}{2}\ln\left(1-\frac{1}{N-S+1}\right).$$

Substituting this in (3.2.3) yields

$$\ln\left(\frac{N}{N-S}\right) = \sum_{i=0}^{S-1}\frac{1}{N-i} + \frac{1}{2}\ln\left(1+\frac{1}{N}\right) + \frac{1}{2}\ln\left(1-\frac{1}{N-S+1}\right) + O\left(\frac{1}{(N-S)^3}\right). \tag{3.2.4}$$

Using

$$\ln\left(1+\frac{1}{x}\right) = \frac{1}{x} - \frac{1}{2}\left(\frac{1}{x}\right)^2 + O\left(\frac{1}{x^3}\right) \qquad |x| > 1,$$

it follows from (3.2.4) that

$$\ln\left(\frac{N}{N-S}\right) = \sum_{i=0}^{S-1}\frac{1}{N-i} - \frac{S}{2N(N-S+1)} + \frac{1}{4}\left[\frac{1}{(N-S+1)^2} - \frac{1}{N^2}\right] + O\left(\frac{1}{(N-S)^3}\right).$$

Formula (3.2.2a) can now be written as

$$\frac{\partial \ln L^*(N, \underline{\theta})}{\partial N} = \ln\left(\frac{N}{N-S}\right) + \ln \pi_0 + er(N, S),$$

where

$$er(N, S) = \frac{S-1}{2N(N-S+1)} - \frac{1}{4}\left[\frac{1}{(N-S+1)^2} - \frac{1}{N^2}\right] + O\left(\frac{1}{(N-S)^3}\right).$$

Clearly an approximation to Equation (3.2.2a) is

$$\ln\left(\frac{N}{N-S}\right) + \ln \pi_0 = 0,$$

which implies

$$N = \frac{S}{1-\pi_0} .$$

Given this equation it follows that

$$\frac{N-S}{\pi_0} = \frac{S}{1-\pi_0} .$$

Therefore the original pseudo-likelihood equations can be replaced by the approximations

(3.2.5a) $$N = \frac{S}{1-\pi_0} ,$$

(3.2.5b) $$\frac{\partial \ln L^*(N, \underline{\theta})}{\partial \theta_j} = \sum_{i=1}^{t} f_i \frac{1}{\pi_i} \frac{\partial \pi_i}{\partial \theta_j} + \frac{S}{1-\pi_0} \frac{\partial \pi_0}{\partial \theta_j} = 0 \qquad j = 1,\ldots,r.$$

Equations (3.2.5b) and (3.2.1) are identical, therefore they have the common solution $\tilde{\underline{\theta}}$, implying that $\tilde{N}$ and $\tilde{\underline{\theta}}$ are approximately equal to $\hat{N}$ and $\hat{\underline{\theta}}$. The error committed by this approximation is measured entirely by $er(\tilde{N}, S)$. The smaller $er(\tilde{N}, S)$ is, the smaller the error made by using $\tilde{N}$ and $\tilde{\underline{\theta}}$ in place of $\hat{N}$ and $\hat{\underline{\theta}}$. Note that it will be easier to find $\tilde{N}$ and $\tilde{\underline{\theta}}$ as this only requires solving Equations (3.2.5b) for $\tilde{\underline{\theta}}$ and these equations do not depend upon N.

Conditions under which $er(\tilde{N}, S)$ will be small relative to $\ln(\tilde{N}/\tilde{N}-S) + \ln(\tilde{\pi}_0)$ are not clear because of the term $O(1/(\tilde{N}-S)^3)$. As a useful rule of thumb it is suggested that this term can be ignored whenever $\tilde{N}-S \geq 10$. Then $er(\tilde{N}, S)$ can be taken as

$$er(\tilde{N}, S) \doteq \frac{S-1}{2\tilde{N}(\tilde{N}-S+1)} - \frac{1}{4}\left[\frac{1}{(\tilde{N}-S+1)^2} - \frac{1}{(\tilde{N})^2}\right].$$

As $\tilde{N}-S$ increases and $S/\tilde{N}$ decreases, $er(\tilde{N}, S)$ decreases. Even for $\tilde{N}-S = 10$, $S/\tilde{N} = .9$ and $S \geq 30$ (which has high probability by assumption a), $er(\tilde{N}, S) \doteq .04$. If an adjusted estimate $\tilde{N}_1$ is computed as

$$\tilde{N}_1 = \frac{S}{1-\tilde{\pi}_0 \exp(er(\tilde{N}, S))} \doteq \frac{S}{1-.1(1-.04)} = \frac{S}{.896},$$

then $\tilde{N}/\tilde{N}_1 = .9/.896 = 1.0045$. It is concluded that the conditions $\tilde{N}-S \geq 10$, $S/\tilde{N} \leq .9$, $S \geq 30$ are sufficient to insure that $\tilde{N}$ is a good approximation to $\hat{N}$.

3.3 Some Large Sample Variance-Covariance Formulae

Approximations for variances and covariances of $\hat{N}, \hat{\theta}_1, \ldots, \hat{\theta}_r$ and $\tilde{N}, \tilde{\theta}_1, \ldots, \tilde{\theta}_r$ will now be developed. In both cases the basic approach is to compute the Fisher information matrix and use its inverse as the asymptotic, or large sample dispersion (variance-covariance) matrix. With $\hat{N}$ and $\hat{\underline{\theta}}$ this approach leads directly

to an approximation for $V(\hat{N})$. With $\tilde{N}$ and $\underline{\tilde{\theta}}$ the approach is to first compute the dispersion matrix of $\underline{\tilde{\theta}}$ from the distribution $P\{f_1, \ldots, f_t | S, \underline{\theta}\}$. Then $V(\tilde{N}|S)$ is found by applying the usual propagation of error method (Demming, 1943) to $\tilde{N} = S/(1-\pi_0(\underline{\tilde{\theta}}))$. The approach used with $\tilde{N}$ and $\underline{\tilde{\theta}}$ is justifiable as a large sample technique.

The conditional information matrix $I(\underline{\theta}|S)$ is defined as the $r \times r$ matrix of elements $I_{j\ell}(\underline{\theta}|S)$ where

$$I_{j\ell}(\underline{\theta}|S) = -E\left(\frac{\partial^2 \ln L(\underline{\theta}|S)}{\partial\theta_j \partial\theta_\ell}\right) \qquad j, \ell = 1, \ldots, r.$$

The indicated expectation is taken with respect to the distribution $P\{f_1, \ldots, f_t | S, \underline{\theta}\}$.

The $\partial \ln L(\underline{\theta}|S)/\partial\theta_j$ given in (3.2.1) can be rewritten as

$$\frac{\partial \ln L(\underline{\theta}|S)}{\partial\theta_j} = \sum_{i=1}^{t} f_i \left[\frac{1}{\pi_i}\frac{\partial\pi_i}{\partial\theta_j} + \frac{1}{1-\pi_0}\frac{\partial\pi_0}{\partial\theta_j}\right].$$

Then

$$\frac{\partial^2 \ln L(\underline{\theta}|S)}{\partial\theta_j \partial\theta_\ell} = \sum_{i=1}^{t} f_i \left[\frac{-1}{(\pi_i)^2}\left(\frac{\partial\pi_i}{\partial\theta_j}\right)\left(\frac{\partial\pi_i}{\partial\theta_\ell}\right) + \frac{1}{\pi_i}\frac{\partial^2\pi_i}{\partial\theta_j\partial\theta_\ell} + \frac{1}{(1-\pi_0)^2}\left(\frac{\partial\pi_0}{\partial\theta_j}\right)\left(\frac{\partial\pi_0}{\partial\theta_\ell}\right) + \frac{1}{1-\pi_0}\frac{\partial^2\pi_0}{\partial\theta_j\partial\theta_\ell}\right].$$

Using $E(f_i|S) = S\pi_i/(1-\pi_0)$, $i = 1, \ldots, t$, it follows that

$$I_{j\ell}(\underline{\theta}|S) = \frac{S}{1-\pi_0}\left[\sum_{i=1}^{t} \frac{1}{\pi_i}\left(\frac{\partial \pi_i}{\partial \theta_j}\right)\left(\frac{\partial \pi_i}{\partial \theta_\ell}\right) - \frac{1}{1-\pi_0}\left(\frac{\partial \pi_0}{\partial \theta_j}\right)\left(\frac{\partial \pi_0}{\partial \theta_\ell}\right) - \sum_{i=0}^{t} \frac{\partial^2 \pi_i}{\partial \theta_j \partial \theta_\ell}\right].$$

Because $\sum_{i=0}^{t} \pi_i = 1$ for all $\underline{\theta} \in \Theta$, the term involving second partials is identically zero. Adding and subtracting $1/\pi_0(\partial \pi_0/\partial \theta_j)(\partial \pi_0/\partial \theta_\ell)$ inside the brackets leads to the final formula

$$(3.3.1) \quad I_{j\ell}(\underline{\theta}|S) = \frac{S}{1-\pi_0}\left[\sum_{i=0}^{t} \frac{1}{\pi_i}\left(\frac{\partial \pi_i}{\partial \theta_j}\right)\left(\frac{\partial \pi_i}{\partial \theta_\ell}\right) - \frac{1}{\pi_0(1-\pi_0)}\left(\frac{\partial \pi_0}{\partial \theta_j}\right)\left(\frac{\partial \pi_0}{\partial \theta_\ell}\right)\right],$$

$$j, \ell = 1, \ldots, r.$$

To get a convenient expression for $I(\underline{\theta}|S)$ define an $r \times r$ matrix $A = [a_{j\ell}]$ by

$$a_{j\ell} = \sum_{i=0}^{t} \frac{1}{\pi_i}\left(\frac{\partial \pi_i}{\partial \theta_j}\right)\left(\frac{\partial \pi_i}{\partial \theta_\ell}\right) \qquad j, \ell = 1, \ldots, r.$$

Define an $r \times 1$ vector $\underline{b}$ by

$$\underline{b} = -\frac{1}{\pi_0}\left[\frac{\partial \pi_0}{\partial \theta_1}, \ldots, \frac{\partial \pi_0}{\partial \theta_r}\right]';$$

finally, define a scalar

$$c = \frac{1-\pi_0}{\pi_0}.$$

It follows from (3.3.1) that

$$I(\underline{\theta}|S) = \frac{S}{1-\pi_0}\left[A - \frac{1}{c}\underline{b}\underline{b}'\right] . \tag{3.3.2}$$

Assuming the inverse exists, a large sample approximation to the dispersion matrix $D(\tilde{\underline{\theta}}|S)$ is given by $I^{-1}(\underline{\theta}|S)$:

$$D(\tilde{\underline{\theta}}|S) = \frac{1-\pi_0}{S}\left[A - \frac{1}{c}\underline{b}\underline{b}'\right]^{-1} . \tag{3.3.3}$$

An approximation to $V(\tilde{N}|S)$ is obtained by use of the Taylor's series expansion for $\tilde{N}$ about θ, assuming $E(\tilde{\underline{\theta}}) \doteq \underline{\theta}$ and $E(\tilde{N}|S) \doteq S/(1-\pi_0)$:

$$\tilde{N} \doteq \frac{S}{1-\pi_0} + \frac{S}{(1-\pi_0)^2}\left[\frac{\partial \pi_0}{\partial \theta_1}, \ldots, \frac{\partial \pi_0}{\partial \theta_r}\right](\tilde{\underline{\theta}} - \underline{\theta}) .$$

Then

$$V(\tilde{N}|S) \doteq \frac{(S\pi_0)^2}{(1-\pi_0)^4}\underline{b}'D(\tilde{\underline{\theta}}|S)\underline{b} ,$$

or

$$V(\tilde{N}|S) = \left(\frac{S}{1-\pi_0}\right)\left(\frac{1}{c}\right)^2 \underline{b}'\left[A - \frac{1}{c}\underline{b}\underline{b}'\right]^{-1}\underline{b} . \tag{3.3.4}$$

The same technique can be applied to find an approximation to $\underline{Cov}(\tilde{N}, \tilde{\underline{\theta}}|S) = [Cov(\tilde{N}, \tilde{\theta}_1|S), \ldots, Cov(\tilde{N}, \tilde{\theta}_r|S)]'$:

$$\underline{\mathrm{Cov}}\,(\tilde{N}, \underline{\tilde{\theta}} \mid S) \doteq -\frac{1}{c}\left[A - \frac{1}{c}\underline{b}\,\underline{b}'\right]^{-1}\underline{b}.$$

All of the above formulae for variances and covariances are conditional on the observed value of S. This is desirable if N is a nuisance parameter and only inferences about $\underline{\theta}$ are being made. This also seems reasonable with regard to inferences about N itself since $\tilde{N} = S + \tilde{f}_0$ (where $\tilde{f}_0 = S\tilde{\pi}_0/(1-\tilde{\pi}_0)$) and S is known after sampling. This consideration implies that inferences about N should be based on the conditional distribution of $\tilde{N}$ given S. Unfortunately this conditional probability distribution is unsatisfactory for making inferences about N, a point which is developed in the next paragraph.

From large sample theory it follows that $E(\tilde{N}|S) \doteq S/(1-\pi_0) = (S/E(S))N$. Furthermore $\tilde{N}$ is approximately distributed as a normal random variable with mean and variance $(S/E(S))N$ and $V(\tilde{N}|S)$ respectively. In this conditional distribution of $\tilde{N}$, N is confounded with the (conditional) parameter $S/E(S)$. If $S/E(S)$ was approximately 1 with high probability this confounding would cause no problems. But in fact $S \sim b(N, E(S)/N)$ so it is clear that there can be high probability that $S/E(S)$ will not be close enough to 1 to be considered unity. Nor can $S/E(S)$ be estimated except as unity. It is concluded that the conditional probability model is not satisfactory for making inferences about N

because the model does not relate $\tilde{N}$ to N in a useful manner.

Consider, for example, confidence interval construction. With the conditional distribution of $\tilde{N}$ it is possible to construct confidence intervals for $E(\tilde{N}|S)$, but this is not the parameter of interest. In the sequence of repetitions of the experiment giving the same value of S the relative frequence of coverage of N by these intervals will be less than the nominal confidence level. An appropriate way to handle the problem is to work with the unconditional distribution of $\tilde{N}$ wherein $E(\tilde{N}) = E_S E(\tilde{N}|S) \doteq N$.

An approximation to the unconditional variance of $\tilde{N}$ can be found by evaluating $E(\tilde{N}-N)^2$ with respect to $P\{f_1, \ldots, f_t | N, \underline{\theta}\}$. This will be done later. An interesting alternative is to formally compute the information matrix for $\hat{N}$ and $\hat{\underline{\theta}}$ from $L^*(N, \underline{\theta})$. It will be shown that this leads to appropriate formula for $V(\hat{N})$, thus corroborating the usefulness of treating a discrete parameter as if it were continuous.

Define $I(N, \underline{\theta})$ as the $(r+1) \times (r+1)$ matrix of elements $I_{j\ell}(N, \underline{\theta})$ where

$$I_{00}(N, \underline{\theta}) = -E\left[\frac{\partial^2 \ln L^*(N, \underline{\theta})}{\partial N^2} \right],$$

$$I_{0\ell}(N, \underline{\theta}) = -E\left[\frac{\partial^2 \ln L^*(N, \underline{\theta})}{\partial N \partial \theta_\ell} \right] \quad \ell = 1, \ldots, r,$$

$$I_{j\ell}(N, \underline{\theta}) = -E\left[\frac{\partial^2 \ln L^*(N, \underline{\theta})}{\partial\theta_j \partial\theta_\ell} \right] \qquad j, \ell = 1, \ldots, r.$$

From (3.2.2a) and (3.2.2b) it follows that

$$(3.3.5) \qquad \frac{\partial^2 \ln L^*(N, \underline{\theta})}{\partial N^2} = \begin{cases} 0 & S = 0, \\ -\sum_{i=0}^{S-1} \frac{1}{(N-i)^2} & S > 0, \end{cases}$$

$$\frac{\partial^2 \ln L^*(N, \underline{\theta})}{\partial N \partial\theta_\ell} = \frac{1}{\pi_0} \frac{\partial \pi_0}{\partial\theta_\ell} \qquad \ell = 1, \ldots, r,$$

and

$$\frac{\partial^2 \ln L^*(N, \underline{\theta})}{\partial\theta_j \partial\theta_\ell} = -\frac{N-S}{(\pi_0)^2} \left(\frac{\partial \pi_0}{\partial\theta_j}\right)\left(\frac{\partial \pi_0}{\partial\theta_\ell}\right) + \frac{N-S}{\pi_0} \frac{\partial^2 \pi_0}{\partial\theta_j \partial\theta_\ell}$$

$$- \sum_{i=1}^{t} f_i \frac{1}{(\pi_i)^2} \left(\frac{\partial \pi_i}{\partial\theta_j}\right)\left(\frac{\partial \pi_i}{\partial\theta_\ell}\right) + \sum_{i=1}^{t} f_i \frac{1}{\pi_i} \frac{\partial^2 \pi_i}{\partial\theta_j \partial\theta_\ell},$$

$$j, \ell = 1, \ldots, r.$$

The expectations of the second two sets of formulae above can be easily determined:

$$I_{0\ell}(N, \underline{\theta}) = -\frac{1}{\pi_0} \frac{\partial \pi_0}{\partial\theta_\ell} \qquad \ell = 1, \ldots, r,$$

$$I_{j\ell}(N, \underline{\theta}) = N \sum_{i=0}^{t} \frac{1}{\pi_i} \left(\frac{\partial \pi_i}{\partial \theta_j}\right)\left(\frac{\partial \pi_i}{\partial \theta_\ell}\right) \qquad j, \ell = 1, \ldots, r.$$

By noting the previous definitions of the quantities A, $\underline{b}$ and c it is seen that

$$I(N, \underline{\theta}) = \begin{bmatrix} I_{00}(N, \underline{\theta}) & \underline{b}' \\ \underline{b} & NA \end{bmatrix} . \tag{3.3.6}$$

The quantity $I_{00}(N, \underline{\theta})$ equals

$$\sum_{S=1}^{N} \left[\sum_{i=0}^{S-1} \left(\frac{1}{N-i}\right)^2 P\{S \mid N, \underline{\theta}\} \right].$$

As an approximation to $\sum_{i=0}^{S-1} 1/(N-i)^2$, Lewontin and Prout (1956) suggest the formula $S/N(N-S+1)$. This is a good approximation except when S is very close to N, such as $S = N, N-1, \ldots, N-5$. Certainly, for $S \leq N-10$ it appears to be an excellent approximation. By assumption (b), $P\{S = N \mid N, \underline{\theta}\} \doteq 0$, and N is large enough (as a consequence of assumption a) that values of S very near N also have very small probability. When the approximation

$$\sum_{i=0}^{S-1} \frac{1}{(N-i)^2} \doteq \frac{S}{N(N-S+1)}$$

is substituted into $I_{00}(N, \underline{\theta})$ it will not matter that it is poor for values of S near N because these values of S have very small probability.

It follows that a good approximation for $I_{00}(N, S)$ is

$$I_{00}(N, S) \doteq \sum_{S=1}^{N} \frac{S}{N(N-S+1)} \binom{N}{S} (1-\pi_0)^S \pi_0^{N-S}$$

$$= \frac{1-\pi_0}{N\pi_0} \left(1-(1-\pi_0)^N\right) .$$

But $(1-\pi_0)^N = P\{S = N \mid N, \underline{\theta}\} \doteq 0$, thus the final result is

$$I_{00}(N, \underline{\theta}) \doteq \frac{1-\pi_0}{N\pi_0} = \frac{c}{N} . \tag{3.3.7}$$

Combining (3.3.7) and (3.3.6) gives

$$I(N, \underline{\theta}) \doteq \begin{bmatrix} \frac{c}{N} & \underline{b}' \\ \underline{b} & NA \end{bmatrix} .$$

It is worth noting that when N is known the matrix NA is the exact information matrix for $\underline{\theta}$ in the corresponding multinomial distribution. Similarly, if π_0 is known the minimum variance unbiased estimator of N is $S/(1-\pi_0)$ with variance N/c.

Assuming that $I^{-1}(N, \underline{\theta})$ exists, an approximation to the dispersion matrix $D(\hat{N}, \hat{\underline{\theta}})$ is

$$(3.3.8) \qquad D(\hat{N}, \hat{\underline{\theta}}) = \begin{bmatrix} \frac{N}{\lambda} & -\frac{1}{\lambda}\underline{b}'A^{-1} \\ -\frac{1}{\lambda}A^{-1}\underline{b} & \frac{1}{N}(A-\frac{1}{c}\underline{b}\underline{b}')^{-1} \end{bmatrix}$$

where $\lambda = c - \underline{b}'A^{-1}\underline{b}$. Using the formula

$$[A - \frac{1}{c}\underline{b}\underline{b}']^{-1} = A^{-1} + \frac{1}{\lambda}A^{-1}\underline{b}\underline{b}'A^{-1} ,$$

$D(\hat{N}, \hat{\underline{\theta}})$ may be written as

$$\begin{bmatrix} \frac{N}{c} + \frac{N}{\lambda c}\underline{b}'A^{-1}\underline{b}' & -\frac{1}{\lambda}\underline{b}'A^{-1} \\ -\frac{1}{\lambda}A^{-1}\underline{b} & \frac{1}{N}A^{-1} + \frac{1}{N\lambda}A^{-1}\underline{b}\underline{b}'A^{-1} \end{bmatrix}$$

As pointed out previously N/c is the variance of $S/(1-\pi_0)$. Thus $(N/\lambda c)\underline{b}'A^{-1}\underline{b}$ is the additional variance attributed to lack of knowledge of π_0 or rather, in this case, lack of knowledge of $\underline{\theta}$. Similarily the term $(1/N\lambda)A^{-1}\underline{b}\underline{b}'A^{-1}$ is that portion of the dispersion matrix of $\hat{\underline{\theta}}$ attributable to lack of knowledge of N.

Formulae for unconditional variances and covariances of $\tilde{N}, \tilde{\theta}_1, \ldots, \tilde{\theta}_r$ can be developed from the results available on $D(\tilde{\underline{\theta}}|S)$, $V(\tilde{N}|S)$ and $\underline{Cov}(\tilde{N}, \tilde{\underline{\theta}}|S)$. For $V(\tilde{N})$ write

$$E(\tilde{N}-E(\tilde{N}))^2 = E_S E[(\tilde{N}-E(\tilde{N}|S))^2|S] + E_S[E(\tilde{N}|S)-N]^2 ,$$

thus (approximately)

$$V(\tilde{N}) = E_S V(\tilde{N}|S) + V(\frac{S}{1-\pi_0}) ,$$

$$= N(\frac{1}{c})^2 \underline{b}'[A - \frac{1}{c}\underline{b}\,\underline{b}']^{-1}\underline{b} + \frac{N}{c} ,$$

$$= N(\frac{1}{c})^2(\underline{b}'A^{-1}\underline{b})(1+\frac{\underline{b}'A^{-1}\underline{b}}{\lambda}) + \frac{N}{c}$$

$$= N\frac{1}{c}\frac{\underline{b}'A^{-1}\underline{b}}{\lambda} + \frac{N}{c} ,$$

and finally

(3.3.9) $$V(\tilde{N}) = \frac{N}{\lambda} .$$

The unconditional dispersion matrix of $\tilde{\underline{\theta}}$ is given by $D(\tilde{\underline{\theta}}) = E(\tilde{\underline{\theta}}-E(\tilde{\underline{\theta}}))(\tilde{\underline{\theta}}-E(\tilde{\underline{\theta}}))'$. For large samples $E(\tilde{\underline{\theta}}|S) \doteq \underline{\theta}$, therefore $E(\tilde{\underline{\theta}}) \doteq E(\tilde{\underline{\theta}}|S)$. It follows that

$$D(\tilde{\underline{\theta}}) \doteq E_S[E(\tilde{\underline{\theta}}-E(\tilde{\underline{\theta}}|S))(\tilde{\underline{\theta}}-E(\tilde{\underline{\theta}}|S))'|S],$$

$$= E_S D(\tilde{\underline{\theta}}|S),$$

and finally

(3.3.10) $$D(\tilde{\underline{\theta}}) \doteq \frac{1}{N}[A - \frac{1}{c}\underline{b}\,\underline{b}']^{-1} .$$

A similar approach leads to

$$\underline{Cov}(\tilde{N}, \tilde{\underline{\theta}}) = E_S \underline{Cov}(\tilde{N}, \tilde{\underline{\theta}}|S) \doteq -\frac{1}{c}[A - \frac{1}{c}\underline{b}\,\underline{b}']^{-1}b$$

$$= -\frac{1}{c}[A^{-1} + \frac{1}{\lambda}A^{-1}\underline{b}\,\underline{b}'A^{-1}]\underline{b} \quad =$$

$$= -\frac{1}{c}[A^{-1}\underline{b} + \frac{1}{\lambda} A^{-1}\underline{b}(\underline{b}'A^{-1}\underline{b})] ,$$

$$= -\frac{1}{c} A^{-1}\underline{b}[1+\frac{c-\lambda}{\lambda}] ,$$

and finally

$$\underline{\mathrm{Cov}}(\tilde{N}, \tilde{\underline{\theta}}) \doteq -\frac{1}{\lambda} A^{-1}\underline{b} . \tag{3.3.11}$$

From formulae (3.3.9), (3.3.10) and (3.3.11) it is seen that $D(\tilde{N}, \tilde{\underline{\theta}}) \doteq D(\hat{N}, \hat{\underline{\theta}})$ where the latter matrix is given by (3.3.8). This approximate equality has been shown under assumptions (a) and (b) given at the start of Section 3.2. Under these assumptions the approximations $\tilde{N} \doteq \hat{N}$ and $\tilde{\underline{\theta}} \doteq \hat{\underline{\theta}}$ should generally be good, therefore it is expected that the dispersion matrices of these estimators will be approximately the same.

The procedure for evaluating $D(\tilde{N}, \tilde{\underline{\theta}})$ is justified as a large sample technique. The approach of treating N as a continuous variable and using $I^{-1}(N, \underline{\theta}) = D(\hat{N}, \hat{\underline{\theta}})$ as the large sample dispersion matrix of $\hat{N}$ and $\hat{\underline{\theta}}$ is not rigorously justifiable. That this approach gives useful results in this instance is proven by the results $\tilde{N} \doteq \hat{N}$, $\tilde{\underline{\theta}} \doteq \hat{\underline{\theta}}$ and $D(\tilde{N}, \tilde{\underline{\theta}}) \doteq D(\hat{N}, \hat{\underline{\theta}})$.

3.4 The Method of Scoring

A common iterative method for finding ML estimators is the method of scoring (Rao, 1965; Kale, 1961, 1962). Because this

method is used in Chapter 4 it will be briefly outlined here.

Assume that $L(\underline{\theta})$ is a likelihood function for $\underline{\theta}$ in some given parameter space; and, assume that $L(\underline{\theta})$ has second partial derivatives with respect to all components of $\underline{\theta}$. The Maximum Likelihood estimator $\hat{\underline{\theta}}_{MLE}$ is usually found by solving the set of equations

$$\frac{\partial \ln L(\underline{\theta})}{\partial \underline{\theta}} = \begin{bmatrix} \frac{\partial \ln L(\underline{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ln L(\underline{\theta})}{\partial \theta_r} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} .$$

When explicit solutions are not possible $\hat{\underline{\theta}}_{MLE}$ is often found by the Newton-Raphson iterative procedure based on the Taylor's series expansion for $\partial \ln L(\underline{\theta})/\partial \underline{\theta}$:

$$\frac{\partial \ln L(\underline{\theta})}{\partial \underline{\theta}} \doteq \frac{\partial \ln L(\underline{\theta}_0)}{\partial \underline{\theta}} + [\Lambda(\underline{\theta}_0)](\underline{\theta}-\underline{\theta}_0) , \tag{3.4.1}$$

where

$$\Lambda(\underline{\theta}) = \left[\frac{\partial^2 \ln L(\underline{\theta})}{\partial \theta_j \partial \theta_\ell} \right] ,$$

is an $r \times r$ matrix, and $\underline{\theta}_0$ is a fixed point.

In the method of scoring the matrix $\Lambda(\underline{\theta})$ is replaced by its expectation:

$$E(\Lambda(\underline{\theta})) = [\, E \frac{\partial^2 \ln L(\underline{\theta})}{\partial\theta_j \partial\theta_\ell} \,] = -I(\underline{\theta}) \,.$$

The method of scoring for finding $\hat{\underline{\theta}}_{MLE}$ is specified by the sequence of steps $\hat{\underline{\theta}}_0, \hat{\underline{\theta}}_1, \ldots,$ where $\hat{\underline{\theta}}_0$ is a starting value presumed to be near $\hat{\underline{\theta}}_{MLE}$, and $\hat{\underline{\theta}}_i$, $i = 1, 2, \ldots$ are defined by

$$\hat{\underline{\theta}}_{i+1} = \hat{\underline{\theta}}_i + I^{-1}(\hat{\underline{\theta}}_i) \frac{\partial \ln L(\hat{\underline{\theta}}_i)}{\partial \underline{\theta}} \,. \tag{3.4.2}$$

In practice only a few iterations are carried out and the final value of $\hat{\underline{\theta}}_{i+1}$ is taken as the value of $\hat{\underline{\theta}}_{MLE}$. Termination of iteration is often determined by criteria such as $\| \hat{\underline{\theta}}_{i+1} - \hat{\underline{\theta}}_i \| < \epsilon$ or $\| \partial \ln L(\hat{\underline{\theta}}_{i+1}) / \partial \underline{\theta} \| < \epsilon$ for some small $\epsilon > 0$.

Formula (3.4.2) specifies the essence of the method of scoring for any likelihood function assuming the required mathematical operations can be carried out. When $L(\underline{\theta})$ derives from a multinomial sampling distribution the method of scoring has a useful advantage over the Newton-Raphson procedure. The terms $\partial^2 \ln L(\underline{\theta}) / \partial\theta_j \partial\theta_\ell$ contain the second partial derivatives $\partial^2 \pi_i / \partial\theta_j \partial\theta_\ell$, $i = 0, 1, \ldots, t$. By taking expectations these higher order partial derivatives drop out and the elements of $I(\underline{\theta})$ involve only first partial derivatives of the cell probabilities.

A general advantage of the method of scoring is that when the iteration is terminated not only is $\hat{\underline{\theta}}_{MLE}$ determined but so is an estimate of its dispersion matrix as $I^{-1}(\hat{\underline{\theta}}_{MLE})$.

4. MAXIMUM LIKELIHOOD ESTIMATION OF POPULATION SIZE WHEN CAPTURE PROBABILITIES HAVE A BETA DISTRIBUTION

4.1 The Approach Used to Obtain the Solution

Throughout this chapter it is assumed that $F(p)$ is a Beta distribution. The results of Chapter 3 will be applied to ML estimation of population size given this assumption. Thus, $\underline{\theta} = (\alpha, \beta)'$ for $\alpha > 0$ and $\beta > 0$, and the cell probabilities $\pi_i(\alpha, \beta)$ are given by formula (2.4.1). Because of the complexity of these cell probabilities as functions of α and β, closed form solution for the ML estimators is not possible. It is noted that modelling capture probabilities as following a Beta distribution is not new (e.g., Holgate, 1966; Cormack, 1966; Eberhardt, 1969), however no one has previously done very much with this model.

Initially it was attempted to find the approximate ML estimators $\tilde{\alpha}$, $\tilde{\beta}$ and $\tilde{N} = S_t/(1-\pi_0(\tilde{\alpha}, \tilde{\beta})$ by using the method of scoring with the conditional likelihood function $L(\alpha, \beta | S_t) = P\{f_{1t}, \ldots, f_{tt} | S_t, \alpha, \beta\}$. The computer program necessary for this procedure was written by the author using formulae (2.4.2) through (2.4.7) for computing the cell probabilities and their partial derivatives. This program was then tried with some simulated livetrapping data from Burnham and Overton (1969), for which the true values of N, α and β were

known. It was discovered that satisfactory convergence to $\tilde{\alpha}$ and $\tilde{\beta}$ did not occur even with the true α and β as starting values.

By examining the likelihood surface more closely with these same data it was discovered that there was a virtual plateau in the vicinity of $\tilde{\alpha}$ and $\tilde{\beta}$. Thus the surface $L(\alpha, \beta | S_t)$ was in these cases, and presumably in general, ill conditioned. While it was apparent at this point that the ML estimators were not going to be very satisfactory, it still seemed worthwhile to find them.

It was decided to compute the exact ML estimators $\hat{N}_{MLE}$, $\hat{\alpha}_{MLE}$ and $\hat{\beta}_{MLE}$ based on expression (3.1.2). This approach was successful because for any fixed (integer) value of N, the ML estimators $\hat{\alpha}(N)$ and $\hat{\beta}(N)$ are easily found by the method of scoring. Before elaborating on this procedure it is worthwhile to briefly look at estimation of α and β when N is known.

The discrete probability distribution determined by $\pi_0(\alpha, \beta), \ldots, \pi_t(\alpha, \beta)$ has been known for many years and has a variety of names (Johnson and Kotz, 1969). It will be referred to here as the Beta-binomial distribution. Skellam (1948) gives the formulae for the method of moments estimators of α and β, these formulae are given in the Appendix. In the same paper Skellam also gives formulae for an iterative solution of the Maximum Likelihood equations, but he doubts if the labor involved in finding these more efficient estimators is worth expending.

Shenton (1950) has derived an approximate formula for the efficiency of the method of moments estimators for the Beta-binomial distribution. He concludes that the method of moments "rarely has low efficiency."

More recently, the Beta-binomial model is used in a paper by Chatfield and Goodhardt (1970). They consider the ML estimators "hard to find," and they use instead the method of mean and zeros. In this method f_{0t} and C_t are equated to their expectations and the resulting equations solved numerically for α and β. This method and its efficiency are examined in the Appendix.

Chatfield and Goodhardt use the method of mean and zeros because their data are strongly reverse J shaped (a larger value of f_{0t}/N), and they believe this method has higher efficiency than the method of moments when the underlying distribution has a large value of π_0. Because most livetrapping data are also strongly reverse J shaped their conjecture was of interest. As shown in the Appendix there is some truth to this conjecture; however, the difference in efficiencies between the two methods of estimation does not appear to be large except in certain extreme cases.

In 1948 Skellam may have been correct in saying the likelihood equations are not worth solving when the method of moments is reasonably efficient. Now that computers are readily accessible it is easy to find the exact ML estimators. Consequently, it is not

necessary to be concerned about the efficiency of possible alternative estimators, except when considering the choice of starting values in numerical solution of the likelihood equations. On the basis of Shenton's work and the Appendix it is concluded that the method of moments estimators are satisfactory as starting values in the method of scoring for $\hat{\alpha}(N)$ and $\hat{\beta}(N)$.

All of the formulae necessary to carry out the method of scoring with the Beta-binomial distribution have been given at various places in this thesis. They will be restated here for clarity. The cell probabilities can be computed for formulae (2.4.3) and (2.4.2):

$$\pi_0 = \prod_{i=1}^{t} \left(\frac{\beta+i-1}{\alpha+\beta+i-1}\right),$$

$$\pi_{i+1} = \pi_i \left[\frac{(t-i)(\alpha+i)}{(i+1)(\beta+t-i-1)}\right] \qquad i = 0, 1, \ldots, t-1.$$

The partial derivatives of the π_i can be computed from formulae (2.4.4) through (2.4.7):

$$\frac{\partial \pi_0}{\partial \alpha} = -\pi_0 \sum_{i=0}^{t} \frac{1}{(\alpha+\beta+i-1)},$$

$$\frac{\partial \pi_i}{\partial \alpha} = \frac{\pi_{i+1}}{\pi_i}\left(\frac{\partial \pi_i}{\partial \alpha}\right) + \frac{\pi_{i+1}}{\alpha+i} \qquad i = 0, 1, \ldots, t-1,$$

$$\frac{\partial \pi_0}{\partial \beta} = \alpha \pi_0 \sum_{i=1}^{t} \frac{1}{(\beta+i-1)(\alpha+\beta+i-1)} ,$$

$$\frac{\partial \pi_{i+1}}{\partial \beta} = \frac{\pi_{i+1}}{\pi_i} \left[\frac{\partial \pi_i}{\partial \beta} - \frac{\pi_i}{\beta+t-i-1} \right] \qquad i = 0, 1, \ldots, t-1.$$

The information matrix for the Beta-binomial distribution is given by

$$I(\alpha, \beta) = \begin{bmatrix} \sum_{i=0}^{t} \frac{1}{\pi_i} \left(\frac{\partial \pi_i}{\partial \alpha}\right)^2 & \sum_{i=0}^{t} \frac{1}{\pi_i} \left(\frac{\partial \pi_i}{\partial \alpha}\right)\left(\frac{\partial \pi_i}{\partial \beta}\right) \\ \sum_{i=0}^{t} \frac{1}{\pi_i} \left(\frac{\partial \pi_i}{\partial \alpha}\right)\left(\frac{\partial \pi_i}{\partial \beta}\right) & \sum_{i=0}^{t} \frac{1}{\pi_i} \left(\frac{\partial \pi_i}{\partial \beta}\right)^2 \end{bmatrix}.$$

With a random sample of size N from this distribution the likelihood equations are

$$\frac{\partial \ln L(\alpha, \beta)}{\partial \alpha} = \sum_{i=0}^{t} f_{it} \frac{1}{\pi_i} \left(\frac{\partial \pi_i}{\partial \alpha}\right) = 0,$$

$$\frac{\partial \ln L(\alpha, \beta)}{\partial \beta} = \sum_{i=0}^{t} f_{it} \frac{1}{\pi_i} \left(\frac{\partial \pi_i}{\partial \beta}\right) = 0.$$

The ML estimators $\hat{\alpha}(N)$ and $\hat{\beta}(N)$ can be found by the method of scoring. Let $\hat{\alpha}_0$ and $\hat{\beta}_0$ be initial values, then

$$\begin{bmatrix} \hat{\alpha}_{i+1} \\ \hat{\beta}_{i+1} \end{bmatrix} = \begin{bmatrix} \hat{\alpha}_i \\ \hat{\beta}_i \end{bmatrix} + I^{-1}(\hat{\alpha}_i, \hat{\beta}_i) \begin{bmatrix} \partial \ln L(\hat{\alpha}_i, \hat{\beta}_i)/\partial\alpha \\ \partial \ln L(\hat{\alpha}_i, \hat{\beta}_i)/\partial\beta \end{bmatrix},$$

$i = 0, 1, \ldots$ generates improved approximations to $\hat{\alpha}(N)$ and $\hat{\beta}(N)$. In the program written by the author the iterations are terminated when $|\partial \ln L(\alpha, \beta)/\partial\alpha| < .0001$ and $|\partial \ln L(\alpha, \beta)/\partial\beta| < .0001$.

The algorithm for the exact ML estimators of N, α and β is as follows. Choose an integer N equal to or slightly larger than S_t. Let $f_{0t} = N - S_t$. Then $f_{0t}, f_{1t}, \ldots, f_{tt}$ are considered to be the frequencies of observations from a random sample of size N from the Beta-binomial distribution. Find the ML estimates $\hat{\alpha}(N)$ and $\hat{\beta}(N)$ using the method of scoring. Then compute $\phi(N) = \ln L(N, \hat{\alpha}(N), \hat{\beta}(N)) + K$, where K is a constant:

$$(4.1.1) \qquad \phi(N) = \sum_{i=0}^{S_t - 1} \ln(N-i) + \sum_{i=0}^{t} f_{it} \ln \pi_i(\hat{\alpha}(N), \hat{\beta}(N)).$$

Continue in this manner to compute $\phi(N+1), \phi(N+2), \ldots,$ until $\hat{N}_{MLE}$ is found such that

$$\phi(\hat{N}_{MLE}) = \max_{N \in \{S_t, S_t+1, \ldots\}} \phi(N).$$

Finally

$$\hat{\alpha}_{MLE} = \hat{\alpha}(\hat{N}_{MLE}) \quad \text{and} \quad \hat{\beta}_{MLE} = \hat{\beta}(\hat{N}_{MLE}).$$

This algorithm is not necessarily the most efficient one, especially if a large range of values of N must be searched for the maximum of $\phi(N)$. However it does work quite well because of the following feature. In the application of this algorithm starting values $\hat{\alpha}_0$ and $\hat{\beta}_0$ are supplied only for the initial value of N. Thereafter the values $\hat{\alpha}(N)$ and $\hat{\beta}(N)$ are excellent starting values to use in the method fo scoring for finding $\hat{\alpha}(N+1)$ and $\hat{\beta}(N+1)$. Because of this feature it is not time consuming to find the exact ML estimators.

The dispersion matrix of these estimators is taken as $D(\hat{N}, \hat{\alpha}, \hat{\beta})$ given by (3.3.8), and an estimate of this matrix is obtained by evaluating it at the estimated values of N, α and β.

It will be shown in the next section that the ML estimator of N is not very satisfactory. In fact, there is almost an identifiability problem for values of N, α and β which seem likely to hold in livetrapping studies. Consequently, brief consideration will be given to the question of identifiability with the class of probability distributions $P\{f_{1t}, \ldots, f_{tt} | N, \alpha, \beta\}$ for $t \geq 3$.

Identifiability is usually discussed in the context of mixtures of distributions (e.g., Blischke, 1963) where it refers to the unique characterization of a given mixture. For a parametric class of probability distributions identifiability should mean a one-to-one relationship between the parameter space and the class of distributions.

Let $P_j = P\{f_{1t}, \ldots, f_{tt} | N_j, \alpha_j, \beta_j\}$, for $j = 1, 2$. It can be shown that if $P_1\{A\} = P_2\{A\}$ for all borel sets A in t-dimensional Euclidean space, then $N_1 = N_2$, $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$. This does not seem to be a strong enough result here; in fact, it is derivable from the following result.

Let $E_j(f_{it}) = N_j \pi_i(\alpha_j, \beta_j)$, $j = 1, 2$ and $i = 1, \ldots, t$. If $E_1(f_{it}) = E_2(f_{it})$, $i = 1, \ldots, t$, then $N_1 = N_2$, $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$.

Proof: It is assumed that

$$N_1 \pi_i(\alpha_1, \beta_1) = N_2 \pi_i(\alpha_2, \beta_2)$$

which implies

$$\frac{\pi_i(\alpha_1, \beta_1)}{\pi_i(\alpha_2, \beta_2)} = \frac{N_2}{N_1} = c > 0 \qquad i = 1, \ldots, t.$$

Let $(i)^{[h]} = i(i-1)\ldots(i-h+1)$ and consider

$$\sum_{i=1}^{t} (i)^{[h]} \pi_i(\alpha_1, \beta_1) = c \sum_{i=1}^{t} (i)^{[h]} \pi_i(\alpha_2, \beta_2) ,$$

for $h = 1, 2, 3$. This leads to the three equations,

$$\frac{\alpha_1}{\alpha_1 + \beta_1} = c \frac{\alpha_2}{\alpha_2 + \beta_2} ,$$

$$\frac{(\alpha_1+1)\alpha_1}{(\alpha_1+\beta_1+1)(\alpha_1+\beta_1)} = c\,\frac{(\alpha_2+1)\alpha_2}{(\alpha_2+\beta_2+1)(\alpha_2+\beta_2)}$$

and

$$\frac{(\alpha_1+2)(\alpha_1+1)\alpha_1}{(\alpha_1+\beta_1+2)(\alpha_1+\beta_1+1)(\alpha_1+\beta_1)} = c\,\frac{(\alpha_2+2)(\alpha_2+1)\alpha_2}{(\alpha_2+\beta_2+2)(\alpha_2+\beta_2+1)(\alpha_2+\beta_2)} \quad .$$

Eliminating the unknown constant c leads to the two equations

$$\frac{\alpha_1+j}{\alpha_1+\beta_1+j} = \frac{\alpha_2+j}{\alpha_2+\beta_2+j} \qquad j = 1, 2.$$

It follows that

$$\alpha_1\beta_2 - \beta_1\alpha_2 = j(\beta_1-\beta_2) \qquad j = 1, 2$$

and thus $\beta_1 = \beta_2$. As $\beta_1 > 0$ is required it follows that $\alpha_1 = \alpha_2$ and finally $\pi_i(\alpha_1, \beta_1) = \pi_i(\alpha_2, \beta_2)$, $i = 1, \ldots, t$, so $N_1 = N_2$.

4.2 The Unsatisfactory Nature of Maximum Likelihood Estimation in this Problem

Examples of ML estimation of N, α and β are given in Tables 1 through 5. The estimated standard deviations are obtained by evaluating $\sqrt{V(\tilde{N})}$ at the estimated parameter values, where $V(\tilde{N})$ given by (3.3.9) approximates the unconditional variance of $\hat{N}_{MLE}$. The estimated expected capture frequencies are computed as $\hat{E}(f_{it}) = \hat{N}_{MLE}\pi_i(\hat{\alpha}_{MLE}, \hat{\beta}_{MLE})$, $i = 0, 1, \ldots, t$.

Edwards and Eberhardt (1967) have looked at several estimators

of population size for two capture-recapture studies. In one of these studies, 135 wild cottontails were placed in a 40 acre enclosure and livetrapping was conducted on this penned population for 18 days. These data are given in Table 1. It is seen from Table 1 that $\hat{N}_{MLE} = 253$, with an estimated standard deviation (SD) of 243.72. Examination of the sequence of points $\phi(N)$ for $100 \leq N \leq 300$ shows the extreme flatness of the likelihood function over this wide range of values of N. Thus while $\widehat{SD}(\hat{N}_{MLE}) = 243.72$ may be an overestimate, it is clear that $SD(\hat{N}_{MLE})$ is large.

Using $N = 135$ and fitting the capture frequencies with the Beta-binomial model gives ML estimates $\hat{\alpha} = 1.2557$ and $\hat{\beta} = 20.20469$, with an estimated dispersion matrix

$$\widehat{D}(\hat{\alpha}, \hat{\beta}) = \begin{bmatrix} .1606 & 2.5572 \\ 2.5572 & 46.0643 \end{bmatrix},$$

and the estimated expected capture frequencies shown below:

i	$\hat{E}(f_{it})$	f_{it}
0	60.9	59
1	37.0	43
2	19.6	16
3	9.7	8
4	4.5	6
5	2.0	0
6	.8	2
7	.3	1
8	.1	0
9	.0	0
⋮	⋮	⋮
18	.0	0

Comparing f_{it} with $\hat{E}(f_{it})$, the Chi-square goodness of fit statistic value is 2.5091 at 3 degrees of freedom. Using these improved parameter estimates gives $\widehat{SD}(\hat{N}_{MLE}) = 44.66$.

There is no basis for saying the Beta-binomial model of livetrapping does not fit these data, and yet the estimate $\hat{N}_{MLE} = 253$ is quite poor. The estimate of its SD does not appear to be very good either.

Table 2 gives the results of another livetrapping study reported on by Edwards and Eberhardt in which N is not known. Tables 3 and 4 examine data from two livetrapping studies (N not known) reported by Nixon, Edwards and Eberhardt (1967). Finally, in Table 5 simulated data are used for which $N = 100$, $\alpha = 1$ and $\beta = 15.6667$. From Table 5 it is seen that $\widehat{SD}(\hat{N}_{MLE}) = 97.32$, while theoretically $SD(\hat{N}_{MLE}) = 52.43$. It is possible that $\widehat{SD}(\hat{N}_{MLE})$ has a positive bias; however, it is clear that the $SD(\hat{N}_{MLE})$ is so large in this case that $\hat{N}_{MLE}$ is unreliable.

Tables 2, 3 and 4 also indicate a large $SD(\hat{N}_{MLE})$, especially when it is remembered that $N \geq S_t$. The extreme flatness of $\phi(N)$ around $\hat{N}_{MLE}$ corroborates these large estimated standard deviations. The basic problem seems to be that it is possible to find parameter points (N_1, α_1, β_1) and (N_2, α_2, β_2) which are quite different and yet the differences $|E_1(f_{it}) - E_2(f_{it})|$, $i = 1, \ldots, t$ are small. Examples of this are given in Tables 6 and 7.

In Table 6 the data of Edwards and Eberhardt with known $N = 135$ are used. The expected frequencies are given for N, $\hat{\alpha}(N)$ and $\hat{\beta}(N)$, for $N = 220$, 253 and 300. There is very little difference between these expected frequencies.

Table 7 uses the simulated data repoted in Table 5. The ML estimate of N for these data is 133 (the true value of N is 100), and as stated in Table 7 the minimum Chi-square estimate of N is 236.

The unsatisfactory nature of $\hat{N}_{MLE}$ is further indicated in Table 8. For $N = 100$ this table gives approximate standard deviations of $\hat{N}_{MLE}$ for a variety of Beta distributions. Also shown in Table 8 is $SD(\hat{N}_{MLE} | \alpha, \beta)$, which is just the standard deviation of $S_t/(1-\pi_0)$ when π_0 is known, and an estimate of $SD(\hat{N}_J)$ where $\hat{N}_J$ is the jackknife estimator discussed in Chapter 6. As explained in Chapter 6 $SD(\hat{N}_J)$ is based on 20 simulated livetrapping studies at each point t, α and β. Finally, $E(S_t)$ is also shown as a guide to interpreting the other table entries.

Define the percent standard deviation of $\hat{N}_{MLE}$ as

$$PSD(\hat{N}_{MLE}) = 100 \frac{SD(\hat{N}_{MLE})}{N} = \frac{10}{\sqrt{N}} SD(\hat{N}_{MLE} | N = 100).$$

The entries for $SD(\hat{N}_{MLE})$ in Table 8 are automatically in percent standard deviation. The $PSD(\hat{N}_{MLE})$ is a much better measure of

the reliability of $\hat{N}_{MLE}$ than is $SD(\hat{N}_{MLE})$ because as N increases, $PSD(\hat{N}_{MLE})$ decreases, even though $SD(\hat{N}_{MLE})$ increases.

From Table 8 it is seen that $\hat{N}_{MLE}$ has a large PSD for values of t, α and β which might reasonably hold for a real live-trapping study, e.g., t ranging from 5 to 15 and $E(p) \in [.12, .06]$. Consider for example $t = 15$, $\alpha = 1$ and $\beta = 7.333$ $(E(p) = .12)$. Then $SD(\hat{N}_{MLE}) = 22.4$ when $N = 100$. In addition consider that $E(S_{15}) = 67.2$ and $\hat{N}_{MLE}$ appears to be quite unsatisfactory. Yet if t, α and β remained the same and N is increased to 10,000, the $PSD(\hat{N}_{MLE}) = 2.24$ --a more acceptable value. In this case $E(S_{15}) = 6,720$, and $SD(\hat{N}_{MLE}) = 224$. Now $\hat{N}_{MLE}$ appears to be satisfactory.

It is concluded that $\hat{N}_{MLE}$ becomes more reliable (satisfactory) with larger N but for the range of N, t, α and β likely to occur in real livetrapping studies it is expected that $\hat{N}_{MLE}$ will often be unsatisfactory, and sometimes completely unreliable.

Because it is known that $\hat{N}_{MLE} \geq S_t$, it is anticipated that the distribution of $\hat{N}_{MLE}$ is skewed to the right. Table 8 indicates that this skewness may be substantial for small to moderate values of N. Thus it is to be expected that $\hat{N}_{MLE}$ has a positive bias in addition to a large PSD in these cases.

It should be noted that the usual Chi-square goodness-of-fit test

is virtually powerless, thus worthless, in this problem. This Beta model for livetrapping appears flexible enough to fit any livetrapping data. Thus it is pointless to use the ML estimates to test the fit of the model to the data.

The model for capture-recapture studies given in Section 2.1 was conceived in an attempt to derive estimators that would be robust to departures from the usual model wherein capture probabilities are assumed constant, i.e., $p_j = p$, $j = 1, \ldots, N$. It was felt that by letting $F(p)$ be an arbitrary Beta distribution an estimator of N could be derived that would have good performance for any set of capture probabilities $p_1, \ldots, p_N$. A robust estimator would thus be obtained, at least with respect to variation in capture probabilities among animals. However, it has been shown that $\hat{N}_{MLE}$ may not be a satisfactory estimator even when the model holds true if N is only moderately large and α and β are unconstrained.

For a fixed value of $E(p)$, it was never anticipated that all values of α and β were equally likely a-priori. It may be that certain types of Beta distributions will adequately model livetrapping experiments. As discussed in the next section there are theoretical and empirical reasons for considering the restricted approach wherein $\alpha = 1$ is assumed.

Table 1. Maximum Likelihood estimation of N, α and β from the livetrapping data of Edwards and Eberhardt (1967) obtained from a confined population of 135 wild cottontails.

The livetrapping data $t = 18$ $S_{18} = 76$ $C_{18} = 142$

i	1	2	3	4	5	6	7	8	...	18
$f_{i\,18}$	43	16	8	6	0	2	1	0	...	0

Parameter	ML Estimate	Estimate of Standard Deviation
N	253.	243.72
α	.37518	.56309
β	11.66072	6.399105

Estimated expected capture frequencies

i	1	2	3	4	5	6	7	8	9	10	...	18
$\hat{E}(f_{i\,18})$	41.7	17.6	8.4	4.1	2.1	1.0	.5	.2	.1	.0	...	.0

Some values of $\hat{\alpha}(N)$, $\hat{\beta}(N)$ and $\phi(N)$

N	$\hat{\alpha}(N)$	$\hat{\beta}(N)$	$\phi(N)$
100	3.41061	39.78111	152.33436
150	.97490	17.54049	155.15068
200	.55144	13.42271	155.48756
225	.45166	12.42944	155.52171
252	.37747	11.68382	155.53040
253	.37518	11.66072	155.53041
254	.37292	11.63790	155.53039
275	.33099	11.21332	155.52698
300	.29183	10.81485	155.51794

Table 2. Maximum Likelihood estimation of N, α and β from the livetrapping data of Edwards and Eberhardt (1967) obtained from a wild population of cottontails of unknown size.

The livetrapping data $t = 8$ $S_8 = 69$ $C_8 = 134$

i =	1	2	3	4	5	6	7	8
$f_{i\,8}$ =	36	15	13	3	1	1	0	0

Parameter	ML Estimate	Estimate of Standard Deviation
N	121.	42.30
α	1.23227	1.28488
β	8.08683	5.48583

Estimated expected capture frequencies

i =	1	2	3	4	5	6	7	8
$\hat{E}(f_{i\,8})$ =	34.2	18.9	9.4	4.1	1.5	.5	.1	.0

Some values of $\hat{\alpha}(N)$, $\hat{\beta}(N)$ and $\phi(N)$

N	$\hat{\alpha}(N)$	$\hat{\beta}(N)$	$\phi(N)$
80	11.19518	44.76740	133.56115
100	2.35204	12.33645	135.71920
120	1.26191	8.31701	135.95412
121	1.23227	8.08683	135.95439
122	1.20431	7.97913	135. 95412
150	.73173	6.14048	135.83570
175	.53943	5.37985	135.68798
200	.42647	4.92865	135.55522

Table 3. Maximum Likelihood estimation of N, α and β from the livetrapping data of Nixon, Edwards and Eberhardt (1967) obtained from a wild population of squirrels in 1962.

The livetrapping data $t = 11$ $S_{11} = 68$ $C_{11} = 139$

i =	1	2	3	4	5	6	7	...	11
$f_{i\,11}$ =	33	16	10	4	2	3	0	...	0

Parameter	ML Estimate	Estimate of Standard Deviation
N	131.	55.00
α	.79752	.77497
β	7.47337	4.09466

Estimated expected capture frequencies

i =	1	2	3	4	5	6	7	8	9	10	11
$\hat{E}(f_{i\,11})$ =	31.8	17.3	9.4	4.9	2.5	1.1	.5	.2	.1	.0	.0

Some values of $\hat{\alpha}(N)$, $\hat{\beta}(N)$ and $\phi(N)$

N	$\hat{\alpha}(N)$	$\hat{\beta}(N)$	$\phi(N)$
80	4.20876	22.40484	120.66209
105	1.39807	10.20478	123.12923
130	.81117	7.53650	123.35698
131	.79752	7.47337	123.35706
132	.78431	7.41221	123.35681
160	.53388	6.24154	123.27730
180	.43378	5.76663	123.19314
200	.36498	5.43752	123.11135

Table 4. Maximum Likelihood estimation of N, α and β from the livetrapping data of Nixon, Edwards and Eberhardt (1967) obtained from a wild population of squirrels in 1963.

The livetrapping data $t = 11$ $S_{11} = 72$ $C_{11} = 223$

i =	1	2	3	4	5	6	7	8	9	10	11
$f_{i\,11}$ =	23	14	9	6	8	7	3	0	2	0	0

Parameter	ML Estimate	Estimate of Standard Deviation
N	104.	22.12
α	.77123	.44692
β	3.20450	1.12960

Estimated expected capture frequencies

i =	1	2	3	4	5	6	7	8	9	10	11
$\hat{E}(f_{i\,11})$ =	20.8	15.1	11.2	8.3	6.0	4.2	2.8	1.8	1.0	.5	.1

Some values of $\hat{\alpha}(N)$, $\hat{\beta}(N)$ and $\phi(N)$

N	$\hat{\alpha}(N)$	$\hat{\beta}(N)$	$\phi(N)$
90	1.15530	3.97044	98.20172
103	.79031	3.24302	98.54869
104	.77123	3.20450	98.54874
105	.75302	3.16768	98.54682
125	.50900	2.66869	98.30503
150	.36038	2.35864	97.89038
200	.22658	2.07439	97.31010

Table 5. Maximum Likelihood estimation of N, α and β from simulated livetrapping data. The data were generated with N = 100, α = 1 and β = 15.667.

The simulated livetrapping data $t = 15$ $S_{15} = 53$ $C_{15} = 89$

i	=	1	2	3	4	5	6	...	15
$f_{i\,15}$	=	32	12	5	2	2	0	...	0

Parameter	ML Estimate	Estimate of Standard Deviation
N	133.	97.32
α	.80440	1.25472
β	17.22885	14.49051

Estimated expected capture frequencies

i	=	1	2	3	4	5	6	7	8	...	15
$\hat{E}(f_{i\,15})$	=	31.0	12.9	5.4	2.2	.8	.3	.1	.0	...	.0

Some values of $\hat{\alpha}(N)$, $\hat{\beta}(N)$ and $\phi(N)$

N	$\hat{\alpha}(N)$	$\hat{\beta}(N)$	$\phi(N)$
65	30.60280	304.64662	95.09474
100	1.61044	25.52182	97.55716
132	.81702	17.36122	97.68589
133	.80440	17.22885	97.68590
134	.79215	17.10031	97.68581
150	.63627	15.45574	97.67426
175	.48559	13.84846	97.63926
200	.39213	12.84131	97.60113
250	.28274	11.65122	97.53466

Table 6. Comparison of observed and expected frequencies for selected values of N, $\hat{\alpha}(N)$ and $\hat{\beta}(N)$ for the data of Edwards and Eberhardt when N is known to be 135.

		Expected Frequencies		
		N = 220	253	300
Observed		$\hat{\alpha}(N)$ = .46867	.37518	.29183
Frequencies		$\hat{\beta}(N)$ = 12.59947	11.66072	10.81485
i	$f_{i\,18}$	$E(f_{i\,18})$	$E(f_{i\,18})$	$E(f_{i\,18})$
1	43	41.13	41.74	42.31
2	16	17.95	17.64	17.33
3	8	8.56	8.38	8.20
4	6	4.19	4.13	4.08
5	0	2.05	2.05	2.06
6	2	.99	1.01	1.03
7	1	.46	.49	.51
8	0	.21	.23	.25
9	0	.09	.10	.11
10	0	.04	.04	.05
11	0	.01	.02	.02
12	0	.01	.01	.01
13	0	.00	.00	.00
⋮	⋮	⋮	⋮	⋮
18	0	.00	.00	.00

Table 7. Comparison of observed and expected frequencies for selected values of N, $\hat{\alpha}(N)$ and $\hat{\beta}(N)$ for the simulated data with N = 100, $\alpha = 1.0$ and $\beta = 15.6667$. N = 236 is the minimum Chi-square estimate of N.

		Expected Frequencies			
		N = 100	133	180	236
Observed		$\hat{\alpha}(N)$ = 1.61044	.80440	.46354	.30294
Frequencies		$\hat{\beta}(N)$ = 25.52184	17.22885	13.61158	11.74649
i	$f_{i\,15}$	$E(f_{i\,15})$	$E(f_{i\,15})$	$E(f_{i\,15})$	$E(f_{i\,15})$
1	32	29.23	30.99	31.97	32.30
2	12	13.87	12.95	12.31	11.90
3	5	5.78	5.38	5.13	5.00
4	2	2.19	2.18	2.17	2.18
5	2	.76	.84	.90	.95
6	0	.24	.31	.36	.40
7	0	.07	.11	.14	.17
8	0	.02	.03	.05	.06
9	0	.00	.01	.02	.02
10	0	.00	.00	.01	.01
11	0	.00	.00	.00	.00
⋮	⋮	⋮	⋮	⋮	⋮
15	0	.00	.00	.00	.00

Table 8. Approximate standard deviations (SD) of $\hat{N}_{MLE}$ when N = 100. Also shown is the estimated standard deviation of the jackknife estimator, $\hat{N}_J$; these estimates were obtained by simulation with N = 100.

t	$E(S_t)$	$SD(\hat{N}_{MLE})$	$SD(\hat{N}_{MLE} \mid \alpha, \beta)$	$SD(\hat{N}_J)$
$\alpha = .3158$	$\beta = 1.0$	$E(p) = .24$		
5	48.2	80.9	10.4	6.7
10	57.6	43.6	8.6	6.3
15	62.5	33.1	7.8	6.3
20	65.6	27.9	7.2	6.4
25	67.9	24.7	6.9	7.6
30	69.6	22.5	6.6	5.2
$\alpha = 1.0$	$\beta = 3.1667$	$E(p) = .24$		
5	61.2	35.9	8.0	7.6
10	76.0	15.1	5.6	6.1
15	82.6	10.0	4.6	5.6
20	86.3	7.7	4.0	5.9
25	88.8	6.4	3.6	5.9
30	90.5	5.5	3.3	4.5
$\alpha = 6.0$	$\beta = 19.0$	$E(p) = .24$		
5	71.7	19.8	6.3	11.2
10	90.0	5.6	3.3	6.7
15	95.9	2.7	2.1	5.7
20	98.1	1.7	1.4	3.5
25	99.0	1.1	1.0	2.8
30	99.5	.8	.7	1.9
$\alpha = .250$	$\beta = 1.0$	$E(p) = .20$		
5	41.2	117.9	12.0	5.7
10	49.8	64.2	10.0	6.3
15	54.4	49.2	9.2	5.7
20	57.5	41.6	8.6	5.6
25	57.7	36.9	8.2	6.0
30	61.5	33.7	7.9	7.2

Table 8. Continued.

t	$E(S_t)$	$SD(\hat{N}_{MLE})$	$SD(\hat{N}_{MLE} \mid \alpha, \beta)$	$SD(\hat{N}_J)$
$\alpha = 1.0$	$\beta = 4.0$	$E(p) = .20$		
5	55.6	46.7	8.9	8.7
10	71.4	19.0	6.3	8.3
15	79.0	12.4	5.2	7.8
20	83.3	9.4	4.5	5.6
25	86.2	7.7	4.0	5.4
30	88.2	6.6	3.6	4.7
$\alpha = 4.0$	$\beta = 16.0$	$E(p) = .20$		
5	63.5	29.8	7.6	12.2
10	83.7	8.9	4.4	7.1
15	91.6	4.7	3.0	5.8
20	95.3	3.0	2.2	4.6
25	97.1	2.1	1.7	3.7
30	98.2	1.6	1.4	2.8
$\alpha = .1905$	$\beta = 1.0$	$E(p) = .16$		
5	33.7	181.7	14.0	6.6
10	41.3	100.1	12.0	7.0
15	45.4	77.1	11.0	5.1
20	48.3	65.5	10.4	5.6
25	50.4	58.4	10.0	5.7
30	52.0	53.4	9.6	7.1
$\alpha = 1.0$	$\beta = 5.250$	$E(p) = .16$		
5	48.8	64.5	10.3	10.2
10	65.6	25.1	7.3	8.1
15	74.1	16.0	5.9	9.2
20	79.2	12.0	5.1	7.6
25	82.6	9.7	4.6	5.6
30	85.1	8.3	4.2	5.5
$\alpha = 4.0$	$\beta = 21.0$	$E(p) = .16$		
5	55.3	44.0	9.0	14.1
10	77.1	13.0	5.5	8.7
15	87.1	6.0	3.9	7.2
20	92.2	4.4	2.9	5.6
25	95.0	3.1	2.3	4.8
30	96.6	2.4	1.9	4.2

Table 8. Continued.

t	$E(S_t)$	$SD(\hat{N}_{MLE})$	$SD(\hat{N}_{MLE} \mid \alpha, \beta)$	$SD(\hat{N}_J)$
$\alpha = .1364$	$\beta = 1.0$	$E(p) = .12$		
5	25.8	307.5	17.0	5.3
10	32.0	171.0	14.6	5.5
15	35.5	132.4	13.5	4.7
20	37.9	113.1	12.8	4.9
25	39.7	101.1	12.3	5.0
30	41.2	92.8	12.0	5.0
$\alpha = 1.0$	$\beta = 7.3333$	$E(p) = .12$		
5	40.5	97.9	12.1	10.5
10	57.7	36.2	8.6	8.8
15	67.2	22.4	7.0	7.8
20	73.2	16.5	6.1	6.1
25	77.3	13.2	5.4	6.9
30	80.4	11.1	4.9	6.1
$\alpha = 3.0$	$\beta = 22.0$	$E(p) = .12$		
5	44.6	74.6	11.1	13.0
10	66.2	22.8	7.2	11.1
15	77.9	12.3	5.3	9.0
20	84.7	8.1	4.3	7.0
25	89.0	5.9	3.5	5.9
30	91.8	4.6	3.0	5.4
$\alpha = .0989$	$\beta = 1.0$	$E(p) = .09$		
5	19.7	506.9	20.2	5.1
10	24.6	283.9	17.5	5.1
15	27.4	220.8	16.3	4.1
20	29.4	189.1	15.5	4.1
25	30.9	169.4	15.0	5.7
30	32.1	155.7	14.5	4.6
$\alpha = 1.0$	$\beta = 10.0000$	$E(p) = .09$		
5	33.1	149.0	14.2	11.0
10	49.7	52.6	10.1	11.3
15	59.7	31.7	8.2	8.9
20	66.4	22.8	7.1	8.6
25	71.2	18.0	6.4	7.5
30	74.8	15.0	5.8	8.0

Table 8. Continued.

t	$E(S_t)$	$SD(\hat{N}_{MLE})$	$SD(\hat{N}_{MLE} \mid \alpha, \beta)$	$SD(\hat{N}_J)$
$\alpha = 9.0$	$\beta = 91.0$	$E(p) = .09$		
5	37.0	106.4	13.1	13.3
10	59.4	29.9	8.3	14.6
15	73.3	14.9	6.0	10.3
20	82.1	9.2	4.1	8.4
25	87.8	6.2	3.7	7.2
30	91.6	4.5	3.0	6.7
$\alpha = 1.0$	$\beta = 15.6667$	$E(p) = .06$		
5	24.2	270.4	17.7	9.8
10	39.0	90.3	12.5	11.6
15	48.9	52.4	10.2	10.0
20	56.1	36.8	8.9	9.6
25	61.5	28.4	7.9	9.9
30	65.7	23.3	7.2	7.9
$\alpha = 3.0$	$\beta = 47.0$	$E(p) = .06$		
5	25.7	225.4	17.0	10.9
10	43.3	67.0	11.4	14.4
15	55.8	35.3	8.9	12.2
20	64.8	22.7	7.4	12.4
25	71.6	16.6	6.3	9.7
30	76.7	12.6	5.5	10.5
$\alpha = 1.0$	$\beta = 24.0$	$E(p) = .04$		
5	17.2	492.5	21.9	8.4
10	29.4	157.7	15.5	11.5
15	38.5	88.7	12.7	12.1
20	45.5	60.8	11.0	13.3
25	51.0	46.1	9.8	11.4
30	55.6	37.1	8.9	10.5
$\alpha = 2.0$	$\beta = 48.0$	$E(p) = .04$		
5	17.8	443.6	21.5	9.1
10	31.3	133.6	14.8	12.9
15	41.7	71.3	11.8	12.4
20	49.9	46.7	10.0	11.5
25	55.5	34.0	8.8	10.6
30	61.8	26.4	7.9	10.4

4.3 The Special Case $\alpha = 1$

For the special case $\alpha = 1$, the ML estimators of N and β are given by the obvious modification of the algorithm used in the general case. For a fixed value of N, $\hat{\beta}(N)$ is the solution of

$$\frac{d \ln L(1, \beta)}{d\beta} = \sum_{i=0}^{t} f_{it} \frac{1}{\pi_i} \frac{d\pi_i}{d\beta} = 0.$$

The method of scoring for finding $\hat{\beta}(N)$ is given by

$$\hat{\beta}_{j+1}(N) = \hat{\beta}_j(N) + \frac{1}{I(\hat{\beta}_j)} \left(\frac{d \ln L(1, \hat{\beta}_j)}{d\beta}\right) \qquad j = 0, 1, \ldots,$$

where

$$I(\beta) = \sum_{i=0}^{t} \frac{1}{\pi_i} \left(\frac{d\pi_i}{d\beta}\right)^2 .$$

From (2.4.3) π_0 is given as

$$\pi_0(\alpha, \beta) = \prod_{i=1}^{t} \left(\frac{\beta+i-1}{\alpha+\beta+i-1}\right) .$$

This expression for π_0 will simplify if α is an integer in the range $1 \leq \alpha \leq t-1$. In particular, for $\alpha = 1$

$$\pi_0 = \frac{\beta}{t+\beta} .$$

Consequently for $\alpha = 1$

$$\pi_{i+1} = \pi_i\left(\frac{t-i}{\beta+t-i-1}\right) \qquad i = 0, 1, \ldots, t-1,$$

$$\frac{d\pi_0}{d\beta} = \frac{t}{(t+\beta)^2},$$

and

$$\frac{d\pi_{i+1}}{d\beta} = \frac{\pi_{i+1}}{\pi_i}\left[\frac{d\pi_i}{d\beta} - \frac{\pi_i}{\beta+t-i-1}\right] \qquad i = 0, 1, \ldots, t-1.$$

Finally, the estimators $\hat{N}_{MLE}$ and $\hat{\beta}_{MLE}$ are defined by

$$L(\hat{N}_{MLE}, 1, \hat{\beta}_{MLE}) = \max_{N \in \{S_t, S_t+1, \ldots\}} L(N, 1, \hat{\beta}(N)).$$

It should rarely, if ever, be necessary to compute these exact ML estimators because with $\alpha = 1$ there exists highly efficient closed form estimators of N and β. Because π_0 simplifies when $\alpha = 1$ this gives

$$E(S_t) = N\frac{t}{t+\beta}.$$

Also

$$E(C_t) = N\frac{t}{1+\beta}.$$

By setting S_t and C_t equal to their expectations, estimators of N and β are found to be

$$\hat{N} = \frac{S_t}{1 - S_t/C_t}\left(1 - \frac{1}{t}\right) \tag{4.3.1}$$

$$\hat{\beta} = \frac{tS_t - C_t}{C_t - S_t}, \tag{4.3.2}$$

provided $S_t \neq C_t$. From $S_t = \sum_{i=1}^{t} f_{it}$ and $C_t = \sum_{i=1}^{t} if_{it}$, it follows that $C_t = S_t$ iff $f_{1t} = S_t$, that is, only if there are no recaptures. If there are any recaptures then $0 \leq \hat{\beta} < +\infty$ and $S_t \leq \hat{N} < +\infty$.

The dispersion matrix $D(\hat{N}, \hat{\beta})$ is approximated by $\Lambda D(S_t, C_t)\Lambda'$ where $D(S_t, C_t)$ is the dispersion matrix of S_t and C_t, and Λ is the Jacobian matrix

$$\begin{bmatrix} \frac{\partial \hat{N}}{\partial S_t} & \frac{\partial \hat{N}}{\partial C_t} \\ \frac{\partial \hat{\beta}}{\partial S_t} & \frac{\partial \hat{\beta}}{\partial C_t} \end{bmatrix}$$

evaluated at $S_t = E(S_t)$ and $C_t = E(C_t)$:

$$\Lambda = \begin{bmatrix} \frac{(t+\beta)^2}{t(t-1)} & -\frac{(1+\beta)^2}{t(t-1)} \\ \frac{(t+\beta)^2(1+\beta)}{Nt(t-1)} & -\frac{(1+\beta)^2(t+\beta)}{Nt(t-1)} \end{bmatrix}.$$

From Section 6.1, if $L_1 = \sum_{i=1}^{t} a_i f_{it}$ and $L_2 = \sum_{i=1}^{t} b_i f_{it}$ are

linear combinations of the capture frequencies, then

$$\mathrm{Cov}(L_1, L_2) = N\left[\sum_{i=1}^{t} a_i b_i \pi_i - \frac{E(L_1)}{N}\frac{E(L_2)}{N}\right].$$

Using this formula $V(S_t)$, $V(C_t)$ and $\mathrm{Cov}(S_t, C_t)$ are easily found in general. For $\alpha = 1$

$$D(S_t, C_t) = \begin{bmatrix} Nt\dfrac{\beta}{(t+\beta)^2} & Nt\dfrac{\beta}{(1+\beta)(t+\beta)} \\ Nt\dfrac{\beta}{(1+\beta)(t+\beta)} & Nt\dfrac{\beta(1+\beta+t)}{(1+\beta)^2(2+\beta)} \end{bmatrix}.$$

It follows that

$$(4.3.3) \qquad D(\hat{N}, \hat{\beta}) = \begin{bmatrix} \dfrac{N\beta}{t(t-1)}\left(t+1 - \dfrac{1+\beta}{2+\beta}\right) & \dfrac{(t+\beta)(1+\beta)^2\beta}{t(t-1)(2+\beta)} \\ \dfrac{(t+\beta)(1+\beta)^2\beta}{t(t-1)(2+\beta)} & \dfrac{(t+\beta)^2(1+\beta)^2\beta}{Nt(t-1)(2+\beta)} \end{bmatrix}.$$

By approximating these estimators by the first three terms of their Taylor's series expansion it follows that

$$E(\hat{N}) \doteq N + \frac{\beta(1+\beta)(t+\beta)}{t(t-1)(2+\beta)},$$

$$E(\hat{\beta}) \doteq \beta - \frac{\beta(1+\beta)(t+\beta)}{Nt(t-1)}\left[\frac{t+\beta}{2+\beta} - \frac{\beta+1}{t-1}\right].$$

Usually with the multinomial distribution N is the sample size and it is assumed known. Then the asymptotic distribution of any

linear combinations L_1 and L_2 of the frequencies is bivariate normal. In the present situation N is not known but it is still true that if N is large, then S_t and C_t will be approximately bivariate normal random variables. Because the estimators $\hat{N}$ and $\hat{\beta}$ are totally differentiable functions of S_t and C_t (except for $S_t = C_t$, a set which has zero asymptotic probability, it follows (Rao, 1965) that $\hat{N}$ and $\hat{\beta}$ are for large N approximately bivariate normal random variables with mean vector $(N, \beta)'$ and dispersion matrix $D(\hat{N}, \hat{\beta})$ given above.

The efficiency of $\hat{N}$ and $\hat{\beta}$ is given in Table 9 for $t = 5(5)30$ and a range of values of β. As discussed in the Appendix, the efficiency is defined as

$$E = \frac{|I^{-1}(N, \beta)|}{|D(\hat{N}, \hat{\beta})|},$$

where

$$I(N, \beta) = \begin{bmatrix} N \sum_{i=0}^{t} \frac{1}{\pi_i} \left(\frac{d\pi_i}{d\beta}\right)^2 & -\frac{t}{\beta(t+\beta)} \\ -\frac{t}{\beta(t+\beta)} & \frac{t}{N\beta} \end{bmatrix}$$

may be taken as the information matrix for N and β when $\alpha = 1$. From the formula for $D(\hat{N}, \hat{\beta})$

$$|D(\hat{N},\hat{\beta})| = \left[\frac{(1+\beta)(t+\beta)\beta}{t}\right]^2 \frac{1}{(t+1)(2+\beta)} .$$

No convenient analytic expression for $|I(N,\beta)|$ appears possible. It is easily seen that the efficiency depends only upon t and β, thus Table 9 is valid for all N large enough so that $D(\hat{N},\hat{\beta})$ is a good approximation to the exact dispersion matrix.

As seen from Table 9 the efficiency is excellent for β and t in the range anticipated to apply to livetrapping studies, say $\beta \geq 5$ which corresponds to $E(p) \leq .167$. It can be expected that the efficiency achieved in practice will be greater than .98.

The question now arises, of how much importance is this special case $\alpha = 1$? This is an empirical question for which no definite answer exists at present. It may be that this case is especially valid and useful as a model for the distribution of capture probabilities. The density function in this case is $f(p) = \beta(1-p)^{\beta-1}$, $0 < p < 1$. It is seen that $f(p)$ is strictly monotone decreasing with $f(0) = \beta$ and $f(1) = 0$ whenever $\beta \geq 1$, that is, whenever $E(p) \leq 5$. For $E(p)$ in the range anticipated for livetrapping this density indicates that most capture probabilities will be small to moderate, with only a small percent of large capture probabilities. Conversely, it is unlikely that there will be many extremely small capture probabilities.

In an attempt to derive an estimate of population size when

capture probabilities varied, Eberhardt, et al. (1963) chose to model the capture frequencies as a sample from a geometric distribution ($P\{Y=i\} = (1-Q)Q^{i-1}$, $i = 1, 2, \ldots$). They did this because it was noted that the geometric model gave a good fit to many capture frequency records. On the basis of this model they suggested the estimator

$$\hat{N}_G = \frac{S_t}{1-S_t/C_t},$$

which will be called the geometric estimator. Note that $\hat{N}_G$ and $\hat{N}$ given in (4.3.1) differ only by the constant term $1-1/t$.

Skellam (1948) has shown that the Beta-binomial distribution converges to a limiting negative-binomial distribution when α is held fixed and β, $t \uparrow +\infty$ subject to β/t is held constant. With $\alpha = 1$ and β and t large enough, the Beta-binomial distribution will be approximately the same as the geometric distribution with $Q = t/(t+\beta+1)$, except for the obvious restriction that only a finite number of integers have positive probability with the Beta-binomial. Thus it is not surprising that $\hat{N}$ and $\hat{N}_G$ are very similar.

Eberhardt (1969) in a paper discussing population estimates based on the capture frequencies, recognized this relationship between the geometric distribution and the Beta-binomial distribution with $\alpha = 1$. However, the only use he made of it was to derive the estimator $\hat{N}$ for $t=2$ from the expressions for $E(f_{12})$ and $E(f_{22})$.

The geometric model was an ad-hoc model introduced on empirical grounds. By taking a more fundamental approach the estimator $\hat{N}$ is derived for a situation where the capture frequencies will approximately fit a geometric distribution. This will be a useful estimator if capture probabilities follow a $\beta(1, \beta)$ distribution. It would be worthwhile to find $\hat{\alpha}_{MLE}$ and $\hat{\beta}_{MLE}$ for a variety of livetrapping data to determine if $\alpha = 1$ is generally a reasonable assumption. For the livetrapping data examined in Tables 1 through 4, a value of $\alpha = 1$ appears tenable.

Table 9. Efficiency of the estimators $\hat{N} = \frac{S_t}{1-S_t/C_t}(1-\frac{1}{t})$ and $\hat{\beta} = \frac{tS_t - C_t}{C_t - S_t}$ when $\alpha = 1$ for selected values of β. The value of N is arbitrary since the efficiency depends only upon t and β.

		Efficiencies					
β	E(p)	t = 5	10	15	20	25	30
1	.50	.920	.878	.843	.827	.817	.809
2	.333	.964	.939	.927	.920	.915	.911
3	.250	.981	.967	.960	.955	.952	.950
4	.20	.989	.980	.975	.972	.970	.969
5	.167	.993	.987	.983	.981	.980	.980
6	.143	.995	.991	.988	.987	.986	.985
7	.125	.997	.993	.991	.990	.989	.989
8	.111	.997	.995	.993	.992	.992	.991
9	.10	.998	.996	.995	.994	.994	.993
10	.091	.998	.997	.996	.995	.995	.994
12	.077	.999	.998	.997	.997	.996	.996
14	.067	.999	.999	.998	.998	.997	.997
16	.059	1.000	.999	.999	.998	.998	.998
18	.053	1.000	.999	.999	.999	.999	.998
20	.048	1.000	.999	.999	.999	.999	.999
30	.032	1.000	1.000	1.000	1.000	1.000	1.000

5. AN EXTENSION OF THE JACKKNIFE METHOD OF BIAS REDUCTION

5.1 Introduction

The jackknife was originally devised by Quenouille (1949, 1956) as a bias reduction technique. Tukey adopted the name "jackknife" for this procedure (see Miller, 1964) and suggested it could be used to obtain approximate confidence intervals. Since then a number of papers have justified this inference procedure for selected situations (Brillinger, 1964; Miller, 1964, 1968; Arvesen, 1969). There have also been papers examining the bias reduction achieved with the jackknife in certain estimation problems (Durbin 1959; Rao, 1965; Rao and Webster, 1966; Mantel, 1967). These papers are only concerned with the elimination of a bias that is $O(\frac{1}{n})$, the resulting estimator may be biased to $O((\frac{1}{n})^2)$.

Quenouille, in his 1956 paper, gave a method for eliminating bias of higher order than $1/n$. Robson and Whitlock (1964) have actually used this procedure, in a slightly modified form. In this chapter a better formula is developed for the elimination of higher order bias. At the same time the procedure is generalized for use with biases that are $O(\frac{1}{g(n)})$. After this work was completed the author discovered other people had been working on an extension of jackknifing (Schucany et al., 1971; Adams et al., 1971).

Independence of development is established by a manuscript submitted March 31, 1971 to the Annals of Mathematical Statistics.

Let $\hat{\theta}_1, \ldots, \hat{\theta}_k$ be estimators of a real valued parameter θ. The problem of combining these estimators to get an improved estimator is a common one in statistics. The usual approach is to assume a linear combination will be used and choose coefficients to achieve a criterion such as minimum variance. Clearly, it is conceivable to combine these estimators to achieve minimum bias or even minimum mean square error. It will be seen that the extended jackknife is just a linear combination of estimators with coefficients chosen to reduce bias. The unique aspect of jackknifing is the nature of the estimators so combined; they are constructed from one initial estimator.

Schucany's approach to bias reduction has been motivated by sequence to sequence transformations for accelerating the convergence of sequences. Consequently, he defines the extended jackknife in a very nonintuitive manner, and ends up failing to get closed form results (nor does he consider generalizing the nature of the biases). My approach is the familiar one of forming a linear combination of estimators to yield an improved estimator.

The nature of the jackknife is as follows: Let $Y_1, \ldots, Y_n$ be a random sample from a distribution involving an unknown, real valued parameter θ. Assume $\hat{\theta}_n = \hat{\theta}_n(Y_1, \ldots, Y_n)$ is an estimator

of θ and assume

$$E(\hat{\theta}_n) = \theta + \frac{a_1}{n} + \frac{a_2}{n^2} + \ldots ,$$

where $a_1, a_2, \ldots$ are constants. Define

$$\hat{\theta}_{(n-1),i} = \hat{\theta}_{(n-1)}(Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n) \qquad i = 1, \ldots, n,$$

and

$$\bar{\hat{\theta}}_{(n-1)} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{(n-1),i} .$$

The jackknife estimator is defined as

$$\hat{\theta}_{J1} = n\hat{\theta}_n - (n-1)\bar{\hat{\theta}}_{(n-1)} .$$

The bias of $\hat{\theta}_n$ is $O(\frac{1}{n})$, but the bias of $\hat{\theta}_{J1}$ is $O((\frac{1}{n})^2)$:

$$E(\hat{\theta}_{J1}) = \theta - \frac{a_2}{n(n-1)} - \frac{a_3}{n(n-1)} (\frac{1}{n} + \frac{1}{n-1}) + \ldots .$$

The jackknife can be used to set an approximate confidence interval on θ by defining n estimators, sometimes called pseudovalues, by

$$\hat{\theta}_n^{(i)} = n\hat{\theta}_n - (n-1)\hat{\theta}_{(n-1),i} \qquad i = 1, \ldots, n.$$

Note that $\hat{\theta}_{J1} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_n^{(i)}$.

The $\hat{\theta}_n^{(i)}$ are identically distributed random variables, and by treating them as if they were iid, an approximate Student's t-statistic can be constructed. When this is done the variance of $\hat{\theta}_{J1}$ is estimated by

$$\frac{\sum_{i=1}^{n} (\hat{\theta}_n^{(i)} - \hat{\theta}_{J1})^2}{n(n-1)} .$$

This inferential aspect of jackknifing will not be considered here; however, some consideration will be given to the possibility of using the jackknife approach to estimate the variance of higher order jackknife estimators.

It is assumed that $\hat{\theta}_n$ has been chosen to estimate θ either because no other estimator is available, or on the basis of some criterion such as sufficiency or Maximum Likelihood. Whatever the case, it is assumed that the only estimators of θ available are the statistics $\hat{\theta}_{(n-1), j_1, \ldots, j_i}$, defined below. The jackknife must be constructed from these estimators, as no other information is available.

Because $Y_1, \ldots, Y_n$ is a random sample, it may be assumed without loss of generality that $\hat{\theta}_n(Y_1, \ldots, Y_n)$ is a symmetric function of its arguments. It is assumed below (see (5.2.1)) that $E(\hat{\theta}_n)$ is such that if $n_1 \neq n_2$, then $E(\hat{\theta}_{n_1}) \neq E(\hat{\theta}_{n_2})$.

Let i be an integer such that $\hat{\theta}_{(n-i)}$ is defined, and let $j_1, \ldots, j_i$ be a combination of i integers from the set $\{1, \ldots, n\}$. For any such combination define $\hat{\theta}_{(n-i), j_1, \ldots, j_i}$ as the estimator based on the $n-i$ random variables remaining after $Y_{j_1}, \ldots, Y_{j_i}$ are dropped from the sample. By assumption the only available unbiased estimators of $E(\hat{\theta}_{(n-i)})$ are these $\hat{\theta}_{(n-i), j_1, \ldots, j_i}$. Thus the MVUE of $E(\hat{\theta}_{(n-i)})$ is the U-statistic (Fraser, 1957)

$$\bar{\hat{\theta}}_{(n-i)} = \frac{1}{\binom{n}{i}} \sum_{1 \le j_1 < \cdots < j_i \le n} \hat{\theta}_{(n-i), j_1, \ldots, j_i} .$$

For notational convenience let $\bar{\hat{\theta}}_{(n)} = \hat{\theta}_n$.

The basis for the jackknife method of bias reduction is the set of estimators $\bar{\hat{\theta}}_{(n-i)}$.

5.2 The Extension

Repeating some of the above, $Y_1, \ldots, Y_n$ is a random sample from a probability distribution depending upon θ. These variables may be vector-valued. Let $\hat{\theta}_n$ be an estimator of θ which is defined for all $n \ge \ell$, where ℓ is a fixed integer ≥ 1. Finally assume

$$(5.2.1) \qquad E(\hat{\theta}_n) = \theta + \sum_{j=1}^{\infty} \frac{a_j}{[g(n)]^j} \qquad n \ge \ell,$$

where $a_1, a_2, \ldots$ are constants.

The function g is assumed to have certain properties:

(5.2.2a) g is strictly monotone increasing and unbounded from above. Thus $g(y) \uparrow + \infty$, implying $E(\hat{\theta}_n) \to \theta$, and

(5.2.2b) $g(\ell) > 0$, thus $g(n) > 0$, $\forall \quad n \geq \ell$.

(5.2.2c) Let c be a constant, then

$$\lim_{y \to \infty} \left[\frac{g(y)}{g(y+c)}\right] = 1.$$

This implies that for any fixed integer j,

$$\lim_{y \to \infty} \left[\frac{g(y)}{g(y+c)}\right]^j = 1 .$$

An example of such a function is $g(y) = (y+\beta)^{\alpha}$, $\alpha > 0$, $y > -\beta$.

Consider a linear combination of the estimators $\overline{\hat{\theta}}_{(n-i)}$, $i = 0, \ldots, k$, where $n-k \geq \ell$:

$$\hat{\theta}_{Jk} = \sum_{i=0}^{k} x_i \overline{\hat{\theta}}_{(n-i)} .$$

Using (5.2.1) it is easy to write the expected value of $\hat{\theta}_{Jk}$. After rearrangement of terms

$$(5.2.3) \qquad E(\hat{\theta}_{Jk}) = \theta\left[\sum_{i=0}^{k} x_i\right] + \sum_{j=1}^{\infty} a_j \left[\sum_{i=0}^{k} \frac{x_i}{(g(n-i))^j}\right] .$$

Let $\underline{x} = (x_0, x_1, \ldots, x_n)'$. It is desired to choose a vector of coefficients, $\underline{x}$, which gives "good" bias reduction.

By looking at (5.2.3) it is apparent that $\underline{x}$ should satisfy $x_0+x_1+\ldots+x_k = 1$. The imposition of k additional, independent linear restraints on $\underline{x}$ will uniquely determine this vector. Presumably the initial terms of the bias expression in (5.2.3) are the dominant ones, consequently a reasonable approach to choosing $\underline{x}$ is to adopt the following $k+1$ restrictions:

$$\sum_{i=0}^{k} x_i = 1 ,$$

$$\sum_{i=0}^{k} \frac{x_i}{(g(n-i))^j} = 0 \qquad j = 1, 2, \ldots, k.$$

This system of equations has a unique solution vector $\underline{x}$, which can be found in closed form. The linear combination of the $\bar{\hat{\theta}}_{(n-i)}$'s generated by the solution to these equations will be symbolized here by $\hat{\theta}_{Jk}$, and called the kth order jackknife based on the initial estimator $\hat{\theta}_n$. For a further justification of $\hat{\theta}_{Jk}$, note that if $a_j = 0$ for all $j > k$, then $E(\hat{\theta}_{Jk}) = \theta$.

Theorem 1. Given $\hat{\theta}_n$ with $E(\hat{\theta}_n)$ as in (5.2.1) and given k such that $n-k \geq \ell$, there exists a unique linear combination of $\bar{\bar{\theta}}_n, \bar{\bar{\theta}}_{(n-1)}, \ldots, \bar{\bar{\theta}}_{(n-k)}$,

$$\hat{\theta}_{Jk} = \sum_{i=0}^{k} x_i \bar{\bar{\theta}}_{(n-i)} , \tag{5.2.4}$$

such that

$$E(\hat{\theta}_{Jk}) = \theta + \frac{(-1)^k a_{k+1}}{g(n)g(n-1)\ldots g(n-k)} + \sum_{j=2}^{\infty} a_{k+j} O\left(\left(\frac{1}{g(n)}\right)^{k+j}\right) . \tag{5.2.5}$$

Furthermore the x_i's can be explicitly given:

$$x_0 = \frac{[g(n)]^k}{\prod_{j=1}^{k} [g(n)-g(n-j)]} ,$$

$$x_i = \frac{(-1)^i [g(n-i)]^k}{\prod_{j=0}^{i-1} [g(n-j)-g(n-i)] \prod_{j=i+1}^{k} [g(n-i)-g(n-j)]} \qquad i = 1,\ldots,k-1, \tag{5.2.6}$$

$$x_k = \frac{(-1)^k [g(n-k)]^k}{\prod_{j=0}^{k-1} [g(n-j)-g(n-k)]} .$$

The proof of this theorm will be given in two parts: first showing the uniqueness of, and deriving the $\underline{x}$ vector, second proving the assertions (implicit in (5.2.5) about the bias of $\hat{\theta}_{Jk}$.

Before going on to the proof, consider the case where $g(y) = y + \beta$, $\beta > -\ell$. This is the only functional form for which the x_i's simplify. In this case

$$\hat{\theta}_{Jk} = \frac{1}{k!} \sum_{i=0}^{k} (-1)^i \binom{k}{i} (n+\beta-i)^k \hat{\theta}_{(n-i)} \,. \tag{5.2.7}$$

Setting $\beta = 0$ gives the form of $E(\hat{\theta}_n)$ Quenouille considered. Thus, in this commonly considered case, here is an explicit kth order jackknife with bias $O((1/n)^{k+1})$ (and it is not the same as the one proposed by Quenouille).

Proof of Theorem 1: The vector of coefficients $\underline{x}$ used in constructing $\hat{\theta}_{Jk}$ is determined by the equation

$$A\underline{x} = \underline{c} \,, \tag{5.2.8}$$

where $\underline{c} = [1, 0, \ldots, 0]'$, and A is a $(k+1)$ by $(k+1)$ matrix:

$$A = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ \frac{1}{g(n)} & \frac{1}{g(n-1)} & \frac{1}{g(n-2)} & \cdots & \frac{1}{g(n-k)} \\ \left(\frac{1}{g(n)}\right)^2 & \left(\frac{1}{g(n-1)}\right)^2 & \left(\frac{1}{g(n-2)}\right)^2 & \cdots & \left(\frac{1}{g(n-k)}\right)^2 \\ \vdots & \vdots & \vdots & & \vdots \\ \left(\frac{1}{g(n)}\right)^k & \left(\frac{1}{g(n-1)}\right)^k & \left(\frac{1}{g(n-2)}\right)^k & \cdots & \left(\frac{1}{g(n-k)}\right)^k \end{bmatrix} .$$

Because A is a Vandermande matrix (Perlis, 1952), its

determinant is well known:

$$|A| = \prod_{0 \le j < i \le k} \left[\frac{1}{g(n-i)} - \frac{1}{g(n-j)} \right].$$

Because of the strict monotonicity of g, it follows that for $j < i$,

$$\left[\frac{1}{g(n-i)} - \frac{1}{g(n-j)} \right] > 0,$$

which implies $|A| > 0$. Thus there is a unique solution to (5.2.8) given by $\underline{x} = A^{-1}\underline{c}$.

Let $A(i+1)$ denote the matrix that results from replacing column $i+1$ of the matrix A by the column vector $\underline{c}$, for $i = 0, 1, \ldots, k$. By Cramer's rule (Perlis, 1952)

$$x_i = \frac{|A(i+1)|}{|A|} \qquad i = 0, \ldots, k.$$

It turns out that $|A(i+1)|$ is proportional to a Vandermode determinant. For example with $i = 0$,

$$|A(1)| = \prod_{j=1}^{k} \frac{1}{g(n-j)} \begin{vmatrix} 1 & 1 & \cdots & 1 \\ \frac{1}{g(n-1)} & \frac{1}{g(n-2)} & \cdots & \frac{1}{g(n-k)} \\ \vdots & \vdots & & \vdots \\ \left(\frac{1}{g(n-1)}\right)^{k-1} & \left(\frac{1}{g(n-2)}\right)^{k-1} & \cdots & \left(\frac{1}{g(n-k)}\right)^{k-1} \end{vmatrix}.$$

The determinant on the LHS of this formula contains most of the

terms of $|A|$:

$$|A(1)| = |A| \frac{[g(n)]^k}{\prod_{j=1}^{k} [g(n)-g(n-j)]} .$$

Thus,

$$x_0 = \frac{[g(n)]^k}{\prod_{j=1}^{k} [g(n)-g(n-j)]} .$$

In the same manner $x_1, \ldots, x_k$ are found; however, for $x_1, \ldots, x_{k-1}$ the computations are tedious and care must be taken not to get confused by notation. The author found it useful in these cases to first transform by $\frac{1}{g} = f$, evaluate x_i, then transform back.

It is convenient to have a single formula for the x_i's. This is achieved by using the conventions

$$\prod_{j=0}^{-1} [g(n-j)-g(n)] \equiv 1 \qquad \text{(used for } i = 0),$$

and

$$\prod_{j=k+1}^{k} [g(n-k)-g(n-j)] \equiv 1 \qquad \text{(used for } i = k).$$

Using these conventions,

$$x_i = \frac{(-1)^i [g(n-i)]^k}{\prod_{j=0}^{i-1} [g(n-j)-g(n-i)] \prod_{j=i+1}^{k} [g(n-i)-g(n-j)]} \qquad i = 0, \ldots, k.$$

Note that from the strict monotonicity of g, it follows that $(-1)^i x_i > 0$. Thus the x_i's alternate in sign. Also $\lim_{n \to \infty} |x_i| = \infty$, which makes it difficult to investigate the asymptotic properties of $\hat{\theta}_{Jk}$.

The bias of $\hat{\theta}_{Jk}$ can be written

$$E(\hat{\theta}_{Jk}) - \theta = \sum_{j=1}^{\infty} a_{k+j} \left[\sum_{i=0}^{k} \frac{x_i}{[g(n-i)]^{k+j}} \right] .$$

Define a triple sequence

$$Q(n,k,j) = \sum_{i=0}^{k} \frac{x_i}{[g(n-i)]^{k+j}} \qquad j \geq 0,\ k \geq 0,\ n-k \geq \ell . \tag{5.2.9}$$

Then

$$E(\hat{\theta}_{Jk}) - \theta = \sum_{j=1}^{\infty} a_{k+j} Q(n,k,j).$$

The $a_{k+1}, a_{k+2}, \ldots$ are constants, thus the investigation of the bias of $\hat{\theta}_{Jk}$ reduces to an investigation of the quantities $Q(n,k,j)$. This task is approached by first deriving a recursive relationship among these quantities.

For all $n-i \geq \ell$, the following is an identity:

$$\frac{1}{g(n-i)} \equiv \frac{1}{g(n)} \left[1 + \frac{g(n) - g(n-i)}{g(n-i)} \right] .$$

Multiply this by $1/[g(n-i)]^{k+j-1}$ and substitute the result in (5.2.9) to derive

$$Q(n,k,j) = \frac{1}{g(n)}\left[Q(n,k,j-1) + \sum_{i=0}^{k} \frac{x_i[g(n)-g(n-i)]}{[g(n-i)]^{k+j}}\right].$$

The second term in the bracket above may be simplified. For $k \geq 1$:

$$\sum_{i=0}^{k} \frac{x_i[g(n)-g(n-i)]}{[g(n-i)]^{k+j}}$$

$$= \sum_{i=1}^{k} \frac{(-1)^i[g(n)-g(n-i)]}{[g(n-i)]^j \prod_{h=0}^{i-1}[g(n-h)-g(n-i)] \prod_{h=i+1}^{k}[g(n-i)-g(n-h)]}.$$

The term in the denominator for $h = 0$ will cancel with the term in the numerator. Then changing the indexing to $i' = i-1$ and $h' = h-1$ gives

$$\sum_{i'=0}^{k-1} \frac{(-1)(-1)^{i'}}{[g(n-1-i')]^j \prod_{h'=0}^{i'-1}[g(n-1-h')-g(n-1-i')] \prod_{h'=i'+1}^{k-1}[g(n-1-i')-g(n-1-h')]}$$

$$= -Q(n-1,k-1,j).$$

Provided $k \geq 1$, $Q(n-1,k-1,j)$ is defined, as originally $n-k \geq \ell$ was required $\Rightarrow (n-1)-(k-1) \geq \ell$ holds. It follows that

$$(5.2.10) \qquad Q(n,k,j) = \frac{1}{g(n)}\left[Q(n,k,j-1) - Q(n-1,k-1,j)\right],$$

$$j \geq 1, \quad k \geq 1, \quad n-k \geq \ell .$$

Some boundary conditions will be needed. First consider $Q(n,k,j)$ when $k = 0$; this makes sense if $\hat{\theta}_{J0} = \hat{\theta}_n$, $x_0 = 1$, and

$$Q(n,0,j) = \frac{1}{[g(n)]^j} \qquad j = 0, 1, \ldots .$$

The case $j = 0$ and $k \geq 1$ is trivial:

$$Q(n,k,0) = 0 \qquad k \geq 1.$$

With these boundary conditions it is possible to solve (5.2.10) for $Q(n,k,j)$ by induction. First, two additional relationships are derived. Note that the presence of n causes no difficulties, it is the variables k and j that are important.

From (5.2.10) write (for $j \geq 2$)

$$Q(n,k,j-1) = \frac{1}{g(n)}\left[Q(n,k,j-2) - Q(n-1,k-1,j-1)\right],$$

and substitute this back into the LHS of (5.2.10) to derive

$$Q(n,k,j) = \frac{1}{[g(n)]^2} Q(n,k,j-2) - \frac{1}{[g(n)]^2} Q(n-1,k-1,j-1)$$

$$- \frac{1}{g(n)} Q(n-1,k-1,j).$$

Continuing in this manner to eliminate the term involving k (rather than $k-1$) gives

$$(5.2.11) \qquad Q(n,k,j) = -\sum_{r=0}^{j-1} \left[\frac{1}{g(n)}\right]^{r+1} Q(n-1,k-1,j-r) \qquad k \geq 1,\ j \geq 1.$$

The term involving j in (5.2.10) is eliminated by the same approach, though it takes more work:

$$(5.2.12) \qquad Q(n,k,j) = \sum_{r=0}^{k} Q(n,r,1)Q(n-r,k-r,j-1) \qquad k \geq 1,\ j \geq 1.$$

The first term in the bias of $\hat{\theta}_{Jk}$ is $a_{k+1}Q(n,k,1)$. To evaluate this term put $j = 1$ in (5.2.11) to derive

$$Q(n,k,1) = -\frac{1}{g(n)}\, Q(n-1,k-1,1) .$$

Induction on k leads immediately to

$$(5.2.13) \qquad Q(n,k,1) = \frac{(-1)^k}{g(n)g(n-1)\ldots g(n-k)} \qquad k \geq 1.$$

To evaluate $Q(n,k,2)$, set $j = 2$ in (5.2.12) and use (5.2.13):

$$Q(n,k,2) = Q(n,k,1)\left[\sum_{r=0}^{k} \frac{1}{g(n-r)}\right] \qquad k \geq 1.$$

It is possible to continue using (5.2.12) to evaluate $Q(n,k,j)$ for $j > 2$, but there is no point in doing this. From (5.2.13) it follows that the leading term in the bias of $\hat{\theta}_{Jk}$ is $(-1)^k a_{k+1}/g(n)\,g(n-1)\ldots g(n-k)$. To finish the proof of Theorem 1 it will suffice to determine the asymptotic behavior of the terms $Q(n,k,j)$.

Lemma 1. When the function g satisfies the conditions (5.2.2a,b,c) and $Q(n,k,j)$ is defined by (5.2.9), then

$$\lim_{n\to\infty}\{[g(n)]^{k+j}Q(n,k,j)\} = (-1)^k\binom{j+k-1}{k} \qquad k \geq 0,\ j \geq 1. \tag{5.2.14}$$

The proof of Lemma 1 is by induction on k, allowing j to be an arbitrary integer. A useful formula in this proof is (Jolley, 1961):

$$\sum_{r=1}^{j}\binom{r+m-2}{m-1} = \binom{j+m-1}{m} \qquad m \geq 1,\ j \geq 1\ .$$

Proof of Lemma 1: If $k = 0$ then $[g(n)]^j Q(n,0,j) = 1$ for all $j \geq 1$. This establishes (5.2.14) for the case $k = 0$, $j \geq 1$.

Let m be an arbitrary integer ≥ 1. From (5.2.11)

$$[g(n)]^{m+j}Q(n,m,j) = -\sum_{r=0}^{j-1} [g(n)]^{m-1+j-r}Q(n-1,m-1,j-r),$$

$$= -\sum_{r=0}^{j-1} \left[\frac{g(n)}{g(n-1)}\right]^{m-1+j-r} [g(n-1)]^{m-1+j-r}Q(n-1,m-1,j-r).$$

Assuming (5.2.14) is true for $k = m-1$ and all $j \geq 1$, it follows that

$$\lim_{n\to\infty} \{[g(n)]^{m+j}Q(n,m,j)\} = (-1)\sum_{r=0}^{j-1} (-1)^{m-1}\binom{j-r+m-2}{m-1},$$

$$= (-1)^m \sum_{r=1}^{j} \binom{r+m-2}{m-1},$$

$$= (-1)^m \binom{j+m-1}{m} \qquad j \geq 1.$$

Because (5.2.14) is true for $k = 0$, which corresponds to $m = 1$, it follows by induction that Lemma 1 is true.

This completes the proof of Theorem 1, since (5.2.5) follows from (5.2.13) and Lemma 1.

The jackknife is often introduced by specifying the sample is partitioned in groups of size $r \geq 1$. It is easy to extend Theorem 1 to the case $r \geq 1$, and it is not required that n/r be an integer.

Let r and k be positive integers such that $n-kr \geq \ell$. For

$1 \le i \le k$, let $j_1, \ldots, j_{ir}$ be a subset of size ir from the integers $1, \ldots, n$. Define $\hat{\theta}_{(n-ir), j_1, \ldots, j_{ir}}$ to be the estimator of θ based on the remaining $(n-ir)$ sample variables after $Y_{j_1}, \ldots, Y_{j_{ir}}$ are deleted, and let

$$\bar{\hat{\theta}}_{(n-ir)} = \frac{1}{\binom{n}{ir}} \sum_{1 \le j_1 < \ldots < j_{ir} \le n} \hat{\theta}_{(n-ir), j_1, \ldots, j_{ir}} .$$

A kth order jackknife is defined by $\hat{\theta}_{Jk, r} = \sum_{i=0}^{k} x_{i, r} \bar{\hat{\theta}}_{(n-ir)}$, where the coefficients $x_{i, r}$ satisfy the $k+1$ independent linear equations

$$\sum_{i=0}^{k} \frac{x_{i, r}}{[g(n-ir)]^j} = \begin{cases} 1 & j = 0, \\ 0 & j = 1, \ldots, k . \end{cases}$$

Theorem 1a. Given $\hat{\theta}_n$, with $E(\hat{\theta}_n)$ as in (5.2.1), and given r, k such that $n - kr \ge \ell$, there exists a unique linear combination of $\bar{\hat{\theta}}_{(n-ir)}$, $i = 0, \ldots, k$, call it $\hat{\theta}_{Jk, r}$ such that

$$E(\hat{\theta}_{Jk, r}) = \theta + \frac{(-1)^k a_{k+1}}{g(n)\, g(n-r) \ldots g(n-kr)} + \sum_{j=2}^{\infty} a_{k+j} O\left(\left[\frac{1}{g(n)}\right]^{k+j}\right) .$$

The coefficients of this linear combination are

$$x_{i,r} = \frac{(-1)^i [g(n-ir)]^k}{\prod_{j=0}^{i-1} [g(n-jr)-g(n-ir)] \prod_{j=i+1}^{k} [g(n-ir)-g(n-jr)]},$$

$$i = 0, \ldots, k.$$

The proof of Theorem 1a is essentially the same as the proof of Theorem 1.

It is doubtful that it would be useful to have $r > 1$ if the goal is bias reduction. To justify this statement, define

$$Q_r(n,k,j) = \sum_{i=0}^{k} \frac{x_{i,r}}{[g(n-ir)]^{k+j}} \qquad k \geq 1,\ j \geq 0.$$

Then the bias of $\hat{\theta}_{Jk,r}$ is

$$E(\hat{\theta}_{Jk,r}) - \theta = \sum_{j=1}^{\infty} a_{k+j} Q_r(n,k,j) .$$

The approach used for $r = 1$ can be used to derive the results

$$Q_r(n,k,1) = \frac{(-1)^k}{g(n)g(n-r)\ldots g(n-kr)},$$

and

$$(5.2.15) \qquad Q_r(n,k,j) = \sum_{i=0}^{k} Q_r(n,i,1) Q_r(n-ir, k-i, j-1),$$

for $k \geq 1$, $j \geq 1$. It is seen that $Q_r(n,k,1) = (-1)^k |Q_r(n,k,1)|$, and it follows from (5.2.15) by induction that

$$Q_r(n,k,j) = (-1)^k \sum_{i=0}^{k} |Q_r(n,i,1)| \, |Q_r(n-ir,k-i,j-1)| ,$$

$$= (-1)^k |Q_r(n,k,j)| \qquad k \geq 1, \; j \geq 1.$$

Because g is strictly monotone increasing it follows that $\min_r |Q_r(n,k,1)| = |Q_1(n,k,1)|$ and by induction

$$\min_r |Q_r(n,k,j)| = |Q_1(n,k,j)| \qquad k \geq 1, \; j \geq 1.$$

This result suggests it would be best to use $r = 1$ when the goal is bias reduction.

5.3 Higher Order Jackknifing Considered as a Recursive Procedure

In the previous section it was shown that removal of biases of higher order than $1/g(n)$ can be achieved with an explicit kth order jackknife. This is desirable for computational purposes, but it gives no hint as to how to extend the notation of pseudovalues, and hence generalize the variance estimation aspect of jackknifing. A recursive formulation of $\hat{\theta}_{Jk}$ is possible.

Define a sequence of jackknife type estimators via

$$\hat{\theta}_n^k = Z_{n,k}\hat{\theta}_n^{k-1} + (1-Z_{n,k})\bar{\hat{\theta}}_{(n-1)}^{k-1}, \tag{5.3.1}$$

for $k = 1, \ldots, n-\ell$ where $\hat{\theta}_n^0 = \hat{\theta}_n$ and $Z_{n,k}$ is a function of n and k only. The function $\bar{\hat{\theta}}_{(n-1)}^{k-1}$ is defined by

$$\bar{\hat{\theta}}_{(n-1)}^{k-1} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{(n-1),i}^{k-1},$$

where $\hat{\theta}_{(n-1),i}^{k-1}$ is the $k-1$ order "jackknife" computed for the $n-1$ sample variables after Y_i has been dropped out. For $k = 1$, $Z_{n,1} = n$ gives the usual first order jackknife.

For $k = 1$, $\hat{\theta}_n^1$ is seen to be a linear combination of $\bar{\hat{\theta}}_{(n)}$ and $\bar{\hat{\theta}}_{(n-1)}$. Assume that $\hat{\theta}_n^k$ is a linear combination of $\bar{\hat{\theta}}_{(n-i)}$ for $i = 0, \ldots, k$, $k \leq n-\ell-1$, and consider

$$\hat{\theta}_n^{k+1} = Z_{n,k+1}\hat{\theta}_n^k + (1-Z_{n,k+1})\bar{\hat{\theta}}_{(n-1)}^k.$$

Now

$$\bar{\hat{\theta}}_{(n-1)}^k = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{(n-1),i}^k,$$

and by assumption each $\hat{\theta}_{(n-1),i}^k$ is a linear combination of $\bar{\hat{\theta}}_{(n-1-r),i}$, $r = 0, \ldots, k$ where $\bar{\hat{\theta}}_{(n-1-r),i}$ is computed by dropping Y_i from the sample and computing the usual $\bar{\hat{\theta}}_{((n-1)-r)}$ estimator from the remaining sample values:

$$\overline{\hat{\theta}}_{(n-1-r),i} = \frac{1}{\binom{n-1}{r}} \sum \hat{\theta}_{(n-1-r),i,j_1,\ldots,j_r},$$

where $j_1,\ldots,j_r$ is a combination of r integers from the set $\{1,\ldots,i-1,i+1,\ldots,n\}$ and the summation extends over all such combinations. Thus for some coefficients b_r

$$\overline{\hat{\theta}}^{k}_{(n-1)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{r=0}^{k} b_r \overline{\hat{\theta}}_{(n-1-r),i},$$

$$= \sum_{r=0}^{k} b_r \left[\sum_{i=1}^{n} \frac{1}{n}\overline{\hat{\theta}}_{(n-1-r),i}\right],$$

$$= \sum_{r=0}^{k} b_r \left[\sum_{i=1}^{n} \frac{1}{n}\frac{1}{\binom{n-1}{r}} \sum \hat{\theta}_{(n-1-r),i,j_1,\ldots,j_r}\right],$$

$$= \sum_{r=0}^{k} b_r \frac{1}{n}\frac{1}{\binom{n-1}{r}} \left[\sum_{i=1}^{n} \sum \hat{\theta}_{(n-(r+1)),i,j_1,\ldots,j_r}\right].$$

The quantity in brackets is seen to be equal to

$$(r+1) \sum \hat{\theta}_{(n-(r+1)),m_1,\ldots,m_{r+1}},$$

for $m_1,\ldots,m_{r+1}$ a combination of $r+1$ integers out of $\{1,\ldots,n\}$, and the summation extends over all such combinations. Thus

$$\bar{\hat{\theta}}^{k}_{(n-1)} = \sum_{r=0}^{k} b_r \bar{\hat{\theta}}_{(n-(r+1))} ,$$

is seen to be a linear combination of $\bar{\hat{\theta}}_{(n-i)}$, $i = 1, \ldots, k+1$, which shows that $\hat{\theta}^{k+1}_{n}$ is a linear combination of $\bar{\hat{\theta}}_{(n-i)}$, $i = 0, \ldots, k+1$.

Because $\hat{\theta}^{1}_{n}$ is a linear combination of $\bar{\hat{\theta}}_{n}$ and $\bar{\hat{\theta}}_{(n-1)}$, it follows by induction that for all $k = 1, \ldots, n-\ell$, there exists coefficients $x_0, \ldots, x_k$ such that

$$\hat{\theta}^{k}_{n} = \sum_{i=0}^{k} x_i \bar{\hat{\theta}}_{(n-i)} .$$

Theorem 2. Assume that $E(\hat{\theta}_n)$ is given by (5.2.1) with g satisfying assumptions (5.2.2a, b, c). Let

$$Z_{n,k} = \frac{g(n)}{g(n)-g(n-k)} ,$$

then $\hat{\theta}^{k}_{n} = \hat{\theta}_{Jk}$. Proof: For $k = 1$ it is obvious that $\hat{\theta}^{1}_{n} = \hat{\theta}_{J1}$. Thus

$$E(\hat{\theta}^{1}_{n}) = \theta + \frac{-a_2}{g(n)g(n-1)} + \sum_{j=2}^{\infty} a_{1+j} Q(n, 1, j) .$$

For arbitrary k such that $k \leq n-\ell-1$ assume that $\hat{\theta}^{k}_{n} = \hat{\theta}_{Jk}$. Thus

$$E(\hat{\theta}_n^k) = \theta + \frac{(-1)^k a_{k+1}}{g(n)\ldots g(n-k)} + \sum_{j=2}^{\infty} a_{k+j} Q(n,k,j).$$

It follows that

$$E(\hat{\theta}_n^{k+1}) = \frac{g(n)E(\hat{\theta}_n^k) - g(n-k-1)E(\hat{\theta}_{(n-1)}^k)}{g(n) - g(n-k-1)},$$

$$= \theta + 0 \cdot a_{k+1} + \sum_{j=2}^{\infty} b_j a_{k+j}$$

(the nature of the coefficients b_j is not important). This shows that the coefficients x_i in

$$\hat{\theta}_n^{k+1} = \sum_{i=0}^{k+1} x_i \hat{\theta}_{(n-i)}$$

must satisfy

$$\sum_{i=0}^{k+1} \frac{x_i}{g(n-i)^j} = \begin{cases} 1 & j = 0, \\ 0 & j = 1, \ldots, k+1, \end{cases}$$

hence, by Theorem 1, $\hat{\theta}_n^{k+1} = \hat{\theta}_{Jk}$.

The value of this recursive expression for $\hat{\theta}_{Jk}$ is it allows an extension of the concept of a pseudovalue: Let

$$\hat{\theta}^{i}_{Jk} = \frac{g(n)\hat{\theta}_{Jk-1} - g(n-k)\hat{\theta}_{Jk-1,i}}{g(n)-g(n-k)},$$

for $i = 1, \ldots, n$, where $\hat{\theta}_{Jk-1,i}$ is the $k-1$ order jackknife estimator computed from the $n-1$ sample elements after Y_i is dropped. Because $\hat{\theta}_{J0} = \hat{\theta}_n$, this is the usual definition of a pseudovalue when $k = 1$. Finally, $\hat{\theta}_{Jk} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}^{i}_{Jk}$ suggests that the quantity $\sum_{i=1}^{n}(\hat{\theta}^{i}_{Jk} - \hat{\theta}_{Jk})^2$ may be used in constructing an estimator of the variance of $\hat{\theta}_{Jk}$.

Consider the simple case $\theta = E(Y)$, $\hat{\theta}_n = \overline{Y}$ and $g(n) = n$. Then $\hat{\theta}_{Jk} = \overline{Y}$,

$$\hat{\theta}_{Jk-1,i} = \overline{Y} + \frac{\overline{Y} - Y_i}{n-1},$$

and

$$\sum_{i=1}^{n}(\hat{\theta}^{i}_{Jk} - \hat{\theta}_{Jk})^2 = \left(\frac{1}{k}\frac{n-k}{n-1}\right)^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2 .$$

If $V(Y) = \sigma^2 < \infty$, then

$$\text{Correlation}(\hat{\theta}^{i}_{Jk}, \hat{\theta}^{j}_{Jk}) = \frac{k^2}{k^2+2k+n} - \frac{1}{n-1} \qquad i \neq j.$$

Moreover

$$\frac{\sum_{i=1}^{n} (\hat{\theta}_{Jk}^{i} - \hat{\theta}_{Jk})^2}{n-1} \xrightarrow{p} \frac{1}{k^2}\sigma^2.$$

Asymptotically the correlations go to zero, but for small n they may be very large. It is seen that an estimate of $V(\hat{\theta}_{Jk})$ should be (in this case)

$$\left(\frac{k}{n-k}\right)^2 \frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta}_{Jk}^{i} - \hat{\theta}_{Jk})^2 .$$

It is reasonable to think similar results will hold for higher order jackknifing in non-trivial cases such as Miller (1964, 1968) and Arvesen (1969) examined.

Expression (5.3.1) is also useful for extending a result of Adams et al. (1971). Adams shows that if $\hat{\theta}_{J1}$ is unbiased, then so is $\hat{\theta}_{J2}$ unbiased for θ. Assume that

$$E(\hat{\theta}_{n}^{k}) = \theta \qquad \forall \quad n \geq \ell + k.$$

Assuming $n-k > \ell$ then

$$E(\hat{\theta}_{n}^{k+1}) = Z_{n,k} E(\hat{\theta}_{n}^{k}) + (1 - Z_{n,k}) E(\bar{\hat{\theta}}_{(n-1)}^{k}) ,$$

$$= Z_{n,k}\theta + (1 - Z_{n,k}) E(\hat{\theta}_{(n-1),1}^{k}) ,$$

$$= Z_{n,k}\theta + (1 - Z_{n,k})\theta = \theta.$$

Thus it is seen that if $E(\hat{\theta}_n^k) = \theta$ for all $n \geq \ell+k$, then all higher order jackknife estimators computed from (5.3.1) will also be unbiased.

In this section it has been shown that any kth order jackknife estimator defined by (5.3.1) is equivalent to some linear combination of the statistics $\bar{\hat{\theta}}_{(n-i)}$, $i = 0, \ldots, k$. Consequently, if $E(\hat{\theta}_n)$, $n \geq \ell$ is known it will always be possible to write down $E(\hat{\theta}_n^k) = \sum_{i=0}^{k} x_i E(\bar{\hat{\theta}}_{(n-i)})$ and it may then be possible to choose the coefficients to achieve good bias reduction. In Section 5.2 this was done assuming $E(\hat{\theta}_n)$ to be, as in (5.2.1), a power series in $1/g(n)$.

6. A NONPARAMETRIC APPROACH TO ESTIMATION OF POPULATION SIZE

6.1 Linear Combinations of the Capture Frequencies as Estimators of Population Size

The main features of the model for multiple capture-recapture studies are population closure and the capture probabilities are a random sample from an arbitrary distribution, F, on the interval $(0, 1]$. In Section 2.3 it was shown that without specifying F the capture frequencies, $f_{1t}, \ldots, f_{tt}$, are a sufficient statistic and they have a multinomial distribution:

$$P\{f_{1t}, \ldots, f_{tt} | N\} = \binom{N}{f_{0t} \cdots f_{tt}} \prod_{i=0}^{t} (\pi_i)^{f_{it}},$$

where the π_i, given by (2.3.4), depend only upon F.

In this chapter a nonparametric procedure for estimating N will be developed. In this procedure F is neither specified nor estimated. The basic motivation for considering such a nonparametric approach is the desirability of having a robust estimator of population size. It will be shown that this goal of robustness has been substantially realized.

If $\hat{N} = \sum_{i=1}^{t} a_i f_{it}$, then it can be expressed in the form $N = S_t + \hat{f}_{0t}$, where $\hat{f}_{0t}$ is an estimator of the number of

individuals never captured. It seems reasonable to use the conditional (on S_t) variance to measure the uncertainty associated with the estimator $\hat{N}$. Unfortunately, the same problem arises here as was discussed in Section 3.3 with regard to ML estimation of N.

The conditional distribution of the frequencies is also multinomial and can be written as

$$P\{f_{1t},\ldots,f_{tt}|S_t\} = \binom{S_t}{f_{1t}\cdots f_{tt}} \prod_{i=1}^{t} \left[\frac{E(f_{it}|S_t)}{S_t}\right]^{f_{it}},$$

where $E(f_{it}|S_t) = (S_t/E(S_t))E(f_{it})$. It follows that

$$E(\hat{N}|S_t) = \frac{S_t}{E(S_t)}\,E(\hat{N}),$$

and

$$V(\hat{N}|S_t) = \sum_{i=1}^{t} (a_i)^2 E(f_{it}|S_t) - \frac{[E(\hat{N}|S_t)]^2}{S_t}.$$

The MVUE of $V(\hat{N}|S_t)$ is

$$\hat{V}(\hat{N}|S_t) = \frac{S_t}{S_t - 1}\left[\sum_{i=1}^{t} (a_i)^2 f_{it} - \frac{(\hat{N})^2}{S_t}\right]. \tag{6.1.1}$$

It is well known that any linear combination of multinomial random variables is asymptotically normally distributed. It follows that for large enough S_t, both the conditional and unconditional

distributions of $\hat{N}$ are approximately normal. In the conditional distribution of $\hat{N}$ the parameter N appears only in conjunction with the multiplier $S_t/E(S_t)$ which is an unknown nuisance parameter. Thus it is not appropriate to construct confidence intervals on N using the estimated conditional variance of $\hat{N}$. It is possible to use the conditional distribution of $\hat{N}$ in testing the equality of two linear combinations of the expected frequencies; the value of this will be shown below.

If $\hat{N}_1 = \sum_{i=1}^{t} a_i f_{it}$ and $\hat{N}_2 = \sum_{i=1}^{t} b_i f_{it}$, then a test of the null hypothesis $E(\hat{N}_1) = E(\hat{N}_2)$ versus $E(\hat{N}_1) \neq E(\hat{N}_2)$ can be conducted conditional on S_t because $E(\hat{N}_1) = E(\hat{N}_2) <=> E(\hat{N}_1 | S_t) = E(\hat{N}_2 | S_t)$. Given $H_0 : E(\hat{N}_1 | S_t) = E(\hat{N}_2 | S_t)$ then

$$\hat{N}_1 - \hat{N}_2 = \sum_{i=1}^{t} (a_i - b_i) f_{it} \dot{\sim} N(0, V(\hat{N}_1 - \hat{N}_2 | S_t))$$

and (6.1.1) can be used to get an estimate of $V(\hat{N}_1 - \hat{N}_2 | S_t)$.

For testing hypotheses of the form $E(\hat{N}) = N_0$ versus $E(\hat{N}) \neq N_0$ a conditional test can not be used because under $H_0 : E(\hat{N}) = N_0$, the distribution of $\hat{N} - N_0$ conditional on S_t is approximately $N((\frac{S_t}{E(S_t)} - 1)N_0, V(\hat{N} | S_t))$. To carry out such a test, or construct confidence limits on N from $\hat{N}$, requires the unconditional probability model for $\hat{N}$. This is unfortunate in that $V(\hat{N})$

depends upon N.

The unconditional distribution of $\hat{N}$ is approximately $N(E(\hat{N}), V(\hat{N}))$, where

$$V(\hat{N}) = \sum_{i=1}^{t} (a_i)^2 E(f_{it}) - \frac{[E(\hat{N})]^2}{N} .$$

If $E(\hat{N})$ is not too different from N, then a reasonably good estimator of $V(\hat{N})$ is

$$\hat{V}(\hat{N}) = \sum_{i=1}^{t} (a_i)^2 f_{it} - \hat{N} , \tag{6.1.2}$$

and $(\hat{N}-E(\hat{N}))/\sqrt{\hat{V}(\hat{N})}$ has approximately a standard normal distribution. Let $\omega_{\alpha/2}$ be the $1 - \alpha/2$ percentile point of the $N(0, 1)$ distribution. If $|N-E(\hat{N})|/\sqrt{\hat{V}(\hat{N})}$ is small (such as $\leq .2$) then the interval

$$[\hat{N}-\omega_{\alpha/2}\sqrt{\hat{V}(\hat{N})}, \hat{N}+\omega_{\alpha/2}\sqrt{\hat{V}(\hat{N})}]$$

should provide a useful approximation to a $1-\alpha$ level confidence interval on N.

Because $V(\hat{N})$ depends upon N it does not appear possible to improve to any significant extent (6.1.2) as an estimator of $V(\hat{N})$. But a slightly improved confidence interval can be found by using the pivotal quantity

$$\frac{\hat{N}-N}{\sqrt{\sum_{i=1}^{t}(a_i)^2 f_{it} - \frac{(\hat{N})^2}{N}}} \sim N(0,1) ,$$

and solving for N_L and N_U such that

$$N_L = \hat{N} - \omega_{\alpha/2}\sqrt{\sum_{i=1}^{t}(a_i)^2 f_{it} - \frac{(\hat{N})^2}{N_L}}$$

and

$$N_U = \hat{N} + \omega_{\alpha/2}\sqrt{\sum_{i=1}^{t}(a_i)^2 f_{it} - \frac{(\hat{N})^2}{N_U}} \quad .$$

Let $N_L^* = \hat{N} - \omega_{\alpha/2}\sqrt{\hat{V}(\hat{N})}$ and $N_U^* = \hat{N} + \omega_{\alpha/2}\sqrt{\hat{V}(\hat{N})}$. Observe that

$$\sum_{i=1}^{t}(a_i)^2 f_{it} - \frac{(\hat{N})^2}{N} = \hat{V}(\hat{N}) + \hat{N}(1 - \frac{\hat{N}}{N}) ,$$

and $N < \hat{N} \Rightarrow \hat{N}(1 - \frac{\hat{N}}{N}) < 0$ while $N > \hat{N} \Rightarrow \hat{N}(1 - \frac{\hat{N}}{N}) > 0$. It follows that

$$\sqrt{\hat{V}(\hat{N})} > \sqrt{\hat{V}(\hat{N}) + \hat{N}(1 - \frac{\hat{N}}{N_L})} ,$$

$$\hat{N} - \omega_{\alpha/2}\sqrt{\hat{V}(\hat{N})} < \hat{N} - \omega_{\alpha/2}\sqrt{\hat{V}(\hat{N}) + \hat{N}(1 - \frac{\hat{N}}{N_L})} ,$$

$$N_L^* < N_L .$$

Similarly it follows that $N_U^* < N_U$. Hence, the improved confidence interval $[N_L, N_U]$ is not symmetric about $\hat{N}$, but is skewed to the right which is appropriate considering that $S_t \leq N$ is known to hold.

In addition to these results for one linear combination of the frequencies, note that if $L_j = \sum_{i=1}^{t} a_{ij} f_{it}$, $j = 1, \ldots, k$ then for large N, $\underline{L} = (L_1, \ldots, L_k)'$ has approximately a multivariate normal distribution (Rao, 1965) with mean vector $E(\underline{L})$, and dispersion matrix $\not\Sigma = [\mathrm{Cov}(L_i, L_j)]$ where

$$\mathrm{Cov}(L_i, L_j) = \sum_{\ell=1}^{t} (a_{\ell i})(a_{\ell j}) E(f_{\ell t}) - \frac{E(L_i)E(L_j)}{N},$$

$$i, j = 1, \ldots, k.$$

6.2 Application of the Jackknife in the Present Problem

As given in Section 2.2 the basic data are the individual capture records which are conveniently expressed in matrix form as $[X_{ji}]$, $j = 1, \ldots, N$, $i = 1, \ldots, t$. In what follows the order of the indexing of the population is unimportant. Let it be assumed that $y_{jt} = \sum_{i=1}^{t} X_{ji}$ is greater than zero for $j = 1, \ldots, S_t$, with $y_{jt} = 0$ for $j = S_t+1, \ldots, N$. Thus the data which are in fact observed are

$$X = \begin{bmatrix} X_{11} & X_{12} \cdots X_{1t} \\ X_{21} & X_{22} \cdots X_{2t} \\ \vdots & \vdots \quad\quad \vdots \\ X_{S_t 1} & X_{S_t 2} \cdots X_{S_t t} \end{bmatrix}. \tag{6.2.1}$$

For application of the jackknife the most appropriate units of sampling effort appear to be days. Let the sample be represented as $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_t$, where $\underline{X}_i$ is column i of X, i.e., $\underline{X}_i$ gives the capture records for day i. Let the initial estimator $\hat{N}_{J0}$ be S_t, the number of individuals known to be in the population. As shown at the end of Section 3.1, S_t is the nonparametric ML estimator of N. Clearly S_t is biased and this bias decreases as t increases. It will be assumed that $E(S_t) = N + \frac{a_1}{t} + \frac{a_2}{t^2} + \ldots$, for some constants $a_1, a_2, \ldots$. This should be at least approximately valid for most reasonable distributions of capture probabilities.

The results of Chapter 5 will be used to compute $\hat{N}_{J1}$ through $\hat{N}_{J5}$, making the identifications $n \equiv t$, $\hat{\theta}_n \equiv S_t$ and $g(t) = t$. From (5.2.7)

$$\hat{N}_{Jk} = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (t-j)^k \bar{\hat{\theta}}_{(t-j)}. \tag{6.2.2}$$

It remains only to compute the U-statistics $\bar{\hat{\theta}}_{(t-j)}$, $j = 1, \ldots, 5$.

Let

$Z_{1;i}$ = the number of individuals seen exactly once, that one time being on day i, $i = 1, \ldots, t$.

Then

$$\hat{\theta}_{(t-1),\,i} = S_t - Z_{1;i},$$

and

$$\bar{\hat{\theta}}_{(t-1)} = \frac{1}{t}\sum_{i=1}^{t}\hat{\theta}_{(t-1),\,i} = S_t - \frac{1}{t}\sum_{i=1}^{t} Z_{1;i} = S_t - \frac{1}{t}f_{1t}.$$

Setting $k = 1$ in (6.2.2) it follows that

$$\hat{N}_{J1} = S_t + \frac{t-1}{t}f_{1t}.$$

In general define

$Z_{j;i_1,\ldots,i_j}$ = the number of individuals captures exactly j times ($j \leq t$), once on each day $i_1, \ldots, i_j$ (and not captured on any other days), where $i_1, \ldots, i_j$ is a combination of the integers $1, \ldots, t$.

Because each individual has its own capture history, it is counted once and only once in the set of numbers $Z_{j;i_1,\ldots,i_j}$, $j = 1, \ldots, t$. In particular it is seen that

$$f_{jt} = \sum_{\{i_1,\ldots,i_j\} \subseteq \{1,\ldots,t\}} Z_{j;i_1,\ldots,i_j},$$

where the notation associated with the summation sign indicates summation is to be over all combinations of j integers out of the set of integers $\{1,\ldots,t\}$.

If days $i_1,\ldots,i_j$ are dropped from the sample, then the number of individuals captured at least once in the remaining sample is

$$\hat{\theta}_{(t-j),i_1,\ldots,i_j} = S_t - \sum_{r=1}^{j}\left[\sum_{\{m_1,\ldots,m_r\} \subseteq \{i_1,\ldots,i_j\}} Z_{r;m_1,\ldots,m_r}\right].$$

The summation in brackets is over $\binom{j}{r}$ possible combinations of integers $\{m_1,\ldots,m_r\} \subseteq \{i_1,\ldots,i_j\}$. Finally, it follows that

$$\bar{\hat{\theta}}_{(t-j)} = S_t - \frac{1}{\binom{t}{j}} \sum_{\{i_1,\ldots,i_j\} \subseteq \{1,\ldots,t\}} \sum_{r=1}^{j}\left[\sum_{\{m_1,\ldots,m_r\} \subseteq \{i_1,\ldots,i_j\}} Z_{r;m_1,\ldots,m_r}\right],$$

$$\bar{\hat{\theta}}_{t-j)} = S_t - \frac{1}{\binom{t}{j}} \sum_{r=1}^{j} \binom{t-r}{j-r} \sum_{\{m_1,\ldots,m_r\} \subseteq \{1,\ldots,t\}} Z_{r;m_1,\ldots,m_r},$$

$$(6.2.3) \qquad \bar{\bar{\theta}}_{(t-j)} = S_t - \frac{1}{\binom{t}{j}} \sum_{r=1}^{j} \binom{t-r}{j-r} f_{rt} .$$

There is only one difficult step in the above sequence of equations. To justify it, observe that for any $\{m_1, \ldots, m_r\} \subseteq \{1, \ldots, t\}$, there are $j-r$ integers to be chosen from the remaining $t-r$ integers to complete the set $\{i_1, \ldots, i_j\}$ containing $\{m_1, \ldots, m_r\}$.

Equation (6.2.3) may be substituted into (6.2.2) but this accomplishes nothing useful. It appears necessary to derive the formula for $\hat{N}_{Jk}$ for given k by straightforward but lengthy algebraic manipulations. For $k = 1, 2, 3, 4, 5$ this has been done and the results are given in Table 10.

Given the expected frequencies, or the distribution F, the results of Section 6.1 can be used to compute $E(\hat{N}_{Jk})$ and $V(\hat{N}_{Jk})$. This has been done by the author for a variety of Beta distributions of capture probabilities, with the conclusion that the sequence $\hat{N}_{J1}, \ldots, \hat{N}_{J5}$ generally has decreasing bias and increasing variance.

The pattern generally found in applying the jackknife to live-trapping data is exemplified by computing $\hat{N}_{Jk}$ and $\hat{SD}(\hat{N}_{Jk}) = \sqrt{\hat{V}(\hat{N}_{Jk})}$ for the data of Edwards and Eberhardt given in Table 1 (in this study $N = 135$).

Table 10. Jackknife estimators of population size starting with S_t as the initial estimator, and assuming $E(S_t) = N + \frac{a_1}{t} + \frac{a_2}{t^2} + \ldots$.

$$\hat{N}_{J1} = S_t + \left(\frac{t-1}{t}\right) f_{1t}$$

$$\hat{N}_{J2} = S_t + \left(\frac{2t-3}{t}\right) f_{1t} - \frac{(t-2)^2}{t(t-1)} f_{2t}$$

$$\hat{N}_{J3} = S_t + \left(\frac{3t-6}{t}\right) f_{1t} - \left(\frac{3t^2-15t+19}{t(t-1)}\right) f_{2t} + \frac{(t-3)^3}{t(t-1)(t-2)} f_{3t}$$

$$\hat{N}_{J4} = S_t + \left(\frac{4t-10}{t}\right) f_{1t} - \left(\frac{6t^2-36t+55}{t(t-1)}\right) f_{2t} + \left(\frac{4t^3-42t^2+148t-175}{t(t-1)(t-2)}\right) f_{3t} - \frac{(t-4)^4}{t(t-1)(t-2)(t-3)} f_{4t}$$

$$\hat{N}_{J5} = S_t + \left(\frac{5t-15}{t}\right) f_{1t} - \left(\frac{10t^2-70t+125}{t(t-1)}\right) f_{2t} + \left(\frac{10t^3-120t^2-485t-660}{t(t-1)(t-2)}\right) f_{3t}$$

$$- \left(\frac{(t-4)^5-(t-5)^5}{t(t-1)(t-2)(t-3)}\right) f_{4t} + \frac{(t-5)^5}{t(t-1)(t-2)(t-3)(t-4)} f_{5t}$$

k	$\hat{N}_{Jk}$	$\hat{SD}(\hat{N}_{Jk})$
0	76	---
1	116.61	8.89
2	141.45	14.90
3	158.58	21.93
4	170.28	31.11
5	176.45	43.45

6.3 A Proposed Estimation Procedure

Because the bias of $\hat{N}_{Jk}$ generally is decreasing while the variance is increasing as k increases, it is anticipated that the mean square error of $\hat{N}_{Jk}$ will initially decrease, then rise again. By examining the theoretical mean square error of $\hat{N}_{Jk}$ for $k = 1, \ldots, 5$ over a variety of distributions for $5 \leq t \leq 30$, it was observed that the minimum was usually achieved at $k = 1, 2,$ or 3. The exact $\hat{N}_{Jk}$ which achieved the minimum mean square error varied considerably according to the distribution of capture probabilities and the value of t. Accordingly, no rule can be formulated independent of the data to specify a value of k such that $\hat{N}_{Jk}$ should be used for any given study. An objective procedure is needed whereby the data themselves can be used to indicate which $\hat{N}_{Jk}$ should be used as the estimator of N for that study. The following procedure is proposed.

Test the hypothesis that there is no difference between the expected values of $\hat{N}_{J1}$ and $\hat{N}_{J2}$, i.e., test

$$H_{01}: E(\hat{N}_{J2} - \hat{N}_{J1}) = 0 \qquad \text{vs.}$$

$$H_{a1}: E(\hat{N}_{J2} - \hat{N}_{J1}) \neq 0$$

(a two sided test is used because the direction of the biases of $\hat{N}_{J1}$ and $\hat{N}_{J2}$ is not known for certain). If H_{01} is not rejected this is interpreted as evidence that the change (decrease) in bias effected by using $\hat{N}_{J2}$ rather than $\hat{N}_{J1}$ is small relative to the variance of $\hat{N}_{J2}$. Given the generally smaller variance of $\hat{N}_{J1}$, it is concluded that there is no reason to use $\hat{N}_{J2}$, rather $\hat{N}_{J1}$ should be taken as the estimator of N.

If H_{01} is rejected this is interpreted as evidence of significant bias reduction relative even to the increased variance of $\hat{N}_{J2}$. The estimator $\hat{N}_{J2}$ should be preferred to $\hat{N}_{J1}$. But further bias reduction may be possible. Before accepting $\hat{N}_{J2}$ as the estimator to be used with the study at hand, test

$$H_{02}: E(\hat{N}_{J3} - \hat{N}_{J2}) = 0 \qquad \text{vs.}$$

$$H_{a2}: E(\hat{N}_{J3} - \hat{N}_{J2}) \neq 0.$$

If H_{02} is not rejected, use $\hat{N}_{J2}$, otherwise continue the process in the obvious manner. Let that estimator chosen by this process be called $\hat{N}_J$, the jackknife estimator.

The general procedure for choosing $\hat{N}_J$ is as follows:

Test the hypotheses

$$H_{0i}: E(\hat{N}_{Ji+1} - \hat{N}_{Ji}) = 0 \quad \text{vs.}$$

$$H_{ai}: E(\hat{N}_{Ji-1} - \hat{N}_{Ji}) \neq 0 ,$$

sequentially for $i \leq 4$, and choose $\hat{N}_J = \hat{N}_{Ji}$ such that H_{0i} is the first null hypothesis not rejected. These null hypotheses are not expected to be true, rather the procedure is a reasonably objective guide for choosing an estimator for any given study. There is room for judgment in this choice, for example it may enter via the choice of significance level for a given test.

The actual test of H_{0i} is based on the fact that $\hat{N}_{Ji+1} - \hat{N}_{Ji}$ is a linear combination of the capture frequencies, therefore for some coefficients $a_1, \ldots, a_t$, $\hat{N}_{Ji+1} - \hat{N}_{Ji} = \sum_{i=1}^{t} a_i f_{it}$. It follows from (6.1.1) that

$$\hat{V}(\hat{N}_{Ji+1} - \hat{N}_{Ji} | S_t) = \frac{S_t}{S_t - 1}\left[\sum_{i=1}^{t} (a_i)^2 f_{it} - \frac{(\hat{N}_{Ji+1} - \hat{N}_{Ji})^2}{S_t}\right].$$

Given H_{0i},

$$T_i = \frac{\hat{N}_{Ji+1} - \hat{N}_{Ji}}{\sqrt{\hat{V}(\hat{N}_{Ji+1} - \hat{N}_{Ji} | S_t)}}$$

has approximately a $N(0, 1)$ distribution.

Given that $\hat{N}_J = \hat{N}_{Jk}$ has been chosen as the estimator of N,

then an estimate of $V(\hat{N}_J)$ is given by (6.1.2) with $\hat{N} = \hat{N}_{Jk}$. This will be an unbiased estimate of $V(\hat{N}_{Jk})$ if $E(\hat{N}_{Jk}) = N$. For large S_t

$$\frac{\hat{N}_{Jk} - E(\hat{N}_{Jk})}{\sqrt{\hat{V}(\hat{N}_{Jk})}} \sim N(0, 1),$$

and assuming $|N - E(\hat{N}_{Jk})| / \sqrt{\hat{V}(\hat{N}_{Jk})}$ is small it follows that

$$\frac{\hat{N}_{Jk} - N}{\sqrt{\hat{V}(\hat{N}_{Jk})}} \sim N(0, 1)$$

which allows approximate confidence intervals for N to be constructed. As discussed in Section 6.2 improved confidence intervals can be constructed using

$$\frac{\hat{N}_{Jk} - N}{\sqrt{\sum_{i=1}^{t} (a_i)^2 f_{it} - \frac{(\hat{N}_{Jk})^2}{N}}} \sim N(0, 1),$$

where

$$\hat{N}_{Jk} = \sum_{i=1}^{t} a_i f_{it}.$$

The procedure of testing the hypotheses H_{0i}, $i = 1, 2, 3, 4$, should be viewed as a very useful guide to the choice of $\hat{N}_J$. Obviously there is no significance level α such that if H_{0i-1} is

rejected at this level, and H_{0i} is not rejected, then $\hat{N}_J = \hat{N}_{Ji}$ is clearly indicated. It is anticipated that the significance levels actually achieved, $P_i = P\{|T_i| \geq |T_i^*|\}$, will be increasing. Here $T_i \sim N(0, 1)$, and T_i^* is the observed test statistic value. If P_{i-1} is small, such as smaller than .05, while P_i is much larger than .05 it is reasonable to take $\hat{N}_J = \hat{N}_{Ji}$. One possible procedure of course is to carry out all the tests at the 5% level.

When these hypothesis tests are carried out for data of Edwards and Eberhardt shown in Table 1, the results are

Null hypothesis	T_i^*	P_i
H_{01}	4.053	<.0001
H_{02}	2.071	.0383
H_{03}	1.071	.2842
H_{04}	.417	.6766

This suggests $\hat{N}_{J3}$ as the estimate to use for these data.

Analytic investigation of the properties of $\hat{N}_J$ is theoretically possible because the approximate distribution of $(T_1, T_2, T_3, T_4)'$ is multivariate normal with mean vector elements

$$E(T_i) = \frac{E(\hat{N}_{Ji+1} - \hat{N}_{Ji})}{\sqrt{E_{S_t} V(\hat{N}_{Ji+1} - \hat{N}_{Ji} | S_t)}} \qquad i = 1, 2, 3, 4,$$

and a dispersion matrix Σ, the elements of which can also be

computed for any given distribution of capture probabilities. Let $q_i = P\{\hat{N}_J = \hat{N}_{Ji}\}$, and let q_5 be the probability that all H_{0i} are rejected so no estimator is chosen. Assume a fixed significance level α and let $\omega_{\alpha/2}$ be the $1 - \alpha/2$ percentile point of the standard normal distribution. Then

$$q_1 = P\{|T_1| \leq \omega_{\alpha/2}\},$$

$$q_2 = P\{|T_2| \leq \omega_{\alpha/2}, |T_1| > \omega_{\alpha/2}\},$$

$$q_3 = P\{|T_3| \leq \omega_{\alpha/2}, |T_1| > \omega_{\alpha/2}, |T_2| > \omega_{\alpha/2}\},$$

$$q_4 = P\{|T_4| \leq \omega_{\alpha/2}, |T_i| > \omega_{\alpha/2}, i = 1, 2, 3\},$$

and

$$q_5 = P\{|T_i| > \omega_{\alpha/2}, i = 1, 2, 3, 4\}.$$

It is seen that $\sum_{i=1}^{5} q_i = 1$. Assuming that q_5 is small a good approximation to $E(\hat{N}_J)$ should be

$$\sum_{i=1}^{4} \frac{q_i}{1-q_5} E(\hat{N}_{Ji}).$$

Similarly, other properties of $\hat{N}_J$ could be found.

It is easy to find q_1, but q_2, q_3 and q_4 are not easily found, as these latter quantities basically require the ability to evaluate the multivariate normal distribution function for 2, 3, or 4 variates.

To get an idea of the properties of $\hat{N}_J$, this estimation procedure has been applied to some simulated livetrapping data from Burnham and Overton (1969). Table 11 gives the results of this study. For these simulated livetrapping studies N was always 100. There may be some loss of generality because of this for it is seen that the bias of $\hat{N}_{Jk}$ is proportional to N, while $SD(\hat{N}_{Jk})$ is proportional to $\sqrt{N}$.

These simulated data were generated as follows. For a given distribution of capture probabilities, a random sample $p_1, \cdots, p_{100}$ was drawn to represent the population. Livetrapping was then simulated for 30 days. Twenty such studies were done for each of a variety of distributions, a different set of capture probabilities being used each time. Three types of distributions were used: Beta = $B(\alpha, \beta)$, uniform on $(0, \theta)$, (symbolized as $U(0, \theta)$), and the constant distribution $C(\theta)$ which assigns probability 1 to θ. This last class of (degenerate) distributions correspond to the model wherein all individuals have constant and equal capture probabilities.

For this simulation evaluation the hypothesis tests were all conducted at the 5% level. A problem which arises when doing this is that sometimes all H_{0i} are rejected when $t > 5$. When this happened the frequencies, the estimators $\hat{N}_{Jk}$ and the test statistic values were printed out and examined. Under certain circumstances $\hat{N}_J < S_t$ is possible. These cases were also examined in detail. In

Table 11. Simulation evaluation of the jackknife estimation procedure for N = 100. All hypothesis tests conducted at the 5% significance level. The entries are based on 20 simulated livetrapping studies carried out to 30 days with each distribution of capture probabilities. Entries are explained in the text.

t	k	$\hat{N}_J$	$\hat{SD}(\hat{N}_J)$	Coverage	S_t	$\hat{N}_{SH}$	$\hat{SD}(\hat{N}_{SH})$
Distribution = B(.3158, 1.0) E(p) = .24							
5	1.7	65.7	6.7	1	48.1	48.3	5.4
10	1.4	72.6	6.3	3	57.0	53.2	4.7
15	1.4	77.7	6.3	4	62.2	56.7	4.4
20	1.4	80.5	6.4	5	65.1	59.1	4.4
25	1.7	86.1	7.6	6	67.9	61.0	4.3
30	1.2	80.2	5.2	3	69.1	62.5	4.3
Distribution = B(1.0, 3.1667) E(p) = .24							
5	1.6	83.8	7.6	9	59.3	65.8	6.6
10	1.1	91.6	6.1	14	74.1	72.4	5.2
15	1.1	95.3	5.6	17	80.8	76.9	4.5
20	1.2	99.7	5.9	19	85.3	79.2	3.7
25	1.2	101.9	5.9	15	87.7	81.2	3.4
30	1.0	99.2	4.5	17	89.7	82.9	3.1
Distribution = B(6.0, 19.0) E(p) = .24							
5	2.2	115.0	11.2	14	71.8	92.8	10.8
10	1.1	110.6	6.7	15	89.7	93.9	4.2
15	1.2	109.3	5.7	19	95.9	95.6	2.7
20	1.0	103.8	3.5	20	98.1	96.6	2.0
25	1.0	102.6	2.8	20	99.1	97.3	1.6
30	1.0	101.3	1.9	18	99.5	97.7	1.3
Distribution = U(0, .48) E(p) = .24							
5	1.8	99.5	9.4	18	67.0	80.4	9.6
10	1.1	98.5	6.2	19	81.3	82.3	5.1
15	1.0	99.2	5.0	18	87.2	84.8	4.0
20	1.0	99.8	4.5	20	90.3	86.6	3.3
25	1.0	99.3	4.0	20	92.1	87.9	2.8
30	1.0	99.3	3.6	18	93.2	89.0	2.5

Table 11. Continued.

t	k	$\hat{N}_J$	$\hat{SD}(\hat{N}_J)$	Coverage	S_t	$\hat{N}_{SH}$	$\hat{SD}(\hat{N}_{SH})$
Distribution = C(.24) E(p) = .24							
5	2.2	122.4	12.0	14	75.5	102.0	7.2
10	1.1	111.0	6.4	9	93.5	100.4	3.2
15	1.0	106.9	4.3	14	98.2	100.2	2.2
20	1.0	102.7	2.5	20	99.7	100.3	1.6
25	1.0	100.9	1.4	20	99.9	100.2	1.2
30	1.0	100.2	.6	19	99.9	100.1	1.1
Distribution = B(.25, 1.0) E(p) = .2							
5	1.5	56.7	5.7	0	43.3	43.1	5.6
10	1.5	67.0	6.3	1	51.9	48.0	5.4
15	1.3	69.0	5.7	1	55.8	50.9	5.4
20	1.3	72.0	5.6	2	59.3	53.3	5.4
25	1.3	75.1	6.0	2	61.5	55.1	5.4
30	1.6	80.2	7.2	5	63.4	56.6	5.3
Distribution = B(1.0, 4.0) E(p) = .2							
5	2.1	87.6	8.7	12	55.5	66.8	7.0
10	1.5	97.0	8.3	16	72.1	72.5	4.8
15	1.4	101.7	7.8	18	79.8	79.5	4.1
20	1.0	98.8	5.6	20	84.2	79.2	3.6
25	1.1	100.5	5.4	20	87.3	81.3	3.3
30	1.0	99.4	4.7	19	89.1	83.0	3.0
Distribution = B(4.0, 16.0) E(p) = .2							
5	2.6	110.7	12.2	16	64.8	89.6	11.4
10	1.1	105.9	7.1	19	83.1	89.4	5.0
15	1.0	105.9	5.8	18	91.4	91.6	3.5
20	1.0	104.9	4.6	17	94.9	93.1	3.0
25	1.0	103.2	3.7	18	96.8	94.3	2.6
30	1.0	102.1	2.8	18	98.0	99.0	2.2
Distribution = U(0, .4) E(p) = .2							
5	2.3	96.6	10.6	17	58.6	77.5	8.4
10	1.1	97.1	6.7	18	76.4	80.7	5.7
15	1.1	97.9	5.8	16	83.2	82.9	4.5
20	1.0	96.9	4.7	15	86.6	84.4	4.0
25	1.0	97.5	4.3	15	89.1	85.7	3.8
30	1.1	98.5	4.3	15	90.4	86.7	3.7

Table 11. Continued.

t	k	$\hat{N}_J$	$\widehat{SD}(\hat{N}_J)$	Coverage	S_t	$\hat{N}_{SH}$	$\widehat{SD}(\hat{N}_{SH})$
Distribution = C(.2) E(p) = .2							
5	2.4	116.8	12.5	17	67.7	100.1	11.0
10	1.1	113.7	7.6	11	88.9	99.0	4.2
15	1.1	105.8	5.1	18	95.8	99.1	2.6
20	1.0	103.6	3.2	19	98.6	99.4	1.7
25	1.0	101.6	2.1	20	99.4	99.5	1.3
30	1.0	100.9	1.3	19	99.9	99.6	1.1
Distribution = B(.1905, 1.0) E(p) = .16							
5	2.1	51.4	6.6	0	35.1	35.9	3.5
10	1.7	58.9	7.0	2	42.3	39.8	2.6
15	1.2	57.0	5.1	0	46.0	42.2	2.6
20	1.4	61.3	5.6	2	48.5	44.0	2.7
25	1.3	63.8	5.7	2	51.2	45.7	2.8
30	1.7	69.2	7.1	4	52.9	47.0	2.9
Distribution = B(1.0, 5.25) E(p) = .16							
5	2.4	79.2	10.2	10	46.5	63.2	9.6
10	1.4	89.2	8.1	9	64.0	67.5	4.5
15	1.6	101.0	9.2	16	73.4	71.8	3.5
20	1.1	98.5	7.6	18	79.2	75.0	2.9
25	1.0	97.7	5.6	17	83.0	77.4	2.9
30	1.1	99.1	5.5	18	85.7	79.3	2.9
Distribution = B(4.0, 21.0) E(p) = .16							
5	3.4	110.0	14.1	19	55.4	87.0	13.2
10	1.3	108.8	8.7	18	77.9	90.0	5.8
15	1.1	110.3	7.2	15	87.7	91.6	4.4
20	1.0	107.6	5.6	16	92.6	92.9	3.6
25	1.0	106.1	4.8	19	95.3	93.9	3.1
30	1.0	104.7	4.2	19	96.8	94.7	2.6
Distribution = U(0, .32) E(p) = .16							
5	3.0	97.8	12.3	17	53.5	77.4	10.5
10	1.2	98.6	7.9	18	72.1	80.0	4.9
15	1.2	101.1	6.9	19	80.6	82.3	3.6
20	1.0	98.3	5.4	19	84.6	83.8	3.0
25	1.0	98.5	4.9	18	87.4	85.0	2.9
30	1.0	98.8	4.6	18	89.2	86.0	2.8

Table 11. Continued.

t	k	$\hat{N}_J$	$\hat{SD}(\hat{N}_J)$	Coverage	S_t	$\hat{N}_{SH}$	$\hat{SD}(\hat{N}_{SH})$
Distribution = C(.16) E(p) = .16							
5	3.3	118.2	14.6	16	58.9	100.7	10.0
10	1.2	114.6	8.5	15	83.2	99.9	4.8
15	1.1	112.5	6.6	9	93.2	100.3	2.9
20	1.0	107.7	4.8	19	96.8	100.0	2.0
25	1.0	105.0	3.6	18	99.0	100.1	1.5
30	1.0	102.6	2.5	18	99.4	100.0	1.3
Distribution = B(.1364, 1.0) E(p) = .12							
5	1.8	36.5	5.3	0	25.5	26.3	5.3
10	1.4	44.2	5.5	1	32.2	30.2	5.4
15	1.2	46.0	4.7	0	36.1	32.6	5.7
20	1.2	49.3	4.9	1	39.0	34.5	6.1
25	1.6	50.4	5.0	1	40.3	35.9	6.3
30	1.2	51.7	5.0	1	41.5	37.0	6.3
Distribution = B(1.0, 7.333) E(p) = .12							
5	2.7	74.4	10.5	8	39.3	59.0	9.7
10	1.5	84.9	8.8	5	57.0	64.0	6.6
15	1.3	90.9	7.8	10	66.7	67.9	6.2
20	1.0	90.0	6.1	12	72.8	70.6	6.2
25	1.2	97.6	6.9	18	77.7	73.4	5.2
30	1.1	98.1	6.1	19	81.2	75.7	4.8
Distribution = B(3.0, 22.0) E(p) = .12							
5	3.3	93.2	13.0	17	45.0	80.1	18.3
10	1.7	104.7	11.1	16	64.6	81.6	8.7
15	1.3	108.1	9.0	17	76.6	84.7	6.6
20	1.1	105.8	7.0	19	83.5	86.6	5.0
25	1.0	104.3	5.9	18	87.8	88.1	4.7
30	1.0	104.5	5.4	18	91.0	89.4	4.0
Distribution = U(0, .24) E(p) = .12							
5	3.5	92.2	13.2	15	43.9	79.2	16.2
10	1.7	101.0	10.5	18	64.2	80.9	10.0
15	1.2	101.0	8.0	18	74.6	81.8	6.2
20	1.3	105.3	7.5	18	81.2	83.5	4.3
25	1.2	103.2	6.6	18	84.8	84.7	3.8
30	1.1	101.2	5.5	18	87.4	85.7	3.3

Table 11. Continued.

t	k	$\hat{N}_J$	$\hat{SD}(\hat{N}_J)$	Coverage	S_t	$\hat{N}_{SH}$	$\hat{SD}(\hat{N}_{SH})$
Distribution = C(.12) E(p) = .12							
5	3.8	110.6	14.9	17	48.6	111.8	33.4
10	1.8	119.7	11.4	12	73.3	101.6	13.5
15	1.2	114.9	8.3	12	85.8	99.8	7.1
20	1.0	109.2	6.0	13	91.7	99.2	5.6
25	1.0	107.3	5.2	15	95.7	99.5	4.5
30	1.0	105.9	4.1	19	97.9	99.6	3.6
Distribution = B(1.0, 10.1111) E(p) = .09							
5	3.1	69.3	11.0	8	33.2	64.5	25.5
10	1.9	89.5	11.3	10	51.4	66.1	11.4
15	1.4	90.1	8.9	13	60.8	67.5	8.7
20	1.4	94.5	8.6	11	67.1	69.5	7.4
25	1.2	94.8	7.5	13	71.9	71.6	6.8
30	1.3	97.6	8.0	14	75.1	73.1	6.5
Distribution = B(9.0, 91.0) E(p) = .09							
5	3.8	88.3	13.3	17	37.2	110.5	48.5
10	2.3	112.0	14.6	16	58.6	92.7	12.4
15	1.4	111.3	10.3	17	73.2	93.4	8.8
20	1.2	112.4	8.4	15	82.4	94.3	7.0
25	1.0	112.6	7.2	10	88.8	95.5	5.3
30	1.0	112.0	6.7	12	92.4	96.1	4.4
Distribution = U(0, .18) E(p) = .09							
5	3.7	79.1	12.6	12	34.2	84.2	22.9
10	2.4	105.4	14.4	18	54.1	78.4	10.5
15	1.3	98.5	8.8	14	66.7	80.0	8.0
20	1.3	102.4	8.6	17	74.0	81.4	6.3
25	1.1	101.1	6.9	20	79.4	83.1	5.7
30	1.0	99.8	6.2	20	84.4	83.8	5.0
Distribution = C(.09) E(p) = .09							
5	3.9	94.5	14.0	18	38.7	117.1	34.2
10	2.5	123.4	16.2	18	60.9	104.7	15.2
15	1.5	117.1	10.8	14	75.6	101.2	9.9
20	1.1	115.6	8.4	10	84.7	101.2	6.6
25	1.0	111.6	6.9	12	89.9	100.4	4.7
30	1.0	111.1	6.0	13	94.0	100.5	3.3

Table 11. Continued.

t	k	$\hat{N}_J$	$\hat{SD}(\hat{N}_J)$	Coverage	S_t	$\hat{N}_{SH}$	$\hat{SD}(\hat{N}_{SH})$
Distribution = B(10, 15, 6667) E(p) = .06							
5	3.4	50.8	9.8	0	23.0	47.1	11.1
10	2.2	74.7	11.6	8	38.2	54.0	8.6
15	1.6	82.9	10.0	9	49.3	60.0	8.0
20	1.5	87.7	9.6	9	56.3	62.9	7.2
25	1.6	92.7	9.9	11	61.5	65.1	6.9
30	1.3	88.9	7.9	10	65.1	66.5	6.6
Distribution = B(3.0, 47.0) E(p) = .06							
5	3.6	61.0	10.9	3	24.5	85.0	52.6
10	2.6	92.6	14.4	18	41.8	81.2	20.7
15	1.9	98.6	12.2	20	54.1	79.9	9.6
20	1.8	108.5	12.4	17	63.5	82.6	8.4
25	1.4	104.7	9.7	18	70.1	83.5	7.4
30	1.2	106.3	10.5	17	75.5	84.8	6.0
Distribution = U(0, .12) E(p) = .06							
5	3.8	61.8	11.1	4	24.6	98.0	45.1
10	2.6	94.7	14.5	17	42.2	83.0	14.8
15	1.9	98.3	12.3	18	54.0	80.9	9.7
20	1.5	97.1	10.0	12	62.0	79.9	8.1
25	1.2	96.3	8.1	16	69.0	80.9	6.4
30	1.1	98.0	7.5	18	73.1	81.3	5.5
Distribution = C(.06) E(p) = .06							
5	3.9	68.4	11.8	6	26.7	122.5	65.8
10	2.6	103.8	15.3	18	45.5	101.6	22.4
15	1.8	106.3	12.3	16	59.3	96.7	13.0
20	1.6	114.4	11.6	15	70.1	98.6	11.1
25	1.2	111.5	9.0	17	77.6	98.3	7.9
30	1.0	111.5	7.7	17	83.4	98.7	6.5
Distribution = B(1.0, 24.0) E(p) = .04							
5	3.2	40.2	8.4	0	17.5	44.3	21.0
10	2.4	65.7	11.5	6	29.4	58.3	27.8
15	2.1	79.2	12.1	10	39.0	59.1	13.5
20	2.2	90.5	13.3	13	45.8	60.9	11.4
25	1.8	89.2	11.4	14	51.6	62.8	10.3
30	1.6	91.1	10.5	11	56.5	64.2	8.7

Table 11. Continued.

t	k	$\hat{N}_J$	$\widehat{SD}(\hat{N}_J)$	Coverage	S_t	$\hat{N}_{SH}$	$\widehat{SD}(\hat{N}_{SH})$
Distribution = B(2.0, 48.0) E(p) = .04							
5	3.4	44.3	9.1	0	17.7	66.9	31.8
10	2.7	72.2	12.9	8	30.7	67.3	23.1
15	2.2	80.2	12.4	12	40.3	67.6	12.5
20	1.8	89.8	11.5	13	48.8	71.1	10.3
25	1.6	92.2	10.6	11	55.6	73.7	8.7
30	1.5	98.3	10.4	13	61.8	75.2	7.5
Distribution = U(0, .08) E(p) = .04							
5	3.4	43.1	8.9	0	17.5	72.3	46.0
10	2.3	68.6	11.6	8	30.2	76.5	21.9
15	2.4	89.4	14.0	17	41.4	77.6	14.8
20	2.1	98.2	13.8	15	49.7	78.3	11.1
25	1.6	96.1	11.6	13	56.3	77.7	8.3
30	1.5	97.8	10.4	17	62.5	79.0	5.9
Distribution = C(.04) E(p) = .04							
5	3.5	48.2	9.6	0	18.5	104.3	41.2
10	2.6	82.7	13.8	13	33.3	102.5	34.0
15	2.5	107.6	16.1	20	45.8	101.8	18.1
20	2.2	116.8	15.4	17	55.9	102.0	14.1
25	1.8	115.2	13.5	17	63.6	98.6	12.4
30	1.7	118.9	13.1	14	70.2	98.0	10.2

both cases, as explained below, a subjective decision was then made as to the value $\hat{N}_J$ to be used.

The entries in Table 11 are mostly averages. For each distribution, and each value of t, the 20 values of $\hat{N}_J = \hat{N}_{Jk}$ were averaged as were the 20 values of k. These averages are identified in Table 11 simply as "k" and "$\hat{N}_J$." The standard deviation of $\hat{N}_J$ was computed for each study, and then averaged to obtain the estimate of $SD(\hat{N}_J)$ given under that column heading. For each study the approximate 95% confidence interval $[\hat{N}_J - 2\hat{SD}(\hat{N}_J), \hat{N}_J + 2\hat{SD}(\hat{N}_J)]$ was computed and it was recorded whether or not the interval covered N. The column headed "Coverage" gives the total times, out of 20 possible, that the interval covered N = 100.

The last three columns of Table 11 came from the original study, they are included here to aid in interpreting the properties of $\hat{N}_J$. The Schnabel estimator (Schnabel, 1938), $\hat{N}_{SH}$, is very well known and widely used to estimate population size. Because of this the average of $\hat{N}_{SH}$ over the 20 studies is given in Table 11 so that it may be compared to the estimated $E(\hat{N}_J)$. The estimated $SD(\hat{N}_{SH})$ is also given.

In Table 11, 32 different distributions of capture probabilities are used, with E(p) ranging from .24 to .04. This gives a total of 640 independent studies examined. Each study was examined on days 5(5)30, making a total of 3,840 different, though not independent,

simulated livetrapping studies. For $t = 5$, it is seen that $\hat{N}_{J4} = \hat{N}_{J5}$, which results in H_{04} never being rejected. However, for the 3,020 studies where $t > 5$, there were a total of 111 cases where H_{oi}, $i = 1, 2, 3, 4$ were all rejected at the 5% level. For the 2500 studies where $t > 5$ and $E(p) \in \{.09, .12, .16, .2, .24\}$ there were only 42 such instances of all null hypotheses being rejected, while for $E(p) \in \{.04, .06\}$ there were 69 such occurrences out of 800 studies. For $E(p) \geq .09$ there was no apparent pattern over distribution or days for these 42 cases. For $E(p) \in \{.04, .06\}$, there was no apparent pattern by distribution, but there was a pattern with respect to days of trapping, with 31 cases at $t = 10$, and only 6 cases at $t = 30$.

Based on an examination of these 111 cases when all four null hypotheses were rejected at the 5% level it was concluded that $\hat{N}_{J5}$ should not be taken as $\hat{N}_J$. Rather a choice should be made from $\hat{N}_{J1}$, $\hat{N}_{J2}$ or $\hat{N}_{J3}$. By examining all information available it was often not difficult to make what seemed like a reasonable choice.

In all other cases the objective procedure arrived at a decision $\hat{N}_J = \hat{N}_{Jk}$ for $k < 5$. However, in some of these cases it happened that $\hat{N}_J \leq S_t$. All of these cases were examined, there were 119 of them, and almost without exception this only occurs when $S_t \geq 85$ and usually S_t was ≥ 90, that is 90% of the population had been seen. In terms of distributions 81 cases occurred for constant

distributions with $E(p) \geq .09$. Only 38 cases occurred for all other distributions. When $\hat{N}_J \leq S_t$ did happen it was always the case that $\hat{N}_{J2} \leq S_t \leq \hat{N}_{J1}$ (note that $S_t \leq \hat{N}_{J1}$ is always true), so that it is sufficient to take $\hat{N}_J = \hat{N}_{J1}$, which always provided a good estimate in these cases.

By examining Table 11 it is seen that $\hat{N}_J$ is quite robust. For those distributions examined it only performs poorly when F is a $B(\alpha, 1)$ type distribution. But for these distributions of capture probabilities no estimators examined were found to have good properties. It is also seen that the standard deviation of $\hat{N}_J$ is of similar magnitude to that of $\hat{N}_{SH}$.

The order of the jackknife chosen by this procedure is seen to vary. The chief factors appear to be number of days of trapping and average capture probability. At day 5, k may easily be 2, 3 or 4. But by day 10, k is probably 1, 2 or 3. Beyond day 10 it is doubtful if $\hat{N}_{J4}$ should ever be used.

The coverage of N by the approximate 95% confidence intervals constructed with this procedure is seen to be fairly good, especially considering the range of distributions of capture probabilities examined. Table 11 indicates that the nominal confidence level will not be achieved if the absolute value of the bias of $\hat{N}_J$ is as large or larger than the standard deviation of $\hat{N}_J$. Often this is not the case and then the frequency of coverage of N appears to be 70%

or more. In general, ignoring the $B(\alpha, 1)$ distributions, the coverage appears to be 50% or better. Furthermore, slightly improved coverage might be achieved by using the improved confidence intervals previously discussed.

Most estimators do well under particular circumstances. For example, $\hat{N}_{SH}$ performs well if capture probabilities are constant over individuals. However, when capture probabilities vary it is apparent from Table 11 that $\hat{N}_{SH}$ has a negative bias which may be quite large. The jackknife estimator is often biased but this bias may be either positive or negative. Consequently, $\hat{N}_J$ tends not to have a large absolute bias. In fact, when the results in Table 11 for $\hat{N}_J$ are averaged over all (28) distributions except the $B(\alpha, 1)$ type the results are

t	$\hat{N}_J$
5	84.2
10	97.6
15	99.7
20	102.2
25	101.6
30	101.7

6.4 Possibilities for Generalizing the Jackknife Estimation Procedure

The jackknife estimator of the previous section may be generalized by retaining S_t as the initial estimator and using a different expression for $E(S_t)$; alternatively, a different initial estimator

of N may be assumed. The assumption that

$$E(S_t) = N + \frac{a_1}{t} + \frac{a_2}{t^2} + \ldots$$

is a useful omnibus assumption which leads to an easy-to-compute, robust estimator. However, if it were the case that capture probabilities were approximately $B(1, \beta)$ variables, then $E(S_t) = N + N\beta/(t+\beta)$ would be approximately true (note that $1/(t+\beta)$ has a formal expansion in powers of $1/t$ for any β, and a valid expansion if $\beta/t < 1$). This generalization is undesirable because $\hat{N}_J$ will depend upon an unknown parameter.

An alternative possibility for generalizing the jackknife procedure is to assume a different initial estimator but retain the assumption that its bias is expressible as a power series in $1/t$. In general let $N = \theta$, and assume

$$\hat{N} = \hat{\theta}_t = \sum_{\ell=1}^{t} a_{\ell t} f_{\ell t}$$

for a set of coefficients $a_{\ell t}$, $\ell = 1, \ldots, t$, $t = 1, 2, \ldots$. Application of the jackknife now requires that the quantities $\bar{\hat{\theta}}_{(t-j)}$ be computed for this new initial estimator.

Define

$f_{\ell, t-j}^{h_1, \ldots, h_{t-j}}$ = the number of individuals captured exactly ℓ times, for $\ell = 1, \ldots, t-j$, in the $t-j$ days $h_1, \ldots, h_{t-j}$ which remain after the j days $i_1, \ldots, i_j$ are dropped from the sample.

Then

$$\hat{\theta}_{(t-j), i_1, \ldots, i_j} = \sum_{\ell=1}^{t-j} a_{\ell, t-j} f_{\ell, t-j}^{h_1, \ldots, h_{t-j}} ,$$

$$\bar{\hat{\theta}}_{(t-j)} = \sum_{\ell=1}^{t-j} a_{\ell, t-j} \left[\frac{1}{\binom{t}{j}} \sum_{\{h_1, \ldots, h_{t-j}\} \subseteq \{1, \ldots, t\}} f_{\ell, t-j}^{h_1, \ldots, h_{t-j}} \right] ,$$

and finally

$$\bar{\hat{\theta}}_{(t-j)} = \sum_{\ell=1}^{t-j} a_{\ell, t-j} \bar{f}_{\ell, t-j} . \tag{6.4.1}$$

To evaluate $\bar{f}_{\ell, t-j}$, define the indicator variables

$W_{\ell; h_1, \ldots, h_{t-j}}^{r}(i) = 1$ if individual i was captured exactly ℓ times in the $t-j$ days $h_1, \ldots, h_{t-j}$, and was captured exactly r times during all t days, for $r = \ell, \ldots, \ell+j$, and $i = 1, \ldots, N$,

$= 0$ otherwise.

Let

$$W^{r}_{\ell;h_1,\ldots,h_{t-j}} = \sum_{i=1}^{N} W^{r}_{\ell;h_1,\ldots,h_{t-j}}(i) .$$

Then $W^{r}_{\ell;h_1,\ldots,h_{t-j}}$ is the total number of individuals captured exactly ℓ times in the $t-j$ days $h_1,\ldots,h_{t-j}$ which were captured exactly r total times. Finally, the term $f^{h_1,\ldots,h_{t-j}}_{\ell,t-j}$ may be partitioned into individuals captured exactly r total times, and exactly ℓ times in the days $h_1,\ldots,h_{t-j}$, for $r = \ell,\ldots,\ell+j$. Thus

$$\bar{f}_{\ell,t-j} = \frac{1}{\binom{t}{j}} \sum_{\{h_1,\ldots,h_{t-j}\}\subseteq\{1,\ldots,t\}} \left[\sum_{r=\ell}^{\ell+j} W^{r}_{\ell;h_1,\ldots,h_{t-j}}\right],$$

$$= \frac{1}{\binom{t}{j}} \sum_{r=\ell}^{\ell+j} \left[\sum_{\{h_1,\ldots,h_{t-j}\}\subseteq\{1,\ldots,t\}} W^{r}_{\ell;h_1,\ldots,h_{t-j}}\right].$$

The term in brackets above may be expanded as

$$\sum_{i\,\epsilon\, y_{it}=r} \left[\sum_{\{h_1,\ldots,h_{t-j}\}\subseteq\{1,\ldots,t\}} W^{r}_{\ell;h_1,\ldots,h_{t-j}}(i)\right]$$

where the first summation is over f_{rt} values of i.

Let i be arbitrary subject only to $y_{it} = r$. Then there exists a fixed set of r days, $\{m_1, \ldots, m_r\}$, such that individual i was captured on these days and on no other days. It follows that $W^r_{\ell;h_1,\ldots,h_{t-r}}(i) = 1$ iff $\{h_1, \ldots, h_{t-j}\}$ contains a subset of size ℓ from $\{m_1, \ldots, m_r\}$ and the other $t-j-\ell$ elements of $\{h_1, \ldots, h_{t-j}\}$ are a subset of the $t-r$ elements of $\{1, \ldots, t\} - \{m_1, \ldots, m_r\}$. There are $\binom{r}{\ell}$ subsets of size ℓ from $\{m_1, \ldots, m_r\}$ and for each of these there are $\binom{t-r}{t-j-\ell}$ remaining choices to complete a combination $h_1, \ldots, h_{t-j}$ for which $W^r_{\ell;h_1,\ldots,h_{t-j}}(i) = 1$. Consequently

$$\sum_{\{h_1,\ldots,h_{t-j}\} \subseteq \{1,\ldots,t\}} W^r_{\ell;h_1,\ldots,h_{t-j}}(i) = \binom{r}{\ell}\binom{t-r}{t-j-\ell}$$

and finally

$$\bar{f}_{\ell,t-j} = \frac{1}{\binom{t}{j}} \sum_{r=\ell}^{\ell+j} \binom{r}{\ell}\binom{t-r}{t-j-\ell} f_{rt} .$$

From (6.4.1),

$$\bar{\hat{\theta}}_{(t-j)} = \sum_{\ell=1}^{t-j} a_{\ell,t-j} \left[\sum_{r=\ell}^{\ell+j} \frac{\binom{r}{\ell}\binom{t-r}{t-j-\ell}}{\binom{t}{j}} f_{rt} \right] .$$

By extending the range of definition of the constants $a_{\ell t}$, this

formula can be rewritten as an explicit linear combination of the frequencies. Let $a_{\ell t} = 0$ if either $\ell \leq 0$, or $\ell > t$. For $\ell = 1, \ldots, t$, and $t = 1, 2, \ldots$, these constants are assumed known already. Also assume that $\binom{m}{n} = 0$ if either $n < 0$, or $n > m$. Using these conventions

$$(6.4.2) \qquad \overline{\hat{\theta}}_{(t-j)} = \sum_{i=1}^{t} \left[\sum_{\ell=i-j} a_{\ell, t-j} \frac{\binom{t}{j}\binom{t-i}{t-j-\ell}}{\binom{t}{j}} \right] f_{it} .$$

It is clear that $\hat{N}_{Jk}$ will be a linear combination of the capture frequencies but it is not at all clear what the coefficients of this combination will be. If the $a_{\ell t}$ are not all equal it may not be convenient to express $\hat{N}_{Jk}$ in closed form. However, if it is desired to use the jackknife estimation procedure developed in Section 6.3 for an initial estimator $\hat{N} = \sum_{\ell=1}^{t} a_{\ell t} f_{\ell t}$, then (6.4.2) gives the necessary formula for $\overline{\hat{\theta}}_{(t-j)}$.

7. ANOTHER LOOK AT THE MODEL

7.1 Testing the Assumption that Capture Probabilities do not Change

A number of assumptions are part of the model for capture-recapture studies, for example closure, independence of captures over individuals and days, and individual capture probabilities are constant during trapping. Given the first two of these three assumptions are true, a test can be derived for the hypothesis that the capture probabilities of the observed individuals do not change.

Consider an arbitrary individual with capture record $X_{j1}, \ldots, X_{jt}$. Given arbitrary capture probabilities then

$$P\{X_{j1}, \ldots, X_{jt}\} = \prod_{i=1}^{t} (p_{ji})^{X_{ji}} (1-p_{ji})^{1-X_{ji}},$$

and no reduction of the data is possible. But given the hypothesis that $p_{ji} = p_j$, $i = 1, \ldots, t$, then $y_{jt} = \sum_{i=1}^{t} X_{ji}$ is a minimal sufficient statistic for p_j. It follows that if $p_{ji} = p_j$ the conditional distribution of $X_{j1}, \ldots, X_{jt}$, given y_{jt}, is independent of p_j:

$$(7.1.1) \qquad P\{X_{j1}, \ldots, X_{jt} \mid y_{jt} = k\} = \frac{1}{\binom{t}{k}} \qquad k = 0, 1, \ldots, t,$$

where the X_{ji} are 0 or 1 only and their sum is k.

Let the individuals which have been captured at least once be indexed from 1 through S_t. Define the conditional random variables $\underline{X}_{j|k} = (X_{j1}, \ldots, X_{jt})'$ given that $y_{jt} = k$. Given the hypothesis that $p_{ji} = p_j$, $i = 1, \ldots, t$ and $j = 1, \ldots, S_t$, then the distribution of $\underline{X}_{j|k}$ is given by (7.1.1). By the assumption of mutual independence of the X_{ji} variables, it follows that the conditional random variables $\underline{X}_{j|k}$ are independent.

Consider some properties of the components of $\underline{X}_{j|k}$, such as $E(X_{ji}|y_{jt} = k) = P\{X_{ji} = 1|y_{jt} = k\}$. If $X_{ji} = 1$, then there are $\binom{t-1}{k-1}$ ways to complete the vector $\underline{X}_{j|k}$, hence

$$P\{X_{ji} = 1|y_{jt} = k\} = \frac{\binom{t-1}{k-1}}{\binom{t}{k}} = \frac{k}{t},$$

and $E(X_{ji}|y_{jt} = k) = k/t$. Using the same technique it is found that

$$V(X_{ji}|y_{jt} = k) = \frac{k}{t}\left(1 - \frac{k}{t}\right),$$

and

$$\mathrm{Cov}(X_{ji}, X_{j\ell}|y_{jt} = k) = -\frac{k}{t}\left(1 - \frac{k}{t}\right)\frac{1}{t-1} \qquad i \neq \ell .$$

Now define the conditional random variables

z_{ki} = the number of individuals captured on day i that were captured exactly k total times.

Analytically $z_{ki} = \sum_{j \ni y_{jt}=k} X_{ji}$, which is a sum over f_{kt} independent random variables. Also let $\underline{z}_k = (z_{k1}, \ldots, z_{kt})' = \sum_{j \ni y_{jt}=k} \underline{X}_{j|k}$.

Given $p_{ji} = p_j$ for all $j \ni y_{jt} = k$, then

$$E(z_{ki}|f_{kt}) = \frac{k}{t} f_{kt} ,$$

$$V(z_{ki}|f_{kt}) = f_{kt} \frac{k}{t}(1 - \frac{k}{t}) = \sigma_k^2 ,$$

and

$$Cov(z_{ki}, z_{k\ell}|f_{kt}) = -f_{kt} \frac{k}{t}(1 - \frac{k}{t}) \frac{1}{t-1} = -\frac{\sigma_k^2}{t-1} ,$$

for $i \neq \ell$.

Finally $n_i = \sum_{k=1}^{t} z_{ki}$ = total captures on day i. Given $H_0: p_{ji} = p_j$, $j = 1, \ldots, S_t$, then

$$E(n_i|f_{1t}, \ldots, f_{tt}) = \sum_{k=1}^{t} \frac{k}{t} f_{kt} = \frac{C_t}{t} = \bar{n} ,$$

$$V(n_i|f_{1t}, \ldots, f_{tt}) = \sum_{k=1}^{t} f_{kt} \frac{k}{t}(1 - \frac{k}{t}) = \sum_{k=1}^{t} \sigma_k^2 = \sigma^2 ,$$

and

$$\mathrm{Cov}(n_i, n_\ell \mid f_{1t}, \ldots, f_{tt}) = -\frac{\sigma^2}{t-1} \qquad i \neq \ell \ .$$

It is tempting to think that a conventional Chi-square goodness-of-fit test can be used to test H_0 by treating $\sum_{i=1}^{t} (n_i - \bar{n})^2/\bar{n}$ as approximately $\chi^2(t-1)$. This is not the case under the present approach of conditioning on the frequencies. This line of thought is valid only for the hypothesis $p_{ji} = p_j$, all $j \ni y_{jt} = 1$. For the subset of the data for which $y_{jt} = 1$, the conditional distribution of $\underline{z}_1$, given f_{1t}, is multinomial:

$$P\{z_{11}, \ldots, z_{1t} \mid f_{1t}\} = \binom{f_{1t}}{z_{11} \cdots z_{1t}} \prod_{i=1}^{t} \left(\frac{1}{t}\right)^{z_{1i}} .$$

For large enough $E(z_{1i} \mid f_{1t}) = \frac{1}{t} f_{1t}$ it follows that

$$(7.1.2) \qquad \sum_{i=1}^{t} \frac{\left(z_{1i} - \frac{1}{t} f_{1t}\right)^2}{\frac{1}{t} f_{1t}} \sim \chi^2(t-1) \ .$$

If f_{kt} is large enough that a test of $H_0: p_{ji} = p_j$, all $j \ni y_{jt} = k$, seems feasible then given this null hypothesis,

$$(7.1.3) \qquad \sum_{i=1}^{t} \frac{\left(z_{ki} - \frac{k}{t} f_{kt}\right)^2}{\frac{k}{t} f_{kt}} \quad \frac{t-1}{k-1} \sim \chi^2(t-1) \ ,$$

for $k = 1, \ldots, t-1$.

Under the general hypothesis that $p_{ji} = p_j$, $j = 1, \ldots, S_t$, then

$$(7.1.4) \qquad \frac{\sum_{i=1}^{t} (n_i - \bar{n})^2}{\sum_{k=1}^{t} f_{kt} \frac{k}{t}(1 - \frac{k}{t})} \frac{t-1}{t} \doteq \chi^2(t-1) .$$

This test is conditional on the frequencies $f_{1t}, \ldots, f_{tt}$, while the individual tests indicated by (7.1.3) are conditional on just f_{kt}. Proofs of these results will now be given.

Let $\underline{1}_t$ be a $t \times 1$ vector every element of which is 1. Then

$$E(\underline{z}_k | f_{kt}) = \frac{k}{t} f_{kt} \underline{1}_t ,$$

and letting the dispersion matrix of $\underline{z}_k$ be $\not{\Sigma}_k$, then

$$\not{\Sigma}_k = \sigma_k^2 [\frac{t}{t-1} I - \frac{1}{t-1} \underline{1}_t \underline{1}_t'] .$$

The rank of $\not{\Sigma}_k$ is $t-1$.

Because $\underline{z}_k = \sum_{j \ni y_{jt} = k} \underline{X}_{j|k}$ is the sum of f_{kt} independent, identically distributed vectors, it follows (Rao, 1965) that, for large enough f_{kt},

$$\underline{z}_k \doteq \text{MVN}(\frac{k}{t} f_{kt} \underline{1}_t, \not{\Sigma}_k) ,$$

i.e., approximately multivariate normal.

Let $\underline{n} = (n_1, \ldots, n_t)'$, then

$$\underline{n} = \sum_{k=1}^{t} \underline{z}_k = \sum_{k=1}^{t} \left[\sum_{j \ni y_{jt} = k} \underline{X}_{j|k} \right]$$

is the sum of S_t independent vectors from t different distributions. For large S_t

$$\underline{n} \sim \text{MVN}(\bar{n}\, \underline{1}_t, \not{\Sigma}) ,$$

where

$$\not{\Sigma} = \sigma^2 \left[\frac{t}{t-1} I - \frac{1}{t-1} \underline{1}_t \underline{1}_t' \right] .$$

Proof: Let p_j, $j = 1, 2, \ldots$ be a sequence of iid random variables from the distribution $F(p)$, $p \in (0, 1]$. Let X_{ji} be independent Bernoulli(p_j) random variables, $i = 1, \ldots, t$. Then the conditional distribution of $\underline{X}_{j|y_j=k}$ is given by (7.1.1) and

$$P\{y_j = k\} = \binom{t}{k} (p_j)^k (1-p_j)^{t-k} .$$

Let $f_m(k) = \sum_{j=1}^{m} 1_{\{k\}}(y_j)$ = the number of times $y_j = k$, for $j = 1, \ldots, m$. Consider

$$\frac{1}{m}\sum_{j=1}^{m} P\{1_{\{k\}}(y_i)=1\} = \binom{t}{k}\frac{1}{m}\sum_{j=1}^{m}(p_j)^k(1-p_j)^{t-k}.$$

Because $p_j \sim$ iid it follows that

$$\frac{1}{m}\sum_{j=1}^{m}(p_j)^k(1-p_j)^{t-k} \to \int_0^1 p^k(1-p)^{t-k}dF$$

a.s., and this latter quantity is greater than zero because $P\{p > 0\} = 1$. It is concluded that

$$(7.1.5) \qquad \sum_{j=1}^{m} P\{1_{\{k\}}(y_i)=1\} \to +\infty \quad \text{a.s.}$$

The event $f_m(k) \to +\infty$ occurs iff infinitely many of the independent events $\{1_{\{k\}}(y_i)=1\}$ occur. It follows from the Borel-Cantelli lemma and (7.1.5) that $P\{f_m(k) \to +\infty\} = 1$, $k = 1, \ldots, t$.

The $\underline{z}_k$ vectors are mutually independent and because $f_m(k) \to +\infty$ a.s. it follows that $\underline{z}_k \dot{\sim} \text{MVN}(\frac{k}{t} f_m(k)\underline{1}_t, \not{\Sigma}_k)$ for large m, and finally $\underline{n} = \sum_{k=1}^{t} \underline{z}_k \dot{\sim} \text{MVN}(\bar{n}\,\underline{1}_t, \not{\Sigma})$ for large m.

To establish (7.1.3) and (7.1.4) consider a random vector $\underline{X} = (X_1, \ldots, X_t)' \sim \text{MVN}(\bar{X}\,\underline{1}_t, \not{\Sigma})$, where $\bar{X} = \frac{1}{t}\sum_{i=1}^{t} X_i$, and $\not{\Sigma} = \sigma^2[\frac{t}{t-1} I - \underline{1}_t \underline{1}_t']$ for arbitrary $\sigma^2 > 0$. Then for any generalized inverse $\not{\Sigma}^-$ (Rao, 1965):

$$(\underline{X} - \overline{X}\underline{1}_t)' \not\Sigma^{-} (\underline{X} - \overline{X}\underline{1}_t) \sim \chi^2(t-1) .$$

Let $\underline{X}^* = (X_1, \ldots, X_{t-1})'$, and let $\not\Sigma^{-}$ be the t by t matrix whose upper (t-1) by (t-1) submatrix is given by

$$\frac{t-1}{t\sigma^2} [I + \underline{1}_{t-1} \underline{1}'_{t-1}],$$

and the remaining elements of $\not\Sigma^{-}$ are all zero. Then

$$(\underline{X}-\overline{X}\underline{1}_t)' \not\Sigma^{-} (\underline{X}-\overline{X}\underline{1}_t) = (\underline{X}^*-\overline{X}\underline{1}_{t-1})' [I+\underline{1}_{t-1}\underline{1}'_{t-1}](\underline{X}^*-\overline{X}\underline{1}_{t-1}) \frac{t-1}{t\sigma^2}$$

$$= [\| \underline{X}^*-\overline{X}\underline{1}_{t-1} \| + \| \underline{1}'_{t-1}(\underline{X}^*-\overline{X}\underline{1}_{t-1}) \|] \frac{t-1}{t\sigma^2}$$

$$= \left[\sum_{i=1}^{t-1} (X_i-\overline{X})^2 + (X_t-\overline{X})^2 \right] \frac{t-1}{t\sigma^2}$$

$$= \frac{\sum_{i=1}^{t} (X_i-\overline{X})^2}{\sigma^2} \frac{t-1}{t} .$$

Interpreting $\underline{X}$ as $\underline{n}$ and σ^2 as $\sum_{k=1}^{t} f_{kt} \frac{k}{t}(1-\frac{k}{t})$ gives (7.1.4). Interpreting $\underline{X}$ as $\underline{z}_k$ and σ^2 as $\sigma_k^2 = f_{kt} \frac{k}{t}(1-\frac{k}{t})$ gives (7.1.3).

This approach to testing the hypothesis that capture probabilities do not change is based on a number of assumptions. If any one

of these assumptions is false the test statistic given in (7.1.4) may be large enough to imply rejection of the hypothesis that $p_{ji} = p_j$, $j = 1, \ldots, S_t$. In particular, the rejection may be a result of lack of closure or failure of independence of captures. Robson (1971) has developed (7.1.4) as a test of association of plant species based on sampling a series of plots, and he suggested to the author it could be used to test independence of captures in livetrapping studies.

Clearly, two different hypothesis can not be tested simultaneously by the same test. If closure and independence seem likely to be true, then the test developed here should be considered as testing that capture probabilities do not change. If no one assumption stands out as weaker, or more doubtful, than the others then (7.1.4) constitutes a general test of the conformity of the data to the specified model.

7.2 A Generalization of the Model Indicating Robustness of the Jackknife Estimator

Capture probabilities may vary over days due to external influences such as weather. It is reasonable to think this sort of variation, if it occurs, is independent of the basic capture probabilities p_j. Furthermore, if this variation is infrequent, or small relative to $V(p)$ the capture frequencies may still have an approximate multinomial distribution.

Assume that $p_{ji} = p_j$ plus a small perturbation that is induced by external factors independent of the value of p_j but which acts on the individuals differentially according to the value of p_j. One way to model this is to let

$$p_{ji} = \begin{cases} p_j + p_j\delta_i & \delta_i \leq 0 \\ p_j + (1-p_j)\delta_i & \delta_i > 0 \end{cases}$$

and assume $\delta_1, \ldots, \delta_t \sim$ iid $G(\delta)$, $\delta \in (-1, 1)$. It is convenient to assume $G(\delta)$ is symmetric about 0, i.e., $1 - G(-\delta) = G(\delta)$.

Then

$$E_\delta(p_{ji}) = \phi(p_j) = p_j + \left(\int_0^1 \delta dG\right)(1-2p_j) ,$$

which is a function of p_j only.

The distribution of the basic variable $[X_{ji}]$ is now

$$P\{[X_{ji}] | N, \underline{P}, \underline{\delta}\} = \prod_{j=1}^{N} \prod_{i=1}^{t} (p_{ji})^{X_{ji}} (1-p_{ji})^{1-X_{ji}} .$$

Averaging over the minor component of variation in capture probabilities gives

$$P\{[X_{ji}] | N, \underline{P}, G\} = \prod_{j=1}^{N} \prod_{i=1}^{t} \phi(p_j)^{X_{ji}} (1-\phi(p_j))^{1-X_{ji}} ,$$

$$= \prod_{j=1}^{N} \phi(p_j)^{y_{jt}} (1-\phi(p_j))^{t-y_{jt}} .$$

Then averaging over $\underline{P}$ gives

$$P\{[X_{ji}]|N,F,G\} = \prod_{i=0}^{t}\left[\int_0^1 \phi(p)^i(1-\phi(p))^{t-i}dF\right]^{f_{it}} .$$

It follows that the capture frequencies are sufficient for the parameters N, F and G.

The unconditional distribution of the frequencies is again multinomial. This can be established with the same argument as used in Section 2.3. The sample space Ω is the same as before, so

$$P\{f_{1t},\ldots,f_{tt}|N,\underline{P},\underline{\delta}\} = \sum_{A} P\{[X_{ji}]|N,\underline{P},\underline{\delta}\}$$

where $A \subseteq \Omega$ is the set of all $\omega \in \Omega$ with corresponding capture frequencies $f_{1t},\ldots,f_{tt}$. Taking expectations over $\underline{\delta}$ and $\underline{P}$,

$$P\{f_{1t},\ldots,f_{tt}|N,F,G\} = \binom{N}{f_{0t}\cdots f_{tt}} \prod_{i=0}^{t} (\pi_i)^{f_{it}},$$

where

$$\pi_i = \int_0^1 \binom{t}{i}\phi(p)^i(1-\phi(p))^{t-i}dF .$$

Because $0 < \int_0^1 \delta dG < 1$ it follows that for $p \in (0,1]$, $0 < \phi(p) < 1$. This clearly implies $\pi_i > 0$, $i = 0,\ldots,t$ and $\sum_{i=0}^{t} \pi_i = 1$.

The jackknife estimation procedure of Chapter 6 can be applied any time the frequencies have a multinomial distribution with "sample size" N. If the frequencies are sufficient, or contain most of the information from the sample about N and $E(S_t) - N$ is approximately a power series in $1/t$ then this jackknife estimation procedure should perform fairly well. Consequently, one reason for investigating generalizations of the model is to see if the jackknife estimation procedure can be extended; or alternatively, to see if the procedure will be robust to certain types of departure from the original model.

The above examination of a generalization of the model shows that the jackknife estimator, $\hat{N}_J$, should have a degree of robustness to small or infrequent perturbations of the capture probabilities $p_1, \cdots, p_N$. This sort of deviation from the model could result in a rejection of the hypothesis that capture probabilities do not change and yet the jackknife estimator may perform satisfactorily.

8. SUMMARY

This thesis was motivated by the need for a robust estimator of population size in livetrapping studies, that is, an estimator with good properties even when capture probabilities vary among animals. Accordingly, a model is postulated in Section 2.1 which involves three assumptions, the most restrictive one being population closure. The main part of the model is the assumption that individual capture probabilities remain constant during trapping, but the set of capture probabilities $p_1, \ldots, p_N$ is a random sample from a probability distribution F on $(0, 1]$. Finally, it is assumed that individual captures are independent events.

Given this model the capture frequencies $f_{1t}, \ldots, f_{tt}$, are a sufficient statistic for N and F, and these frequencies have a multinomial distribution with cell probabilities depending only upon F.

Initially, a parametric approach is investigated by assuming F is a Beta distribution, $B(\alpha, \beta)$. In Chapter 3 consideration is given to some aspects of the general problem of ML estimation with truncated multinomial data. It was shown that $\hat{N} \doteq \tilde{N}$ under mild conditions, where both of these estimators are approximations to $\hat{N}_{MLE}$. Formulae are developed for the variance of $\tilde{N}$ and then

$V(\hat{N}_{MLE})$ may be approximated by $V(\tilde{N})$. Finally, it is shown that formally treating N as a continuous parameter in the likelihood and computing the information matrix leads to the appropriate variance formulae.

Because the conditional likelihood surface $L(\alpha, \beta | S_t)$ is ill conditioned, the estimator $\tilde{N}$ is not easily found. Consequently, the exact ML estimator, $\hat{N}_{MLE}$, is computed using the unconditional likelihood $L(N, \alpha, \beta)$. For any fixed N greater than S_t, $f_{1t}, \ldots, f_{tt}$ are the frequencies from a sample of size N from a Beta-binomial distribution. By using the method of scoring with starting values derived by the method of moments the ML estimators $\hat{\alpha}(N)$ and $\hat{\beta}(N)$ are easily found. Then by examining the sequence of points $L(N, \hat{\alpha}(N), \hat{\beta}(N))$ the value of $\hat{N}_{MLE}$ is found.

This parametric estimator of N is unsatisfactory for values of N, α and β likely to apply to real data. For values of N in the range 100 to 200 the standard error of $\hat{N}_{MLE}$ is often quite large relative to N. Furthermore, there is no reason to believe that $\hat{N}_{MLE}$ will be robust to violations of the assumed model such as slight variations in individual capture probabilities over days, or if F is not a Beta distribution. It should, however, be valuable to compute $\hat{\alpha}_{MLE}$ and $\hat{\beta}_{MLE}$ to obtain some idea of the types of Beta distributions which would best serve as models for capture probabilities. This might indicate that the assumption $\alpha = 1$ is

generally reasonable; this case is of interest because the $B(1,\beta)$ distributions are especially convenient and plausible as distributions of capture probabilities.

In Chapter 6 a nonparametric estimation procedure is developed which is valid whenever the capture frequencies have a multinomial distribution and $E(S_t)$ is, approximately, expressible as $N + \frac{a_1}{t} + \frac{a_2}{t^2} + \ldots$ where $a_1, a_2, \ldots,$ do not depend upon t. In developing this estimation procedure attention is first restricted to the class of estimators of N which are linear functions of the capture frequencies. If $\hat{N} = \sum_{i=1}^{t} b_i f_{it}$, then for large enough N, $\hat{N}$ has an approximate normal distribution, the variance of which can be estimated from the frequencies alone. If this estimator has small bias relative to its standard error then approximate confidence intervals on N can be constructed. An extension of Quenouille's jackknife method of bias reduction is used to generate some linear combinations of the frequencies which are reasonable to consider as estimators of N.

Chapter 5 gives this extension of the jackknife method of bias reduction. The main result of that chapter: If $Y_1, \ldots, Y_n$ is a random sample, and $\hat{\theta}_n(Y_1, \ldots, Y_n)$ is an estimator of θ such that

$$E(\hat{\theta}_n) = \theta + \frac{a_1}{g(n)} + \frac{a_2}{[g(n)]^2} + \ldots ,$$

where $a_1, a_2, \ldots$ do not depend upon n and g satisfies certain properties, then

$$\hat{\theta}_{Jk} = \sum_{i=0}^{k} x_i \hat{\bar{\theta}}_{(n-i)}$$

satisfies

$$E(\hat{\theta}_{Jk}) = \theta + \frac{(-1)^k a_{k+1}}{g(n)\ldots g(n-k)} + O([\frac{1}{g(n)}]^{k+2}) ,$$

where the coefficients $x_0, \ldots, x_k$ depend only upon g.

If this result is applied to estimation of N when the initial estimator $\hat{N} = \hat{\theta}_t$ is a linear combination of the frequencies then the resulting jackknifed estimators are again linear combinations of the frequencies. Explicit results are computed for the initial estimator $\hat{N} = S_t$, assuming $E(S_t) = N + \frac{a_1}{t} + \frac{a_2}{t^2} + \ldots$, for $k = 1, \ldots, 5$. A procedure is then given for selecting one of the $\hat{N}_{Jk}$ as the estimator to use in any given livetrapping study: The hypotheses $H_{0i}: E(\hat{N}_{Ji+1} - \hat{N}_{Ji}) = 0$ are tested and $\hat{N}_J = \hat{N}_{Jk}$ is chosen where $\hat{N}_{Jk}$ is the first estimator such that $H_{01}, \ldots, H_{0k-1}$ are rejected and H_{0k} is not rejected.

The estimator $\hat{N}_J$, i.e., this estimation procedure, is evaluated by applying it to some simulated livetrapping data. It is shown that this procedure generally arrives at a choice of $\hat{N}_J$ for $k = 1, 2,$ or 3. This estimator is quite robust to variations in

capture probabilities among individuals in the sense that the absolute value of the bias of $\hat{N}_J$ is not generally large. This is quite different from estimators such as the Schnabel which is unbiased if capture probabilities do not vary, but which may have a large negative bias when capture probabilities vary among individuals.

A further indication of robustness for $\hat{N}_J$ is derived in Section 7.2 by considering an extension of the model wherein external influences are allowed to cause slight variations in capture probabilities over days. If $V(p)$ dominates these day-to-day variations then the capture frequencies still have an approximate multinomial distribution. It follows that if the hypothesis $p_{ji} = p_j$, $j = 1, \ldots, S_t$, $i = 1, \ldots, t$ is tested and rejected, as discussed in Section 7.1, it does not automatically mean $\hat{N}_J$ is not a suitable estimator of population size.

BIBLIOGRAPHY

Adams, J.E., Gray, H.L., and Watkins, T.A. (1971). An asymptotic characterization of bias reduction by jackknifing. Ann. Math. Statist. 42, 1606-1612.

Arvesen, J.N. (1969). Jackknifing U-statistics. Ann. Math. Statist. 40, 2076-2100.

Blischke, W.R. (1963). Mixtures of discrete distributions. Proceedings of the International Symposium on Classical and Contagious Discrete Distributions, Statistical Publishing Society, Calcutta, 351-372.

Brillinger, D.R. (1964). The asymptotic behavior of Tukey's general method of setting approximate confidence limits (the jackknife) when applied to maximum likelihood estimates. Rev. Int. Statist. Inst. 32, 202-206.

Burnham, K.P. and Overton, W.S. (1969). A simulation study of livetrapping and estimation of population size. Technical Report No. 14, Department of Statistics, Oregon State University.

Chatfield, C. and Goodhardt, G.J. (1970). The beta-binomial model for consumer purchasing behavior. J. R. Statist. Soc. C 19, 240-250.

Cormack, R.M. (1966). A test for equal catchability. Biometrics 22, 330-342.

Cormack, R.M. (1968). The statistics of capture-recapture methods. Oceanogr. Mar. Biol. Ann. Rev. 6, 455-506.

Demming, W.E. (1943). Statistical Adjustment of Data, John Wiley and Sons, Inc., New York.

Durbin, J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. Biometrika 46, 477-480.

Eberhardt, L.L. (1969). Population estimates for recapture frequencies. J. Wildl. Mgmt. 33, 28-39.

Eberhardt, L., Peterle, T.J. and Schofield, R. (1963). Problems in a rabbit population study. Wildl. Monogr. 10.

Edwards, W.R. and Eberhardt, L.L. (1967). Estimating cottontail abundance from livetrapping data. J. Wildl. Mgmt. 31, 87-96.

Feldman, D. and Fox, M. (1968). Estimation of the parameter n in the binomial distribution. J. Amer. Statist. Ass. 63, 150-158.

Fraser, D.A.S. (1957). Nonparametric Methods in Statistics, John Wiley and Sons, Inc., New York.

Holgate, P. (1966). Contributions to the mathematics of animal trapping. Biometrics 22, 925-936.

Johnson, N.L. and Kotz, S. (1969). Discrete Distributions, Houghton Mifflin Co., Boston.

Johnson, N.L. and Kotz, S. (1970). Continuous Univariate Distributions - 2, Houghton Mifflin Co., Boston.

Jolley, L.S.W. (1961). Summation of Series, Dover Publications, Inc. New York.

Kale, B.K. (1961). On the solution of the likelihood equation by iteration processes. Biometrika 48, 452-456.

Kale, B.K. (1962). On the solution of the likelihood equations by iteration processes. The multiparametric case. Biometrika 49, 479-486.

Lewontin, R.C. and Prout, T. (1956). Estimation of the number of different classes in a population. Biometrics 12, 211-223.

Mantel, N. (1967). Assumption-free estimators using U-statistics and a relationship to the jackknife method. Biometrics 23, 567-571.

Miller, R.G., Jr. (1964). A trustworthy jackknife. Ann. Math. Statist. 35, 1594-1605.

Miller, R.G., Jr. (1968). Jackknifing variances. Ann. Math. Statist. 39, 567-582.

Nixon, C.M., Edwards, W.R., and Eberhardt, L.L. (1967). Estimating squirrel abundance from livetrapping data. J. Wildl. Mgmt. 31, 96-101.

Perlis, S. (1952). Theory of Matrices, Addison-Wesley, Reading, Massachusetts.

Quenouille, M.H. (1949). Approximate tests of correlation in time-series. J.R. Statist. Soc. B 11, 68-84.

Quenouille, M H. (1956). Notes on bias in estimation. Biometrika 43, 353-560.

Rao, C.R. (1958). Maximum likelihood estimation for the multinomial distribution. Sankhyā 18, 139-148.

Rao, C.R. (1965). Linear Statistical Inference and its Applications, John Wiley and Sons, Inc., New York.

Rao, J.N.K. (1965). A note on estimation of ratios by Quenouille's method. Biometrika 52, 647-649.

Rao, J.N.K. and Webster, J.T. (1966). On two methods of bias reduction in the estimation of ratios. Biometrika 53, 571-577.

Robson, D.S. (1971). Personal communication by letter dated September 8, 1971.

Robson, D.S. and Whitlock, J.H. (1964). Estimation of a truncation point. Biometrika 51, 33-39.

Schnabel, Z.E. (1938). The estimation of the total fish population of a lake. Amer. Math. Monthly 45, 348-352.

Schucany, W.R., Gray, H.L. and Owen, D.B. (1971). Bias reduction in estimation. J. Amer. Statist. Ass. 66, 524-533.

Shenton, L.R. (1950). Maximum likelihood and the efficiency of the method of moments. Biometrika 37, 111-116.

Skellam, J.G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between sets of trails. J.R. Statist. Soc. B 10, 257-261.

APPENDIX

APPENDIX

The Efficiency of the Method of Moments, and Method of Mean and Zeros for the Beta-Binomial Distribution

Let P_θ be a probability distribution for which the information matrix $I(\underline{\theta})$ exists, where $\underline{\theta} = (\theta_1, \ldots, \theta_r)'$. Let $\hat{\underline{\theta}}$ be any estimator of $\underline{\theta}$ based on a random sample of size N; and let $D(\hat{\underline{\theta}})$ be the dispersion matrix of $\hat{\underline{\theta}}$. The efficiency of $\hat{\underline{\theta}}$ is defined as (Shenton, 1950):

$$E = \frac{|I^{-1}(\underline{\theta})|}{|D(\hat{\underline{\theta}})|} .$$

Now assume $X_1, \ldots, X_r$ are statistics, computed from the sample, such that $E(X_i) = g_i(\underline{\theta})$, $i = 1, \ldots, r$. The method of expectations estimator of $\underline{\theta}$ is defined as the solution to the equations

$$\begin{bmatrix} X_1 \\ \vdots \\ X_r \end{bmatrix} = \begin{bmatrix} g_1(\underline{\theta}) \\ \vdots \\ g_r(\underline{\theta}) \end{bmatrix} = \underline{g}(\underline{\theta}) .$$

Assuming the inverse function exists, $\hat{\underline{\theta}} = \underline{g}^{-1}(X_1, \ldots, X_r)$. The dispersion matrix of $\hat{\underline{\theta}}$ may be approximated by

$$D(\hat{\underline{\theta}}) \doteq \Lambda D(X_1, \ldots, X_r)\Lambda' ,$$

where $D(X_1, \ldots, X_r)$ is the dispersion matrix of $X_1, \ldots, X_r$, and

Λ is the Jacobian matrix,

$$\Lambda = \left. \frac{\partial \underline{g}^{-1}(x_1, \ldots, X_r)}{\partial(X_1, \ldots, X_r)} \right|_{\substack{X_i = E(X_i), \\ i = 1, \ldots, r}}$$

The efficiency of $\hat{\underline{\theta}}$ may be determined from

$$\frac{1}{E} = |I(\underline{\theta})| \, |D(X_1, \ldots, X_r)| \, |\Lambda|^2 .$$

If $\underline{g}^{-1}$ exists but can not be found explicitly, it is still possible to find the Jacobian $|\Lambda|$ implicitly as

$$|\Lambda| = \left| \frac{\partial \underline{g}(\theta_1, \ldots, \theta_r)}{\partial(\theta_1, \ldots, \theta_r)} \right|^{-1} .$$

These definitions and formulae will be used with the Beta-binomial distribution making the identification $\underline{\theta} = (\alpha, \beta)'$, $r = 2$. If Y is a Beta-binomial random variable, then for $i = 0, 1, \ldots, t$ $P\{Y = i | \alpha, \beta\} = \pi_i(\alpha, \beta)$ where π_i is given by (2.4.1). In keeping with the derivation of this distribution as a mixing of a binomial and Beta distribution let $E(p) = \frac{\alpha}{\alpha + \beta}$, then $E(Y) = tE(p)$.

The method of moments estimators of α and β are found as the solution to the equations $m_1 = E(m_1)$ and $m_2 = E(m_2)$ where $m_1 = C_t/N$ and $m_2 = \sum_{i=1}^{t} (i)^2 f_{it}/N$. The expectations of

these statistics are

$$E(m_1) = t\frac{\alpha}{\alpha+\beta},$$

$$E(m_2) = E(m_1)\left[\frac{\beta+t\alpha+t}{\alpha+\beta+1}\right].$$

In the extreme cases $f_{0t} = N$ or $f_{tt} = N$ the method of moment estimators are essentially arbitrary. However, the ML estimators are also arbitrary in these cases. When these extreme cases do not obtain closed form solution for the method of moment estimators exists provided $(m_1/m_2)t + m_1 - m_1 t - t \neq 0$:

$$\alpha^* = \frac{m_1 t - m_2}{\frac{m_2}{m_1}t + m_1 - m_1 t - t},$$

$$\beta^* = z\alpha^*,$$

where $z = (tN/C_t) - 1$ (this result is originally due to Skellam (1948)).

The method of mean and zeros estimators are found as the solution to the equations $f_{0t} = N\pi_0$ and $m_1 = E(m_1)$. In the extreme case $f_{0t} = 0$ these equations have no solution and if $f_{0t} = N$ the solution is arbitrary. For $f_{0t} = 0, \ldots, N-1$, the equations to be solved are

$$\frac{f_{0t}}{N} = \prod_{i=1}^{t} \left(\frac{z\alpha+i-1}{\alpha(z+1)+i-1}\right) = g(\alpha) ,$$

and $\beta = z\alpha$. The derivative of g is

$$\frac{dg(\alpha)}{d\alpha} = -g(\alpha) \sum_{i=1}^{t} \frac{(i-1)}{(\alpha z+i-1)(\alpha(z+1)+i-1)} < 0$$

for all $\alpha > 0$ (because $f_{0t} < N$, it follows $C_t > 0$, thus $0 \leq z < +\infty$). Thus, $g(\alpha)$ is strictly monotone decreasing. By continuity, $g(0) = 1$ and $g(+\infty) = (z/(z+1))^t = (1-C_t/tN)^t > 0$. It follows that a unique solution exists to the equation $f_{0t}/N = g(\alpha)$ iff $f_{0t}/N > (1-C_t/tN)^t$. When a solution exists it can easily be found because of the monotonicity of $g(\alpha)$.

The efficiency of the method of moments for the Beta-binomial distribution has been examined by Shenton (1950). The efficiency of the method of mean and zeros for the Beta-binomial distribution does not appear to have been investigated. Formulae for both efficiencies will be given below. Shenton's formula is not used here because it is an approximate derived before computers were readily available.

When the efficiency of these two estimation methods is computed the sample size, N, drops out of the final result. Without loss of generality then let N be set equal to one. For the method of moments the elements of $D(m_1, m_2)$ are given by

$$V(m_1) = \sum_{i=1}^{t} (i)^2 \pi_i - \left[\sum_{i=1}^{t} i\pi_i\right]^2 ,$$

$$V(m_2) = \sum_{i=1}^{t} (i)^4 \pi_i - \left[\sum_{i=1}^{t} (i)^2 \pi_i\right]^2 ,$$

$$Cov(m_1, m_2) = \sum_{i=1}^{t} (i)^3 \pi_i - \left[\sum_{i=1}^{t} i\pi_i\right]\left[\sum_{i=1}^{t} (i)^2 \pi_i\right] .$$

The Jacobian $|\Lambda|$ for this transformation is conveniently computed implicitly:

$$\frac{1}{|\Lambda|} = \begin{vmatrix} \sum_{i=1}^{t} (i) \frac{\partial \pi_i}{\partial \alpha} & \sum_{i=1}^{t} (i) \frac{\partial \pi_i}{\partial \beta} \\ \sum_{i=1}^{t} (i)^2 \frac{\partial \pi_i}{\partial \alpha} & \sum_{i=1}^{t} (i)^2 \frac{\partial \pi_i}{\partial \beta} \end{vmatrix} .$$

Formulae for $I(\alpha, \beta)$, $\partial \pi_i / \partial \alpha$ and $\partial \pi_i / \partial \beta$ are given in Section 4.1.

A program was written by the author to compute the efficiency of the method of moments for any values of α and β. The efficiency of the method of mean and zeros is computed simultaneously using the formulae

$$D(f_{0t}, m_1) = \begin{bmatrix} (1-\pi_0)\pi_0 & -\pi_0 E(p) \\ -\pi_0 E(p) & \frac{E(p)(1-E(p))(\alpha+\beta+t)}{t(\alpha+\beta+1)} \end{bmatrix},$$

and

$$\frac{1}{|\Lambda|} = \left| \begin{bmatrix} \frac{\partial \pi_0}{\partial \alpha} & \frac{\partial \pi_0}{\partial \beta} \\ \frac{\beta}{(\alpha+\beta)^2} & \frac{-\alpha}{(\alpha+\beta)^2} \end{bmatrix} \right|.$$

It is required that $t \geq 2$ holds in order to estimate both α and β. For $t = 2$ these two estimators and the ML estimator are all equivalent. Thus for $t = 2$ the efficiencies are 1. Table A1 gives the efficiencies of the method of mean and zeros (M+Z) and method of moments (MM) for a variety of values of α and β with $t = 5(5)30$. Also shown in this table is π_0 for each value of t. For the range of α, β and t examined it is evident that the method of moments estimators are efficient enough to use as starting values in an iterative solution for the exact ML estimators. In fact, the method of moments shows good efficiency relative to the method of mean and zeros in all but a few extreme cases (e.g., $\alpha = .1$ and $\beta = 9.9$).

Table A1. Efficiency of the method of mean and zeros (M+Z) and method of moments (MM) for selected Beta-binomial distributions. Also shown is the value of $\pi_0(\alpha, \beta)$.

t	M+Z	MM	π_0	t	M+Z	MM	π_0	t	M+Z	MM	π_0
$\alpha = 10.$ $\beta = 10.$ $E(p) = .5$				$\alpha = 10.$ $\beta = 31.67$ $E(p) = .24$				$\alpha = 10.$ $\beta = 73.33$ $E(p) = .12$			
5	.445	1.000	.047	5	.761	.998	.272	5	.890	.998	.536
10	.113	1.000	.005	10	.469	.994	.086	10	.726	.994	.298
15	.031	.999	.001	15	.282	.991	.031	15	.587	.991	.171
20	.009	.999	.000	20	.168	.989	.012	20	.470	.988	.101
25	.003	.998	.000	25	.100	.987	.005	25	.372	.985	.061
30	.001	.998	.000	30	.060	.985	.002	30	.297	.982	.038
$\alpha = 1.0$ $\beta = 1.0$ $E(p) = .5$				$\alpha = 1.0$ $\beta = 3.167$ $E(p) = .24$				$\alpha = 1.0$ $\beta = 7.333$ $E(p) = .12$			
5	.657	.979	.167	5	.847	.967	.388	5	.935	.952	.595
10	.478	.932	.091	10	.722	.923	.241	10	.862	.894	.423
15	.398	.899	.063	15	.650	.891	.174	15	.811	.854	.328
20	.348	.876	.048	20	.599	.868	.137	20	.770	.823	.268
25	.314	.858	.038	25	.561	.851	.112	25	.736	.800	.227
30	.288	.845	.032	30	.529	.837	.095	30	.708	.781	.196
$\alpha = .1$ $\beta = .10$ $E(p) = .5$				$\alpha = .1$ $\beta = .3167$ $E(p) = .24$				$\alpha = .1$ $\beta = .7333$ $E(p) = .12$			
5	.771	.939	.414	5	.811	.935	.642	5	.872	.928	.766
10	.667	.832	.385	10	.721	.827	.599	10	.807	.823	.716
15	.622	.765	.369	15	.681	.760	.575	15	.778	.757	.688
20	.596	.718	.358	20	.658	.714	.558	20	.762	.712	.669
25	.577	.684	.350	25	.643	.681	.546	25	.751	.678	.655
30	.564	.658	.344	30	.631	.655	.536	30	.743	.653	.643

Table A1. Continued.

t	M+Z	MM	π_0	t	M+Z	MM	π_0	t	M+Z	MM	π_0
α = 10.	β = 156.7	E(p) = .06		α = 10.	β = 990.0	E(p) = .01		α = 10.	β = 9990.	E(p) = .001	
5	.947	.998	.737	5	.992	1.000	.951	5	.999	1.000	.995
10	.864	.996	.548	10	.978	.999	.905	10	.998	1.000	.990
15	.785	.994	.411	15	.964	.999	.861	15	.996	1.000	.985
20	.712	.991	.310	20	.950	.998	.819	20	.995	1.000	.980
25	.644	.989	.236	25	.937	.998	.780	25	.994	1.000	.975
30	.582	.987	.181	30	.923	.997	.743	30	.992	1.000	.970
α = 1.0	β = 15.67	E(p) = .06		α = 1.0	β = 99.0	E(p) = .01		α = 1.0	β = 999.0	E(p) = .001	
5	.973	.957	.758	5	.996	.988	.952	5	1.000	.999	.995
10	.935	.902	.610	10	.991	.969	.908	10	.999	.997	.990
15	.904	.861	.511	15	.985	.952	.868	15	.999	.994	.985
20	.877	.828	.439	20	.980	.936	.832	20	.998	.992	.980
25	.853	.801	.385	25	.974	.921	.798	25	.997	.990	.976
30	.832	.779	.343	30	.969	.907	.767	30	.997	.988	.971
α = .1	β = 1.567	E(p) = .06		α = .1	β = 9.9	E(p) = .01		α = .1	β = 99.9	E(p) = .001	
5	.937	.911	.846	5	.997	.907	.958	5	1.000	.983	.995
10	.901	.803	.797	10	.994	.800	.930	10	1.000	.955	.990
15	.885	.735	.768	15	.992	.726	.909	15	1.000	.930	.986
20	.876	.688	.748	20	.991	.671	.893	20	1.000	.907	.982
25	.871	.653	.732	25	.989	.629	.879	25	1.000	.885	.978
30	.867	.626	.719	30	.989	.595	.867	30	.999	.865	.974