Estimation for the Schnabel census with plants

R. Gormley, I. B. J. Goudie[1,*]

School of Mathematics and Statistics,
University of St Andrews, St Andrews, Scotland KY16 9SS

Abstract

Samples of predetermined sizes are drawn from a closed homogeneous population augmented by a known number of planted individuals. Our results suggest that the Pathak-type estimator of its size that we exhibit should be preferred to alternative possible estimators.

Keywords: capture–recapture; conditionally unbiased estimator; factorial series distribution; Petersen estimator.
2010 MSC: 62P10, 62F10, 91D20

---

1. Introduction

The earliest multi-sample mark-recapture model is the Schnabel census, which we denote as M_f. Under that model a fixed number t of samples, of pre-chosen sizes, is drawn from a closed homogeneous target population P_1 of unknown size $N > 0$. This paper addresses the corresponding plant-capture model M_{fp}, in which we add to P_1 a population P_2 comprising a known number $R \geq 0$ of planted individuals, each marked with a unique tag. The t samples are then drawn from the augmented population, with the size of the j^{th} sample ($j = 1, \ldots, t$) being some pre-chosen constant $s_j > 0$. Individuals from P_1 are also uniquely tagged when they are first caught, so that the complete capture history is known for all individuals that are caught at least once. The behaviour of members of P_2 is assumed to be indistinguishable from that of the target individuals. In the standard mark-recapture model

---

[*] Corresponding author
Email address: `ig@st-andrews.ac.uk` (I. B. J. Goudie)
[1] Tel: +44 (0)1334 463705 Fax: +44 (0)1334 463748

M_0 (Otis et al., 1978), the sample sizes have binomial distributions, but conditioning on those sizes yields the Schnabel census, M_f. Similarly M_{fp} can be obtained by conditioning on the sample sizes in M_p, the extension of M_0 to cater for the presence of plants (Goudie et al., 2007).

The use of planted individuals when estimating population sizes has been discussed in very disparate contexts. For animal populations, Skalski and Robson (1982) and Yip and Fong (1993) considered removal experiments for populations that included pre-marked animals. In a software reliability context, Duran and Wiorkowski (1981) discussed estimating the number of indigenous software errors already present by inserting further errors. Plant-capture has also been used to assist problems of census undercount by estimating the numbers of homeless people. It gained attention from its use in the 1990 U.S. census, and has since been used in a variety of cities in North America (Hopper et al., 2008; McCandless et al., 2012).

Section 2 below summarises the distributional results for M_{fp} given by Goudie and Gormley (2013), who obtained the maximum likelihood estimator (MLE) of N, and shows that unbiased estimation may be provided by an alternative estimator of Pathak-type. As we indicate in section 3, a Petersen-type estimator of N is also a possibility under M_{fp}. In section 4 we compare the various estimators of N using computational results, and then in section 5 consider estimating the standard deviation of the Pathak-type estimator.

2. Unbiased estimation

Without loss of generality, we assume that the target individuals are numbered $i = 1, \ldots, N$ and the planted individuals $i = N+1, \ldots, N+R$. The data from the t samples can then be displayed in a $(N+R) \times t$ matrix $D = (d_{ij})$ in which $d_{ij} = 1$ if individual i is captured in sample j, and $d_{ij} = 0$ otherwise. If $\mathbf{s} = (s_1, \ldots, s_t)$, the total number of different ways of selecting the t samples is $A(N, \mathbf{s}, R) = (N+R)_{s_1} \ldots (N+R)_{s_t} / (s_1! \ldots s_t!)$, where $(N)_s = N(N-1)\ldots(N-s+1)$ when s is a positive integer and $(N)_0 = 1$.

Now let X denote the total number of different individuals captured from population P_1, and define $z = s_1 + \ldots + s_t$, $\tilde{c} = \max\{s_1, \ldots, s_t\}$ and $c = \max(0, \tilde{c} - R)$. Goudie and Gormley (2013) note that X is a sufficient statistic for N, and show that its probability function is

$$p_t(x; \mathbf{s}, N, R) = (N)_x a(x, \mathbf{s}, R) / A(N, \mathbf{s}, R) \qquad x = c, \ldots, \min(z, N), \quad (1)$$

where, if Δ denotes the forward finite difference operator,

$$a(x, \mathbf{s}, R) = \frac{1}{x!}\left[\Delta^x A(N, \mathbf{s}, R)\right]_{N=0} = \frac{1}{x!}\sum_{k=0}^{x}(-1)^k\binom{x}{k}A(x-k, \mathbf{s}, R).$$

The product $x!\,a(x, \mathbf{s}, R)$ can be interpreted as giving the number of ways of drawing the samples so that each of a set of x specified target individuals is captured at least once and any other individuals caught are plants.

The probability function (1) is that of a (range-restricted) factorial series distribution (FSD), as defined by Berg (1974), with series function $A(N, \mathbf{s}, R)$. Using Berg's terminology, the range of this FSD is the set of consecutive integers given by $I \equiv \{x : a(x, \mathbf{s}, R) > 0\} \equiv \{x : c \leq x \leq z\}$. It follows from his results that, for $N \in I$, the unique unbiased estimator of N is

$$\tilde{N}(x) = \begin{cases} x + \{a(x-1, \mathbf{s}, R)/a(x, \mathbf{s}, R)\} & x = c+1, \ldots, z; \\ x & x = c. \end{cases} \tag{2}$$

In the case $R = 0$, the observed value of X must be positive, but, if plants are present and $R \geq \tilde{c}$, the observation $x = 0$ has positive probability as shown in (1). The special case of the estimator $\tilde{N}$ for the situation $R = 0$ was discussed by Berg (1974), who attributed it to Pathak (1964).

Consider now an alternative scenario with a target population of size $N + R$, but no plants. Suppose that we draw a sample of size R followed by t samples with sizes $s_1, \ldots, s_t$. Consider the probability that the z sightings in the t samples include x further distinct individuals, in addition to the R in the initial sample. Clearly this probability equals the above probability $p_t(x; \mathbf{s}, N, R)$ of observing x distinct individuals under our previous scenario with R plants present. Letting $\mathbf{s}^* = (R, \mathbf{s})$, it may be verified that

$$\begin{aligned} a(x, \mathbf{s}, R) &= R!\,a(x+R, \mathbf{s}^*, 0) & (3)\\ p_t(x\,; \mathbf{s}, N, R) &= p_{t+1}(x+R\,; \mathbf{s}^*, N+R, 0). & (4) \end{aligned}$$

These relations may be used to derive results for M_{fp} from those for M_f.

If we define $\mathbf{s}'$ to be $(s_1, \ldots, s_{t-1}, s_t - 1)$ when $s_t > 1$, and $(s_1, \ldots, s_{t-1})$ for $s_t = 1$, then, using, (3), it follows from equation (2.4) of Berg (1976) that

$$s_t a(x, \mathbf{s}, R) = (x + R - s_t + 1)a(x, \mathbf{s}', R) + a(x-1, \mathbf{s}', R).$$

In particular, if $\tilde{\mathbf{s}} = (s_1, \ldots, s_t, 1)$, we have $a(x, \tilde{\mathbf{s}}, R) = (x+R)a(x, \mathbf{s}, R) + a(x-1, \mathbf{s}, R)$, from which it follows that an alternative expression for the estimator (2) is given by $\tilde{N}(x) = -R + a(x, \tilde{\mathbf{s}}, R)/a(x, \mathbf{s}, R)$ $(x = c, \ldots, z)$.

From a computational standpoint, using a recurrence relation to find the value of the estimator $\tilde{N}(x)$ is better than evaluating the function $a(x, \mathbf{s}, R)$. If we temporarily use the fuller notation $\tilde{N}(x; \mathbf{s}, R)$ for the estimate $\tilde{N}(x)$, then, using the argument leading to (3), it may be similarly verified that

$$\tilde{N}(x\,; \mathbf{s}, R) = \tilde{N}(x + R\,; \mathbf{s}^*, 0) - R. \tag{5}$$

The required recurrence relation can now be obtained from equation (2.6) of Berg (1976), using (5). We obtain that $\tilde{N}(x\,; \mathbf{s}, R) - x$ equals

$$\begin{cases} \left(\tilde{N}(x\,; \mathbf{s}', R) - x\right) \dfrac{\tilde{N}(x-1\,; \mathbf{s}', R) + R - s_t + 1}{\tilde{N}(x\,; \mathbf{s}', R) + R - s_t + 1} & c < x < z; \\ \tilde{N}(x-1\,; \mathbf{s}', R) + R - s_t + 1 & x = z; \\ 0 & x = c. \end{cases}$$

It can be shown that, under model M_p (Ashbridge and Goudie, 2009), in which $s_1, \ldots, s_t$ are random, the improvement to $\tilde{N}(x)$ given by the Rao-Blackwell Theorem is the conditionally unbiased estimator (CUE) $\tilde{N}_C(z, x)$ for M_p. Again setting $z = s_1 + \ldots + s_t$, this estimator $\tilde{N}_C(z, x)$ is defined by $\tilde{N}_C(z, 0) = 0$ ($z = 0, \ldots, Rt$), and, for $x = 1, \ldots, N$ and $z = x, \ldots, xt + Rt$, by $\tilde{N}_C(z, x) = x + G(z, x-1, t, Rt)/G(z, x, t, Rt)$, in which $G(z, x, t, \theta)$ is the Gould–Hopper number, or non-central C-number, defined by

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

For $t = 1$, the CUE $\tilde{N}_C(z, x)$ for M_p equals $(R+1)x/(z-x+1)$ (Ashbridge and Goudie, 2009). The implied estimator of the size $N+R$ of the augmented population is the version of the Petersen estimate of $N + R$ due to Chapman (1951), namely $\{(R + 1)(z + 1)/(z - x + 1)\} - 1$. Under M_{fp}, if $R > 0$, the special case of $\tilde{N}(x)$ for $t = 1$ is also Chapman's estimator. For, when $t = 1$, we have $z = s_1$ and it may be verified that $a(x, s_1, R) = G(z, x, 1, R)/z!$. It follows that, for $t = 1$, the estimators $\tilde{N}(x)$ and $\tilde{N}_C(z, x)$ are identically equal. For the case $R = 0$, a similar situation arises for $t = 2$. The equivalence of $\tilde{N}(x)$ and Chapman's estimator for this case was noted by Berg (1976).

3. Other estimators

When plants are present, an estimator of Petersen-type is an obvious candidate to consider. If Y is the number of distinct planted individuals seen, the most basic form of Petersen estimator $\tilde{N}_B$ results from equating the observed proportion y/R of plants to the estimated proportion $x/\tilde{N}_B$ of target individuals seen. Setting $W = X + Y$, we investigate a Chapman-type variant, namely $\tilde{N}_P = (R + 1)x/(w - x + 1)$. The joint distribution of W and X and an expression for the expected value of $\tilde{N}_P$ are derived in the Appendix. By the discussion in section 2, in the special case where $t = 1$, the estimators $\tilde{N}, \tilde{N}_C$ and $\tilde{N}_P$ are all equivalent.

As the true population size N is integer valued, in the comparative study below we consider $\hat{N}(x) = [\tilde{N}(x) + 0.5]$, where the square brackets denote the integer part. The estimators $\hat{N}_C(z, x)$ and $\hat{N}_P(x)$ are similarly obtained by rounding $\tilde{N}_C(z, x)$ and $\tilde{N}_P(x)$ respectively.

The final estimator included in this study is the MLE. When $x = z$, the likelihood function obtained from (1) is a monotone increasing function of N (Goudie and Gormley, 2013). For $x < z$, however, the likelihood has a unique maximum and the value of the MLE $\hat{N}_M(x)$ is given by the largest $N \geq x + 1$ such that $N(N + R - s_1) \ldots (N + R - s_t) > (N - x)(N + R)^t$, or, if this inequality is not satisfied for any such N, by $\hat{N}_M = x$.

4. Comparison of the estimators

As the MLE does not exist when $x = z$, conditioning on the event $E \equiv \{X < z\}$ is essential for comparative analyses involving that estimator. In Table 1 we therefore give the conditional mean μ', given the event E, and the conditional standard deviation σ', again given E, for the Pathak-type estimator $\hat{N}$, the CUE $\hat{N}_C$ for M_p, the MLE $\hat{N}_M$ and the Petersen-type estimator $\hat{N}_P$. To obtain the conditional moments of $\hat{N}_P$ we need the conditional joint distribution of W and X given that $X < z$. This is, in fact, straightforward to obtain since, as the events $\{X = z\}$ and $\{X = W = z\}$ are equivalent, it follows that $P(X < z) = 1 - P(X = W = z)$.

Table 1 addresses both situations where $R > 0$ and those where $R = 0$, when the true size of P_1 is $N = 100$. If R plants are present, by homogeneity, we expect a fraction $N/(N+R)$ of the observations to be from P_1. For a better comparison with the case $R = 0$, we therefore increase the total sample size z when plants are present so that the total number of observations Z_T of target individuals has an expectation matching the value of z when $R = 0$.

When $z > N = 100$, the raw estimator $\tilde{N}$ is unbiased, but it seemed plausible that both conditioning on E and rounding to the nearest integer might make $\hat{N}$ slightly biased. In fact, for $z \geq 50$, Table 1 suggests that the bias of $\hat{N}$ is negligible and at least as small as that of the MLE $\hat{N}_M$. For these values of z, the Petersen-type estimator $\hat{N}_P$, which can only be used when $R > 0$, is effectively unbiased, but suffers from a relatively large σ'. For $z \geq 50$, the CUE $\hat{N}_C$ is usually more precise than $\hat{N}_M$, and on balance is somewhat better in terms of bias. Under M_p, for which $\hat{N}_C$ was designed, the sample sizes are binomially distributed with a constant mean. It is thus unsurprising that $\hat{N}_C$ performs better for equal sample sizes than for sample sizes that show wide variation. Although σ' can be smaller for $\hat{N}_C$ than for $\hat{N}$, this is often a reflection of its smaller mean. Due to its smaller bias, the Pathak-type estimator $\hat{N}$ should be regarded as preferable.

Table 1. *The conditional mean* μ' *and standard deviation* σ' *of each of the four estimators of* N *when* P_1 *has size* N*=100. In the second column,* $\mathbf{u}_K$ *denotes the* K*-vector* $(1, \ldots, 1), \boldsymbol{\alpha} = (14\mathbf{u}_4, 4)$ *and* $\boldsymbol{\beta}$ *= (22, 17, 12, 7, 2).*

			$\hat{N}$		$\hat{N}_C$		$\hat{N}_M$		$\hat{N}_P$	
z	$\mathbf{s}$	R	μ'	σ'	μ'	σ'	μ'	σ'	μ'	σ'
15	$5\mathbf{u}_3$	0	39.0	8.0	36.6	7.0	61.4	18.3	-	-
18	$6\mathbf{u}_3$	20	94.9	47.0	93.7	46.4	130.1	94.4	94.2	56.1
21	$7\mathbf{u}_3$	40	99.7	47.0	99.4	46.8	116.8	70.6	99.9	51.9
25	$5\mathbf{u}_5$	0	83.5	31.2	80.7	29.9	122.0	69.2	-	-
30	$6\mathbf{u}_5$	20	100.0	40.2	98.7	39.9	113.3	57.8	99.8	52.8
35	$7\mathbf{u}_5$	40	100.0	32.1	99.2	31.9	106.7	37.2	100.0	37.2
50	$10\mathbf{u}_5$	0	100.1	28.7	98.5	28.1	107.4	36.3	-	-
60	$12\mathbf{u}_5$	20	100.0	21.3	99.3	21.0	102.8	22.8	100.0	31.7
60	$\boldsymbol{\alpha}$	20	100.0	21.2	100.7	21.4	102.8	22.9	100.0	31.5
60	$\boldsymbol{\beta}$	20	100.0	21.3	104.0	22.2	103.0	23.1	100.0	31.2
70	$14\mathbf{u}_5$	40	100.0	18.5	99.6	18.5	101.6	19.3	100.0	23.1
125	$25\mathbf{u}_5$	0	100.0	8.0	99.5	8.0	100.0	8.1	-	-
150	$30\mathbf{u}_5$	20	100.0	7.4	99.6	7.4	99.8	7.6	100.0	13.8
175	$35\mathbf{u}_5$	40	99.9	7.1	99.7	7.0	99.7	7.0	100.0	10.5

For the smallest total sample size $z = 15$ shown in Table 1, the three estimators, $\hat{N}, \hat{N}_C$ and $\hat{N}_M$, are all strongly negatively biased, with $\hat{N}_M$ being the best of them. For the other cases in which $15 < z < 50$, $\hat{N}_M$ is the most biased, and its positive bias gives it large values of σ'. For these values of z and for $R > 0$, the Petersen-type estimator $\hat{N}_P$, which essentially uses the proportion of plants detected, becomes less biased and more precise as R

increases. Its performance here in terms of bias, however, is matched by that of $\hat{N}$, for which σ' remains somewhat smaller. Indeed its small bias makes $\hat{N}$ the preferred choice for these values of z, even though $\hat{N}_C$ approximates it fairly closely. Thus, except when z is very small, Table 1 suggests that the most reliable of these four estimators is $\hat{N}$, the one of Pathak-type.

The above assessments of the estimators have not been unduly coloured by the need to use conditional moments in order to make comparisons with the MLE $\hat{N}_M$. Omitting $\hat{N}_M$, we re-examined the scenarios used in Table 1 using instead the unconditional mean μ and standard deviation σ. The broad conclusions were unaffected, despite some appreciable changes of detail. In particular, as shown in Table 2, when $R = 0$ and $\mathbf{s} = 5\mathbf{u}_3$, the negative bias of $\hat{N}$, although still large, is less pronounced than in Table 1. For larger z the improvements in the bias become trivial. The estimator $\hat{N}_C$ showed similar behaviour. Overall, inspection of unconditional moments further strengthened the preference for the Pathak-type estimator $\hat{N}$, except, as previously mentioned, when z is very small.

Table 2. *The unconditional mean* μ *and standard deviation* σ *of* $\hat{N}(x)$ *when* P_1 *has size* N*=100, together with the means* σ_u *and* $\bar{s}_B$ *of the estimates of* σ.

$\mathbf{s}$	R	μ	σ	σ_U	$\bar{s}_B$	$\mathbf{s}$	R	μ	σ	σ_U	$\bar{s}_B$
$5\mathbf{u}_3$	0	61.9	26.1	36.5	22.5	$12\mathbf{u}_5$	20	100.0	21.3	20.3	21.6
$6\mathbf{u}_3$	20	99.2	61.8	48.3	57.0	$\boldsymbol{\alpha}$	20	100.0	21.2	20.3	21.2
$7\mathbf{u}_3$	40	99.8	48.2	40.8	50.6	$\boldsymbol{\beta}$	20	100.0	21.3	20.3	21.3
$5\mathbf{u}_5$	0	95.5	55.3	46.8	45.5	$14\mathbf{u}_5$	40	100.0	18.5	18.0	18.6
$6\mathbf{u}_5$	20	100.0	40.8	35.1	45.8	$25\mathbf{u}_5$	0	100.0	8.0	7.9	8.1
$7\mathbf{u}_5$	40	100.0	32.1	29.9	35.3	$30\mathbf{u}_5$	20	100.0	7.4	7.3	7.4
$10\mathbf{u}_5$	0	100.1	28.7	25.7	31.0	$35\mathbf{u}_5$	40	99.9	7.1	7.0	7.0

5. Other considerations

The general variance estimator of a FSD (Berg, 1974, p. 148) provides one estimator of the unconditional variance σ^2 of the Pathak-related estimator $\tilde{N}$ that can also be used for the rounded version $\hat{N}$. When $N \in I$, an unbiased estimator $\hat{v}_U(X)$ of the variance of $\tilde{N}$ is given by $\hat{v}_U(X) = \{h_1(X)\}^2 + h_1(X) - h_2(X)$, where $h_v(x) = a(x - v, \mathbf{s}, R)/a(x, \mathbf{s}, R)$ for $x \geq v$ and zero otherwise. We can therefore estimate the standard deviation of $\hat{N}$ using the square root $\hat{s}_U(X)$ of $\hat{v}_U(X)$. For each of the situations considered in Table 2, the expected value σ_U of $\hat{s}_U$ can be compared with the true value of σ.

A second way to estimate σ is to use a parametric bootstrap. When the population size estimate is $\hat{N}(x)$, a large number of realisations from an augmented population of size $\hat{N}(x) + R$ is simulated, with each realisation comprising t samples. The value of $\hat{N}(X)$ is then calculated for each of these realisations. The value of σ can be estimated by the standard deviation $\hat{s}_B$ of these population size estimates from the resamples.

To produce Table 2, 1000 random observations of X under its unconditional distribution were selected, and, for each of the ensuing estimates $\hat{N}(x)$, 1000 realisations of the sampling process were generated as described above. The mean $\bar{s}_B$ of the resulting 1000 estimates $\hat{s}_B$ is shown in the table. In all but one of the situations considered, the estimator $\hat{s}_U$ is negatively biased, albeit with small bias for large z. In contrast, $\hat{s}_B$ tends to have a positive bias of smaller absolute size. Indeed, for $z > 60$, $\hat{s}_B$ appears almost unbiased. In three of the situations with small sample sizes, however, $\hat{s}_B$ shows negative bias. If there is no major disparity in the size of the biases, overestimation of a point estimator's variance should be seen as preferable to underestimation if over-optimism about the estimator's accuracy is to be avoided. This suggests that $\hat{s}_B$ should be preferred to $\hat{s}_U$ except when $R = 0$ and $z \leq 25$, in which circumstance $\hat{s}_B$ displays negative bias.

For large sample sizes, the probability that the MLE does not exist, which is $P(\bar{E}) = P(X = z)$, can be very small. Indeed, as $X \leq N$, it is exactly zero when $z > N$. On the other hand, if, when $R = 0$, few observations are taken, $P(\bar{E})$ can be substantial. For instance, $P(\bar{E}) = 0.4492$ for the first situation in Table 2. Those who prefer to use the MLE may wish to note that using 20 plants reduces this probability to 0.0108.

More generally, Table 1 suggests that the benefit from using plants depends on how much information is available. When $E(Z_T)$ is 125, providing strong information, increasing the number of plants reduces σ', but only for $\hat{N}_C(x)$ is the mean improved. When $E(Z_T) = 50$ and sample sizes are constant, increasing R removes the slight bias of $\hat{N}$, reduces that of $\hat{N}_C$ and $\hat{N}_M$, and also reduces σ' for all these estimators. For very weak information, when $E(Z_T) < 50$, σ' is often not reduced by using plants. This is unsurprising when the plants lead to larger means which are often accompanied by larger variances. Indeed, using 20 plants reduces the absolute bias of $\hat{N}_M$, and the negative biases of both $\hat{N}$ and $\hat{N}_C$ are much reduced, often with a further improvement if 40 plants are deployed. We conclude that it is for small or moderate sample sizes that using plants is effective.

Appendix

To evaluate exactly the moments of $\tilde{N}_P$ for general t, we require the joint distribution of X and W. Note firstly that, as $A(N+R,\mathbf{s},0) = A(N,\mathbf{s},R)$, equation (1) implies that, for $w = \tilde{c}, \ldots, \min(z, N+R)$, the probability function of W is $(N+R)_w\, b(w,\mathbf{s})/A(N,\mathbf{s},R)$ where $b(w,\mathbf{s}) = a(w,\mathbf{s},0)$. As the conditional distribution of X given W is clearly hypergeometric, the joint probability function $q(w,x;N,R)$ of W and X is given by

$$\frac{(N+R)_w\, b(w,\mathbf{s})}{A(N,\mathbf{s},R)} \cdot \binom{N}{x}\binom{R}{w-x} \Big/ \binom{N+R}{w} = \binom{N}{x}\binom{R}{w-x}\frac{w!\, b(w,\mathbf{s})}{A(N,\mathbf{s},R)}.$$

For convenience, we define this probability function on the set $\{(w,x) : w = 0,\ldots,z; x = 0,\ldots,w\}$, noting that some probabilities are then zero under the usual convention that the combinatorial coefficient $(N)_x/x! = 0$ for $x > N$ and hence also $b(w,\mathbf{s}) = 0$ for $\tilde{c} > w$. The expected value of $\tilde{N}_P$ is

$$\begin{aligned} & \sum_{w=0}^{z}\sum_{x=1}^{w} \frac{(R+1)x}{w-x+1}\binom{N}{x}\binom{R}{w-x}\frac{w!\, b(w,\mathbf{s})}{A(N,\mathbf{s},R)} \\ = \quad & N\sum_{w=0}^{z}\sum_{m=0}^{w-1}\binom{N-1}{m}\binom{R+1}{w-m}\frac{w!\, b(w,\mathbf{s})}{A(N-1,\mathbf{s},R+1)} \\ = \quad & N\{1 - q(0,0;N-1,R+1) - \ldots - q(z,z;N-1,R+1)\}. \end{aligned}$$

Thus $\tilde{N}_P$ is negatively biased with the size of the bias being the product of N and the probability, when the parameters are $N-1$ and $R+1$, that $W = X$, or equivalently that $Y = 0$. An elementary combinatorial argument then shows that the bias of $\tilde{N}_P$ is $-N\, A(N-1,\mathbf{s},0)/A(N,\mathbf{s},R)$. Note, in particular, that if $N \le \tilde{c}$, the joint distribution of W and X is such that, when the parameters are $N-1$ and $R+1$, we must have $X \le N-1 < \tilde{c} \le W$, and hence the event $W = X$ has probability zero, implying that $\tilde{N}_P$ is unbiased.

Acknowledgement

The research work of Richard Gormley has been supported by the Engineering and Physical Sciences Research Council.

References

Ashbridge, J. and Goudie, I. B. J. (2009), 'Conditionally unbiased estimation of population size under plant-capture.', *Commun. Stat.* A **38**, 1–12.

Berg, S. (1974), ‘Factorial series distributions, with applications to capture-recapture problems’, *Scand. J. Stat.* **1**, 145–152.

Berg, S. (1976), ‘A note on the UMVU estimate in a multiple-recapture census’, *Scand. J. Stat.* **3**, 86–88.

Chapman, D. (1951), ‘Some properties of the hypergeometric distribution with applications to zoological sample censuses’, *Univ. Calif. Public. Stat.* **1**, 131–160.

Duran, J. and Wiorkowski, J. (1981), ‘Capture-recapture sampling for estimating software error content.’, *IEEE Trans. Softw. Eng.* **7**, 147–148.

Goudie, I. B. J. and Gormley, R. (2013), ‘Maximum likelihood estimates for the Schnabel census with plants’, *Commun. Stat.* A, to appear .

Goudie, I., Jupp, P. and Ashbridge, J. (2007), ‘Plant-capture estimation of the size of a homogeneous population’, *Biometrika* **94**, 243–248.

Hopper, K., Shinn, M., Laska, E., Meisner, M. and Wanderling, J. (2008), ‘Estimating numbers of unsheltered homeless people through plant-capture and postcount survey methods.’, *Am. J. Public Health* **98**, 1438–1442.

McCandless, L., Patterson, M., Currie, L., Moniruzzaman, A. and Somers, J. (2012), ‘Counting the unsheltered homeless in Edmonton using plant-capture studies from Toronto.’, *Report, Faculty of Health Sciences, Simon Fraser University, Canada* .

Otis, D., Burnham, K., White, G. and Anderson, D. (1978), ‘Statistical inference from capture data on closed animal populations’, *Wildl. Monogr.* **62**, 1–135.

Pathak, P. (1964), ‘On estimating the size of a population and its inverse by capture mark method’, *Sankhya A* **26**, 75–80.

Skalski, J. and Robson, D. (1982), ‘A mark and removal field procedure for estimating population abundance’, *J. Wildl. Manag.* **46**, 741–751.

Yip, P. and Fong, D. (1993), ‘Estimating population size from a removal experiment’, *Stat. Probab. Lett.* **16**, 129–135.