CONDITIONALLY UNBIASED ESTIMATION OF POPULATION SIZE UNDER PLANT-CAPTURE

J. Ashbridge and I. B. J. Goudie*
School of Mathematics and Statistics,
University of St Andrews, St Andrews, Scotland

ABSTRACT

A target population is increased by the insertion of a known number of planted individuals. The standard equal catchability model used in mark-recapture is assumed to be applicable to the augmented population. This paper considers a conditionally unbiased estimator of the unknown size of the target population. Its performance is seen to compare favourably with that of the maximum likelihood estimator and a Petersen-type estimator. Estimators of the standard deviation of the conditionally unbiased estimator are also considered.

Key words: capture–recapture; factorial series distribution; Gould–Hopper number; Petersen estimator.

*Correspondence: I. B. J. Goudie, School of Mathematics and Statistics, University of St Andrews, St Andrews, KY16 9SS, Scotland; E-mail: ig@st-andrews.ac.uk.

1. INTRODUCTION

Each member of a planted population, consisting of a known number $R \geq 0$ of individuals, is given a unique tag before being added to a target population of unknown size $N \geq 0$. This augmented population, which is assumed to be closed, is sampled on t occasions. Whenever a member of the target population is caught for the first time it is similarly tagged to distinguish it from other members of the augmented population. In each of the t samples, any particular member of the augmented population, independently of other individuals and of its previous capture history, is captured with a constant probability p, where $0 < p < 1$.

One factor motivating interest in plant-capture methods has been the need for a variant of mark-recapture methodology that could be based on a single sample (Laska et al., 1988). The approach was deployed in five cities by the U.S. Census Bureau during the 1990 US decennial census to estimate the number of homeless living on the streets (Martin et al., 1997, Laska & Meisner, 1993). Application of plant-capture for this purpose has continued to be recommended (U.S. Department of Housing and Urban Development, 2006). In particular, it is used by the Homeless Outreach Population Estimate (HOPE) survey in New York City, which is now conducted annually.

The use of plant-capture methodology has been also been proposed in other very different contexts. Skalski and Robson (1982) and Yip and Fong (1993) have considered the estimation of animal abundance by introducing pre-marked animals into a population prior to a removal experiment. Assuming that seeded errors had been inserted into a software system, Duran and Wiorkowski (1981) considered fixed sample size and sequential models for estimating the number of indigenous errors already present. For a broad class of discrete-time plant-capture models, Norris and Pollock (2001) provided a likelihood ratio test of the key assumption that the behaviour of the plants matches that of the target population.

Plant-capture estimation under continuous-time sampling has been addressed by Yip (1996), Goudie, Pollock and Ashbridge (1998) and Goudie and Ashbridge (2000). Yip (1995) and Yip, Xi, Fong and Hayakawa (1999) consider the application of plant-capture to software reliability in a continuous-time context.

Under the above assumptions, the target and planted populations act independently, and, apart from the pre-marking of the planted individuals, the standard mark-recapture model M_0 (Otis et al., 1978) describes the behaviour of each population. A conditionally unbiased estimator (CUE) for estimating population size under model M_0 was recommended by Goudie and Ashbridge (2005). In this paper we consider the more general form of this estimator that is appropriate for estimation of N under the present plant-capture model.

Following a review in section 2 below of the relevant distributional results, we define, in section 3, the CUE of the target population size N. In section 4, further properties and the evaluation of this estimator are discussed. In sections 6 and 7 respectively, we describe two other possible estimators of N and present exact computational results comparing the three estimators. One of the alternative estimators is of Petersen-type, and the other is the maximum likelihood estimator (MLE). The existence of the latter was established by Goudie et al. (2007), who also provided a derivation of its asymptotic distribution.

In section 5 below, we consider the relationship between the CUE for the discrete-time model in this paper and that for the corresponding continuous-time model. We investigate in section 8 the extent to which estimation is assisted by the use of planted individuals. In section 9, we discuss two possible estimators of the standard deviation of the CUE.

2. THE SUFFICIENT STATISTICS AND THEIR DISTRIBUTION

We may assume that the individuals in the target population are numbered $i = 1, \ldots, N$ and those in the planted population $i = N+1, \ldots, N+R$. The data from the t samples can then be displayed in a $(N+R) \times t$ matrix $D = (d_{ij})$ such that $d_{ij} = 1$ if individual i is captured in sample j, and $d_{ij} = 0$ otherwise. We let $Z = D_{11} + \ldots + D_{N+R,t}$ denote the total number of captures achieved during the sampling, and X the total number of different individuals from the target population that have been captured.

Let $\mathrm{Bin}(n, p)$, where n is a positive integer, denote the distribution with probability function $\{(n)_v/v!\}\, p^v(1-p)^{n-v}$, for $v = 0, \ldots, n$, where $(n)_0 = 1$ and $(n)_v = n(n-1)\ldots(n-v+1)$ when v is a positive integer. The values of the elements of D can clearly be regarded as the values of $Nt + Rt$ independent Bernoulli random variables, and hence Z has the binomial distribution $\mathrm{Bin}(Nt+Rt, p)$. Goudie et al. (2007) showed that (X, Z) is sufficient for (N, p), and that the conditional probability function of X given Z, for $x = 0, \ldots, \min(N, z)$, is

$$p(x|z) = (N)_x G(z, x, t, Rt)\{(Nt+Rt)_z\}^{-1}, \qquad (1)$$

in which $G(z, x, t, b)$ is defined by

$$G(z, x, t, b) = \frac{z!}{x!} \sum_{k=0}^{x} (-1)^k \binom{x}{k} \binom{b + xt - kt}{z} = \frac{1}{x!} \left[\Delta^x (b + wt)_z\right]_{w=0},$$

where Δ denotes the forward finite difference operator. The function $G(z, x, t, b)$ is a Gould–Hopper number (Gould & Hopper, 1962) or non-central C-number.

Noting that $G(z, x, t, Rt) = 0$ for $z > xt + Rt$, it follows from equation (1) that the joint probability function of Z and X is given by

$$p(z, x) = (N)_x G(z, x, t, Rt)\, p^z (1 - p)^{Nt+Rt-z}/z!, \tag{2}$$

for $x = 0, \ldots, N$ and $z = x, \ldots, xt+Rt$. This result subsumes the case $R = 0$, corresponding to the standard mark-recapture model M_0 in which no plants are used. In this case, the number $G(z, x, t, 0)$ in the joint probability function (2) is simply a (central) C-number.

The Gould–Hopper numbers can be relatively difficult to compute, except when their arguments are small. To avoid this problem, a recursive approach can be used when evaluating the joint probability function (2). Explicitly

$$p(z, x) = p\{(xt + Rt - z + 1)p(z - 1, x) + (N - x + 1)\, t\, p(z - 1, x - 1)\}/\{z(1 - p)\}, \tag{3}$$

subject to the initial conditions that $p(z, 0) = (Rt)_z p^z (1 - p)^{Nt+Rt-z}/z!$ for $z = 0, \ldots, Rt$ and $p(x, x) = (N)_x (t\,p)^x (1 - p)^{Nt+Rt-x}/x!$ for $x = 1, \ldots, N$. Setting $R = 0$ in equation (3) provides the correct form for the corresponding equation for model M_0: the expression given by Goudie and Ashbridge (2005) contained a typographical error. Equation (3) can be proved using the triangular recurrence relation for the Gould–Hopper numbers (Charalambides & Koutras, 1983), which states that, for $z = 0, 1 \ldots$ and $x = 1, 2 \ldots$,

$$G(z + 1, x, t, b) = (xt + b - z)G(z, x, t, b) + t\, G(z, x - 1, t, b), \tag{4}$$

where $G(z, 0, t, b) = (b)_z$ for $z = 0, 1 \ldots$, and $G(0, x, t, b) = 0$ for $x = 1, 2 \ldots$.

3. CONDITIONALLY UNBIASED ESTIMATION OF THE POPULATION SIZE

The CUE for model M_0 recommended by Goudie and Ashbridge (2005) was closely related to an estimator proposed by Charalambides (1981) for a sampling problem. It differed trivially in being rounded to the nearest integer value, and, more substantively, in the nature of one of its arguments. The constant pre-chosen sample size in Charalambides' sampling problem was replaced in model M_0 by the value of a random variable giving the total number of captures. In the same paper, Charalambides also presented an unbiased estimator of N for the problem of sampling from a population of t individuals of each of N types and a control class of a further Rt individuals. The estimator that we are proposing in this paper for the plant-capture model is the corresponding generalisation of the CUE for model M_0.

Consider initially the estimator defined by $\tilde{N}_U(z, 0) = 0$ for $z = 0, \ldots, Rt$ and by

$$\tilde{N}_U(z, x) = x + G(z, x-1, t, Rt)/G(z, x, t, Rt) \quad x = 1, \ldots, N;\ z = x, \ldots, xt + Rt. \tag{5}$$

Note that, for $R = 0$, the Gould–Hopper numbers in expression (5) reduce to C-numbers, and hence (5), when rounded to the nearest integer, is the CUE for model M_0 given in Goudie and Ashbridge (2005).

The conditional distribution (1) has the form of a factorial series distribution (Berg, 1974), with the series function being given by $A(N) = (Nt + Rt)_z$. Provided $N \leq z$, it follows from Berg's results, or from Charalambides (1981), that $\tilde{N}_U(z, x)$ is the unique unbiased estimator of N with respect to the conditional distribution of the data given $Z = z$. As is the case with model M_0, however, the condition $N \leq z$ holds with probability less than unity, and hence this conditional unbiasedness does not imply that $\tilde{N}_U$ is unconditionally unbiased.

As the true population size N is integer valued, we round the value of $\tilde{N}_U$ and define the CUE by $\hat{N}_U(z, x) = [\tilde{N}_U(z, x) + 0.5]$, where the square brackets denote the integer part.

4. FURTHER PROPERTIES OF THE ESTIMATOR

An interesting special case of the estimator $\tilde{N}_U(z,x)$ occurs for the case $t=1$, for which the Gould–Hopper numbers can be found by a simpler combinatorial argument. For general t, consider the $(R+x)\times t$ submatrix $\tilde{D}$ of D corresponding to the R planted individuals and a particular set of x target individuals. It follows from the inclusion-exclusion principle (c.f. Johnson et al., 2005, p. 432) that $(x!/z!)G(z,x,t,Rt)$ is the number of possible submatrices $\tilde{D}$ which have exactly z elements equal to unity, with these being arranged so that at least one such element occurs on each of the x rows corresponding to the target individuals. In the case $t=1$, this number is obviously just $(R)_{z-x}/(z-x)!$, and thus $G(z,x,1,R)=(R)_{z-x}(z)_x/x!$. Hence, for $t=1$, equation (5) reduces to $\tilde{N}_U(z,x)=(R+1)x/(z-x+1)$. The implied estimator of the size $N+R$ of the augmented population equals $\{(R+1)(z+1)/(z-x+1)\}-1$, which is a version of the Petersen estimate of $N+R$. More precisely, it is the variant, due to Chapman (1951), of the estimate of the size of a population known to contain R marked individuals, if a sample of size z containing x unmarked individuals is drawn from it.

Charalambides (1981) gave the estimator $\tilde{N}_U(z,x)$ in an ostensibly different form. In our notation, he expressed the estimator, for $x=1,\ldots,N;\ z=x,\ldots,xt+Rt$, in the form

$$\tilde{N}_U(z,x)=(z/t)-R+[G(z+1,x,t,Rt)/\{t\,G(z,x,t,Rt)\}]. \tag{6}$$

The equivalence of the two forms is easily verified using the recurrence relation (4).

To avoid computation of Gould–Hopper numbers, a recursive approach can also be helpful for evaluating the estimator $\tilde{N}_U$. By generalising Property 5 of Berg (1975), a recursion of triangular form can be established. Explicitly, for $x=1,\ldots,N;\ z=x+1,\ldots,(R+x-1)t$, an expression for $\tilde{N}_U(z,x)$ is given by the equation

$$\{\tilde{N}_U(z,x)-x\}\{t\tilde{N}_U(z-1,x)+Rt-z+1\}=\{\tilde{N}_U(z-1,x)-x\}\{t\tilde{N}_U(z-1,x-1)+Rt-z+1\}.$$

When $x\geq 2$, this can be verified by using equations (5) and (6) to express each factor in terms of Gould–Hopper numbers. When $R > 0$, the recursion is also needed when $x = 1$, where it also holds since, for $z \leq Rt$, $G(z, 0, t, Rt)/G(z-1, 0, t, Rt) = Rt - z + 1$ and $\tilde{N}_U(z, 0) = 0$.

To start the recursion, one boundary condition is given by

$$\tilde{N}_U(z, x) = \begin{cases} x & x = 1, \ldots, N; \quad z = (R + x - 1)t + 1, \ldots, (R + x)t; \\ 0 & x = 0; \quad z = 0, \ldots, Rt. \end{cases}$$

The first of these equations follows from (5) as $G(z, x-1, t, Rt) = 0$ for $z > (R + x - 1)t$. To derive a second boundary condition, note that, as $G(z, x, t, Rt) = 0$ for $z < x$, the triangular recurrence relation (4) implies that $G(z, z, t, Rt) = t^z$ for $z = 1, 2 \ldots$. Taking the expressions for $\tilde{N}_U(z, z)$ and $\tilde{N}_U(z-1, z-1)$ provided by (5) and (6) respectively gives a further recursion

$$\tilde{N}_U(z, z) - \tilde{N}_U(z-1, z-1) = \{Rt + zt - z + 1\}/t \qquad z = 1, 2 \ldots,$$

with $\tilde{N}_U(0, 0) = 0$. This then yields the second boundary condition

$$\tilde{N}_U(z, z) = z(2Rt + tz + t - z + 1)/(2t) \qquad z = 0, 1 \ldots.$$

5. THE CORRESPONDING CONTINUOUS-TIME PROBLEM

A continuous-time analogue of this plant-capture model was considered by Goudie and Ashbridge (2000). Under this process, sightings of any member of the augmented population occurred according to a Poisson process, which could be time-inhomogeneous. The $N + R$ Poisson processes were assumed to be independent and were observed for a pre-determined time. For the case where all individuals have a common intensity function, both the CUE and the relevant distributional results for the continuous-time process can be obtained as limits of the corresponding quantities in the present model.

First note that, as $t \to \infty$, we have $G(z, x, t, Rt)/t^z$ tends to the non-central Stirling number

$S(z,x,R)$ of the second kind, defined by

$$S(z,x,R) = \frac{1}{x!}\sum_{k=0}^{x}(-1)^k\binom{x}{k}(x-k+R)^z\,.$$

It then follows that the limit as $t \to \infty$ of the probability function (1) of the conditional distribution for X given Z is

$$p(x|z) = (N)_x(N+R)^{-z}S(z,x,R) \qquad\qquad x = 0,\,\dots\,,\min(N,z).$$

This is the corresponding conditional distribution in the continuous-time case. Again using the limit of $G(z,x,t,Rt)/t^z$, the limit of the estimator $\tilde{N}_U$, as $t \to \infty$, is given by

$$\tilde{N}_U(z,x) = x + \{S(z,x-1,R)/S(z,x,R)\} \qquad x = 1,\,\dots\,,N; z = x, x+1,\,\dots\,,$$

with $\tilde{N}_U(z,0) = 0$ for $z = 0,1,\,\dots\,$. When rounded to the nearest integer, this is the corresponding CUE in continuous time (Goudie & Ashbridge, 2000). Note that the limiting operation used in this section is equivalent to the one in the sampling problem of Charalambides (1981) in which the number of individuals of each type tends to infinity in such a way that the ratio of the size of the control class to this number tends to a constant.

6. OTHER ESTIMATORS OF THE TARGET POPULATION SIZE

An alternative estimator of N is the (unconditional) MLE $\hat{N}_M$ based on the joint probability function (2) of Z and X. If $h(k) = -k\log k$, Goudie et al. (2007) show that $\hat{N}_M$ is the smallest integer k in the set $\{x, x+1, \ldots\}$ for which

$$\Delta h(kt+Rt) - \Delta h(kt+Rt-z) - \log\{(k-x+1)/(k+1)\} < 0.$$

For the case $R > 0$, let Y denote the number of distinct members of the planted population that are captured during the sampling process. It is then evident that X and Y are independent, having $\text{Bin}(N,\,p^*)$ and $\text{Bin}(R,\,p^*)$ distributions respectively, where $p^* = 1-(1-p)^t$. It follows that $W = X+Y$ has a $\text{Bin}(N+R,\,p^*)$ distribution, and that the conditional distribution of X given W is hypergeometric, with probability function

$$p(x|w) = \binom{N}{x}\binom{R}{y}\Big/\binom{N+R}{w} \qquad\qquad x = \max(0, w-R),\ \ldots\ , \min(N, w).$$

Taking the resulting conditional likelihood function, and equating its values at $N-1$ and N, shows that an approximate conditional MLE of N is given by the Petersen-type estimator Rx/y. To circumvent the problem that this estimator is undefined when $y = 0$, we chose, after some experimentation, the modified form $(R+1)x/(y+1)$. In fact, we will round this modified estimator to the nearest integer, and consider

$$\hat{N}_P(x, y) = [0.5 + \{(R+1)x/(y+1)\}].$$

7. THE COMPARATIVE MERITS OF THE ESTIMATORS

Here, assuming $R > 0$, we compare the performance of the CUE $\hat{N}_U$ with that of the Petersen-type estimator $\hat{N}_P$ and the MLE $\hat{N}_M$. The performance of each estimator is determined by exact calculation of the statistics m and s which are respectively the mean and standard deviation conditional on the event $E \equiv \{Z > X\}$. This conditioning enables a precise comparison to be made. Although it is not required for the estimators $\hat{N}_P$ and $\hat{N}_U$, which are finite with probability one, it is needed for $\hat{N}_M$ if the statistics m and s are to be finite. Results are presented in Table 1 for situations in which $t = 5$, 10, 15 or 20 samples are taken from a population of size $N = 50$, using $R = 5$ or 25 plants when the capture probability $p = 0.05$, 0.1 or 0.2. All the moments of estimators that are given in this paper are conditional on the event E, but, for brevity and to avoid confusion with the more important conditioning on Z, from here on we omit the word 'conditional' on each occasion. For ease of comparison, corresponding results for the case $R = 0$ are also shown in Table 1, but will not be discussed until the next section.

The conditional distributions, given the event E, of the MLE $\hat{N}_M$ and the CUE $\hat{N}_U$ are easy to obtain using the joint probability function (2) of Z and X. For the Petersen-type estimator $\hat{N}_P$, however, we need to obtain the joint conditional distribution of X and Y given E. Firstly, note that the complementary event $\bar{E} \equiv \{Z = X\}$ occurs if and only if both $Y = 0$ and each animal in the target population is seen at most once. It follows that

$$P(\bar{E}) = (1-p)^{Rt}\{(1-p)^t + tp(1-p)^{t-1}\}^N.$$

Now, given that a particular animal is captured, the conditional distribution of the number of times it is captured has a zero-truncated binomial distribution. In particular, the conditional probability of a single capture is $\phi = tp(1-p)^{t-1}/p^*$. So, if we denote $P(E|X = x, Y = y)$ by $\tilde{P}(E)$, it follows, by independence, that $\tilde{P}(E) = 1 - \phi^x$ when $y = 0$. It is clear that $\tilde{P}(E) = 1$ when $y > 0$. In view of the independence of X and Y, the joint conditional distribution of these variables can then be found by Bayes Theorem, since

$$P(X = x, Y = y|E) = P(X = x)P(Y = y)\tilde{P}(E)/P(E),$$

where, as noted above, X and Y have the binomial distributions $\mathrm{Bin}(N, p^*)$ and $\mathrm{Bin}(R, p^*)$ respectively. This joint conditional distribution is defined on the set

$$\{(x, y) : x = 0, \ldots, N;\ y = \max(0, 1-x), \ldots, R\}.$$

Table 1 indicates that the CUE $\hat{N}_U$ should be preferred to the Petersen-type estimator $\hat{N}_P$. Usually the bias of these estimators is negligible, with the only exception being a tendency to negative bias when p, t and R are all small. In most of these latter cases, $\hat{N}_U$ is slightly the better of these two estimators in terms of bias. The standard deviation of $\hat{N}_U$ is consistently smaller than the corresponding standard deviation of $\hat{N}_P$, with the ratio of the latter to the former being particularly large when p is large and R is small.

Compared to the other two estimators, the MLE $\hat{N}_M$ performs poorly in terms of bias. In particular, $\hat{N}_M$ shows a positive bias when p and t are both small, even if R is relatively large. In only one of the 24 cases considered in Table 1 is $\hat{N}_M$ the least biased of the three estimators. Moreover, although the results suggest that the bias of the CUE $\hat{N}_U$ does not increase as t increases, those for $p = 0.1$ show that $\hat{N}_M$ can become more biased as t increases. So $\hat{N}_M$ does not necessarily improve its performance as the amount of information increases. In most of the cases considered in Table 1, the standard deviation of $\hat{N}_U$ is less than that of $\hat{N}_M$, with the ratio of the latter to the former being large when p and t are small. We therefore conclude that, when plants are present, $\hat{N}_U$ should also be preferred to $\hat{N}_M$.

8. THE EFFECT OF THE PLANTED INDIVIDUALS

The CUE $\hat{N}_U$, which we recommend, and the MLE $\hat{N}_M$ may also be deployed under model M_0, which is given by $R = 0$. For each of these estimators, it is therefore of interest to compare the performance when plants are present to that when they are absent.

From Table 1, it can be seen that, under M_0, there is often no problem of bias when using the CUE $\hat{N}_U$. The exceptions occur when t is small and the capture probability p is not large. In these cases, the presence of plants reduces the bias, and most effectively so when the number of plants is relatively large. When $p = 0.05$ and $t = 5$, the improvements are achieved at the price of rather larger standard deviations s, but it should be borne in mind that it is common in population size estimation for negatively biased estimators also to have small standard deviations. When $p = 0.1$, the plants reduce the bias of $\hat{N}_U$ when t is small, and provide a consistent reduction in s for all values of t. When $p = 0.2$, any bias in $\hat{N}_U$ is never more than trivial, but the plants produce a reduction in s when $t = 5$.

For the MLE $\hat{N}_M$, the effect of the plants on the standard deviation is usually beneficial: the only increase in Table 1 occurs for $p = 0.05$ and $t = 5$. In this case, the above remark on the standard deviation of negatively biased estimators is again relevant. The effect of the plants on the mean of $\hat{N}_M$ appears to be more mixed. When $p = 0.1$ or 0.2, the plants reduce the bias of $\hat{N}_M$ when $t = 5$, but, for the longer sampling times, both improvements and small deteriorations in bias occur. When $p = 0.05$, the presence of plants reduces the bias of $\hat{N}_M$, except in the case $t = 5$ where it is slightly worse. It should also be noted, however, that the plants reduce the probability that $\hat{N}_M$ is infinite and thus of no practical use. Of the cases considered in section 7, the greatest reduction in this probability occurs in this case where $p = 0.05$ and $t = 5$. Here the probability of the event E, which is 0.3190 when $R = 0$, falls to 0.0885 when $R = 5$ and to 0.0005 when $R = 25$.

9. ESTIMATING THE STANDARD DEVIATION OF THE PROPOSED ESTIMATOR

The approaches used by Goudie and Ashbridge (2005) to provide estimators of the standard deviation of the CUE for model M_0 can also be applied in the present context. Application of the results of Berg (1974) shows that, for fixed $z \geq N$, there is a unique unbiased estimator of the variance of the raw form $\tilde{N}_U$ of the conditionally unbiased estimator prior to rounding. This unbiased estimator is given by

$$\hat{V}_U(z, x) = \{\tilde{N}_U(z, x) - x\}\{\tilde{N}_U(z, x) - \tilde{N}_U(z, x-1)\}.$$

The square root $\hat{S}_U(z, x)$ of $\hat{V}_U(z, x)$ thus provides a possible estimator of the standard deviation of $\hat{N}_U$. Under the joint distribution of Z and X, the event $\{z \geq N\}$ does not occur with probability one. The estimator $\hat{V}_U(z, x)$ is not therefore unconditionally unbiased. The estimator $\hat{S}_U(z, x)$ is also biased: the extent of the bias, conditional on the event $E = \{Z > X\}$, is examined in Table 2.

Alternatively a parametric bootstrap can be used to provide another estimator $\hat{S}_B(z, x)$ of the standard deviation of $\hat{N}_U$. When the estimate of population size is $\hat{N}_U(z, x)$, realisations are simulated from an augmented population of size $\hat{N}_U(z, x)+R$. Each realisation comprises t samples, in which the constant capture probability is given by $z/\{(\hat{N}_U(z, x) + R)t\}$. When conditioning on the event E is required, realisations for which $Z = X$ can be discarded. For each simulated realisation of the sampling process, the value of the CUE is calculated. The standard deviation $\hat{S}_B(z, x)$ of the estimates from the resamples then provides an estimated standard deviation for the CUE.

In Table 2 we compare these estimators of the standard deviation in each of the situations previously considered in Table 1. The true standard deviation of the CUE $\hat{N}_U$ is again denoted by s, and the expected value of $\hat{S}_U(z,x)$ is given by s_U. The performance of the parametric bootstrap was assessed by randomly selecting 1000 values of (Z,X) under its conditional distribution given E. For each (Z,X), 1000 realisations of the sampling process were generated in the manner described above. The value $\hat{s}_B$ shown in the table is the mean of the resulting 1000 estimates of the standard deviation.

We see from Table 2 that the mean s_U of the estimator $\hat{S}_U(z,x)$ is usually close to the true standard deviation s. Of the situations considered in Table 2, there is, however, only one in which $\hat{S}_U(z,x)$ does not have a negative bias. In contrast, in two-thirds of these situations the bootstrap estimate $\hat{S}_B(z,x)$ is biased positively. Since it is wise not to underestimate the variance of a point estimate, positive bias should usually be regarded as preferable. Moreover, the absolute bias of $\hat{S}_B(z,x)$ is less than that of $\hat{S}_U(z,x)$ in all but one of the cases shown. These arguments would therefore suggest that the bootstrap estimator of the standard deviation of the CUE should be preferred.

10. DISCUSSION

The results presented in section 7 suggest that, for this plant-capture model, the CUE $\hat{N}_U$ should be preferred to the MLE $\hat{N}_M$ and to the Petersen-type estimator $\hat{N}_P$. Indeed, it was seen that both $\hat{N}_U$ and $\hat{N}_P$ are usually less biased than $\hat{N}_M$. The estimator $\hat{N}_U$ also has a tendency to have a smaller standard deviation than $\hat{N}_M$, and is considerably better than $\hat{N}_P$ in this regard. Our recommendation of the CUE mirrors that of Goudie and Ashbridge (2005) for model M_0.

It is also worth noting that our assessment of the performance of $\hat{N}_U$ has been adversely affected by the conditioning on the event $E \equiv \{Z > X\}$ that we used to obtain a precise comparison with $\hat{N}_M$. For instance, Table 1 showed, when $N = 50$, the conditional mean of $\hat{N}_U$, for $p = 0.05$ and $t = 5$, was 40.8 when $R = 5$. Unconditionally, the bias is appreciably reduced, since the corresponding unconditional mean is 46.5. Bearing this in mind, our preference for $\hat{N}_U$ is generally stronger when the information available is weak.

These conclusions imply that it is of most interest to consider the efficacy of the plants under the assumption that $\hat{N}_U$ is being used. Unsurprisingly, the results of section 8 showed that the effect of the plants is most beneficial when t and p are small. In these situations where the information is weak, even a small number of plants can yield useful reductions in bias, with further improvements being achieved as the number of plants is increased. The plants will also tend to yield better precision, except when the comparison is with the spurious precision of severely negatively biased estimators. In practice, though, any decision to increase the number of plants deployed must be balanced against the importance of ensuring that the behaviour of any plants used must match that of members of the target population.

It was argued in section 9 that the parametric bootstrap estimate of the standard deviation of $\hat{N}_U$ should be preferred. If interval estimates are required, the resampling approach also has the advantage of providing bootstrap confidence intervals, as was shown by Buckland and Garthwaite (1991).

REFERENCES

Berg, S. (1974). Factorial series distributions, with applications to capture-recapture problems. *Scand. J. Statist.* 1:145-152.

Berg, S. (1975). Some properties and applications of a ratio of Stirling numbers of the second kind. *Scand. J. Statist.* 2:91-94.

Buckland, S.T. & Garthwaite, P.H. (1991). Quantifying precision of mark-recapture estimates using the bootstrap and related methods. *Biometrics* 47:255-268.

Chapman, D.G. (1951). Some properties of the hypergeometric distribution with applications to zoological censuses. *Univ. Calif. Public. Stat.* 1:131-160.

Charalambides, C.A. (1981). On a restricted occupancy model and its applications. *Biom. J.* 23:601-610.

Charalambides, C.A. & Koutras, M. (1983). On the differences of the generalized factorials at an arbitrary point and their combinatorial applications. *Discrete Mathematics* 47:183-201.

Duran, J.W. & Wiorkowski, J.J. (1981). Capture-recapture sampling for estimating software error content. *IEEE Trans. Software Eng.* 7:147-148.

Goudie, I.B.J. & Ashbridge, J. (2000). A conditionally-unbiased estimator of population size based on plant-capture in continuous time. *Comm. Statist. Theory Methods* 29:2605-2619.

Goudie, I.B.J. & Ashbridge, J. (2005). A conditionally unbiased estimator for the equal-catchability model. *Comm. Statist. Theory Methods* 34:1543-1554.

Goudie, I.B.J., Jupp, P.E. & Ashbridge, J. (2007). Plant-capture estimation of the size of a homogeneous population. *Biometrika* 94:243-248.

Goudie, I.B.J., Pollock, K.H. & Ashbridge, J. (1998). A plant-capture approach for population size estimation in continuous time. *Comm. Statist. Theory Methods* 27:433-451.

Gould, H.W. & Hopper, A.T. (1962). Operational formulas connected with two generalizations of hermite polynomials. *Duke Mathematical Journal* 29:51-63.

Johnson, N.L., Kotz, S. & Kemp, A.W. (2005). *Univariate Discrete Distributions*, 3rd edn. Wiley, New York.

Laska, E. M. & Meisner, M. (1993). A plant-capture method for estimating the size of a population from a single sample. *Biometrics* 49:209-220.

Laska, E. M., Meisner, M. & Siegel C. (1988). Estimating the size of a population from a single sample. *Biometrics* 44:461-472. Correction: *Biometrics* 45:1347.

Martin, E., Laska, E, Hopper, K., Meisner, M. & Wanderling, J. (1997). Issues in the Use of a Plant-Capture Method for Estimating the Size of the Street Dwelling Population. *Journal of Official Statistics* 13:59-73.

Norris, J.L.,III & Pollock, K.H. (2001). Nonparametric MLE incorporation of heterogeneity and model testing into premarked cohort studies. *Environ. Ecol. Statist.* 8:21-32.

Otis, D.L., Burnham, K.P., White, G.C. & Anderson, D.R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs* 62:1-135.

Skalski, J.R. & Robson, D.S. (1982). A mark and removal field procedure for estimating animal abundance. *J. Wildlife Manag.* 46:741-751.

U.S. Department of Housing and Urban Development. (2006) *A Guide to Counting Unsheltered Homeless People (Revised)*, 0ffice of Community Planning and Development.

Yip, P. (1995). Estimating the numbers of errors in a system using a martingale approach. *IEEE Trans. Reliability* 44:322-326.

Yip, P.S.F. (1996). Effect of plant-capture in a capture-recapture experiment. *Comm. Statist. Theory Methods* 25:2025-2038.

Yip, P. & Fong, D.Y.T. (1993). Estimating population size from a removal experiment. *Statistics and Probability Letters* 16:129-135.

Yip, P.S.F., Xi, L.Q., Fong, D.Y.T. & Hayakawa, Y. (1999) Sensitivity-analysis and estimating number-of-faults in removal debugging. *IEEE Trans. Reliability* 48:300-305.

Table 1. *The mean m and standard deviation s of each of the three estimators for a target population of size N=50.*

			t							
			5		10		15		20	
	p	R	m	s	m	s	m	s	m	s
$\hat{N}_P$	0.2	25	50.0	8.7	50.1	4.2	50.0	2.3	50.0	1.3
		5	50.0	18.2	50.0	8.4	50.0	4.6	50.0	2.6
	0.1	25	50.0	15.6	50.0	9.1	50.0	6.3	50.1	4.5
		5	47.9	25.8	50.0	19.1	50.0	12.8	50.0	9.1
	0.05	25	49.9	25.7	50.0	15.9	50.0	11.8	50.0	9.4
		5	37.0	20.8	47.7	25.9	49.6	23.4	50.0	19.5
$\hat{N}_M$	0.2	25	49.8	6.6	49.5	2.8	49.5	1.4	49.4	0.8
		5	50.2	8.1	49.6	2.9	49.5	1.5	49.5	0.8
		0	50.5	9.0	49.6	3.0	49.5	1.5	49.5	0.8
	0.1	25	51.3	14.1	49.9	6.9	49.6	4.3	49.5	3.0
		5	55.4	26.8	50.2	8.3	49.7	4.8	49.6	3.2
		0	57.7	31.5	50.6	9.3	49.8	5.0	49.6	3.2
	0.05	25	56.8	33.1	51.4	14.2	50.3	9.4	49.9	7.0
		5	57.9	36.3	55.2	26.3	51.4	12.8	50.3	8.4
		0	43.3	24.4	57.7	31.7	52.3	16.2	50.6	9.4
$\hat{N}_U$	0.2	25	50.0	6.5	50.0	2.8	50.0	1.5	50.1	0.9
		5	50.0	7.7	50.0	2.9	50.0	1.5	50.0	0.9
		0	49.9	8.3	50.0	3.0	50.0	1.5	50.0	0.9
	0.1	25	50.0	12.9	50.0	6.7	50.0	4.3	50.0	3.0
		5	49.8	17.7	50.0	7.9	50.0	4.7	50.0	3.1
		0	48.2	18.1	50.0	8.5	50.0	4.9	50.0	3.2
	0.05	25	49.8	22.5	50.0	13.0	50.0	9.1	50.0	6.9
		5	40.8	18.7	49.8	17.7	50.0	11.3	50.0	8.0
		0	30.1	13.4	48.6	18.3	49.9	12.7	50.0	8.6

Table 2. *The mean s_U of the estimator $\hat{S}_U(z,x)$ and the estimated mean $\hat{s}_B$ of the estimator $\hat{S}_B(z,x)$ compared with the true standard deviation s, for a target population of size N=50.*

		$R=5$				$R=25$			
		t				t			
p		5	10	15	20	5	10	15	20
0.2	s	7.66	2.91	1.49	0.92	6.46	2.76	1.47	0.91
	s_U	7.26	2.85	1.43	0.78	6.33	2.72	1.41	0.77
	$\hat{s}_B$	7.99	2.94	1.49	0.87	6.52	2.78	1.46	0.87
0.1	s	17.71	7.89	4.71	3.14	12.95	6.75	4.31	2.97
	s_U	15.49	7.52	4.57	3.07	12.18	6.59	4.23	2.93
	$\hat{s}_B$	17.25	8.12	4.79	3.15	13.45	6.83	4.32	2.99
0.05	s	18.72	17.66	11.30	8.01	22.53	13.01	9.07	6.87
	s_U	19.15	15.44	10.44	7.64	19.71	12.26	8.78	6.71
	$\hat{s}_B$	17.21	17.50	11.79	8.37	21.25	13.18	9.17	7.06