Biometrika (2000), **87,** 3, *pp.* 633–649

Printed in Great Britain

A nonparametric approach to the analysis of two-stage mark–recapture experiments

BY SONG XI CHEN

Department of Statistics and Applied Probability, National University of Singapore, Singapore 117 543

songxichen@yahoo.com

AND CHRIS J. LLOYD

Australian Graduate School of Management, Sydney, New South Wales 2052, Australia

chrisl@agsm.unsw.edu.au

SUMMARY

We present a new approach to the analysis of two-stage mark–recapture experiments where individual covariates are available which describe the detectability of individuals in the population. Central to the theory is a single quantity α, which we call the heterogeneity index and which measures the variability in the detectability of individuals. This theory also has implications for the analysis of independent observer line transect surveys. An estimator of α is provided by combining kernel density estimates of the underlying densities of the covariate, conditional on capture histories. We derive expressions for the asymptotic bias and variance and show that our new estimator is as efficient as the Petersen estimator when individuals all have the sames detectability. We also report the results of a simulation study of the new estimator for a range of mark–recapture conditions.

Some key words: Heterogeneity; Kernel density estimation; Weighted distribution.

1. INTRODUCTION

In a standard mark–recapture experiment, animals from a population of interest are captured, marked and then released. At a later occasion after the captured animals have mixed with uncaptured ones, another sample is taken from the population. The proportion of recaptures provides information about the probability of capture, which in combination with total capture numbers provides information on N, the total number in the population; see Seber (1982) for a definitive review.

An important consideration is that capture probabilities may vary across occasions, with capture history or from individual to individual. Failure to account for sources of capture heterogeneity typically leads to underestimation of the population. When individuals' covariates are available, Huggins (1989) and Alho (1990) proposed a logistic regression model for capture probabilities in terms of these covariates. Parameters are estimated from a conditional likelihood and N is then estimated by a Horvitz–Thompson-type estimator. The theory and methods proposed in this paper are best viewed as an alternative to those of Huggins and Alho. The advantages as we see them are that (i) there

is no need to maximise the conditional likelihood, (ii) the approach requires little more than nonparametric density estimation, and (iii) we obtain not only an estimator of the population size but also an estimator of the frequency distribution of the covariate, which may be of interest in practice. Also we show that the effect of individual heterogeneity on estimation of N depends on a single quantity α which measures the variability in the detectability of individuals. Our method of analysis is based directly on estimation of α.

The two-stage mark–recapture experiment is mathematically analogous to an independent observer line transect survey. In this design, two observers independently record sightings of individuals as well as their distance and other covariates from the randomly chosen transect line; see Butterworth & Borchers (1988) and Buckland & Turnock (1992). The observers take the place of capture occasions in a mark–recapture experiment. The reason for the second observer is that perfect detection along the transect need not be assumed. In a conventional single-observer line transect survey, this assumption is required; see Buckland, Anderson et al. (1993) for a comprehensive review. The conditional logistic regression methods of Huggins and Alho have been taken up and refined by Buckland, Anderson et al. (1993, Ch. 4), Manly, McDonald & Garner (1996) and Borchers, Buckland et al. (1998) for independent observer line transect surveys. Chen (1999, 2000) has developed a nonparametric method for the case in which the underlying covariate distribution is uniform. A very general likelihood framework is developed by Borchers, Zucchini & Fewster (1998).

The plan of the paper is as follows. Section 2 establishes notation, suitable for both the mark–recapture experiment and the independent-observer-line-transect survey context. In § 3 we define the index α and in § 4 we give several alternative expressions and interpretations of this index. Sections 5 and 6 present our estimator and give some asymptotic properties. Sections 7 and 8 report results for real and simulated examples. In § 9 we compare our method with other methods in terms of the different likelihood factorisations on which they are based.

2. MATHEMATICAL FRAMEWORK AND NOTATION

Consider a population of N individuals each with measurable covariate vector $X \in \Re^d$, such as age or weight in the mark–recapture experiment context or sighting distance and cluster size in the independent-observer-line-transect survey context.

Two samples S_1 and S_2 are taken. Upper case J denotes a general sampling history, such as '1' for sample 1 and '01' for sample 2 but not sample 1. We denote the number of individuals with history J by n_J. For instance, n_{11} is the number detected in both samples, and so $n := n_1 + n_2 - n_{11}$ is the number of distinct individuals detected. Our data comprise a list of the n detected individuals together with their covariate values x_i and detection histories.

Let f be the density function of the covariates X. We make no assumption about $f(x)$. For instance in the mark–recapture experiment context, when x might be age, we make no assumption about the age distribution of the population. The population values $X_1, \ldots, X_N$ are a simple random sample from $f(x)$. The observed values $x_1, \ldots, x_n$ are a possibly biased sample from the covariate population.

The conditional probability that an individual i with covariate value x_i is in sample S_j is denoted by

$$g_j(x_i) := \mathrm{pr}(i \in S_j | X_i = x_i) \quad (j = 1, 2),$$

and we call g_1 and g_2 the detection functions. We assume that for each individual the two detections are independent. Thus the probability of detection history $J = 11$ for an individual with covariate x is $g_{11}(x) = g_1(x)g_2(x)$, and similarly for g_{10}, g_{01} and g_{00}. The density of X conditional on history J is

$$f_J(x) = g_J(x)f(x)/p_J, \tag{1}$$

where $p_J := \int g_J(x)f(x)\,dx = \mathrm{pr}(i \in S_J)$. We let $p = 1 - p_{00}$ denote the probability that an individual is detected at all. We finally assume independence of individual detections on each sampling occasion. In this case, n and n_J have binomial distributions with parameters (N, p) and (N, p_J) respectively.

3. Likelihood estimation

The data comprise observed covariates and histories $\{(x_i, J_i): i = 1, \ldots, n\}$ for all detected individuals from which the full likelihood may be written down in several ways. Borchers, Zucchini & Fewster (1998) give several partitions of the likelihood, which we further discuss in § 9. The marginal distribution of the totals $(n_{10}, n_{01}, n_{11}, N - n)$ is multinomial and carries information about average sampling probabilities p_J. The corresponding likelihood component is

$$L(N, p_1, p_2, p; \{n_J\}) \propto \frac{N!}{(N-n)!}(p - p_2)^{n_{10}}(p - p_1)^{n_{01}}(p_1 + p_2 - p)^{n_{11}}(1 - p)^{N-n}. \tag{2}$$

Conditional on history totals $\{n_J\}$, the distribution of the histories $\{J_i\}$ is a known distribution and contributes nothing to the likelihood. The likelihood for the covariates $\{x_i\}$ given the histories $\{J_i\}$ and totals $\{n_J\}$ is

$$L(f_{10}, f_{01}, f_{11}; \{x_i\}) | \{J_i\}, \{n_i\}) \propto \prod_{J_i = 10} f_{10}(x_i) \prod_{J_i = 01} f_{01}(x_i) \prod_{J_i = 11} f_{11}(x_i). \tag{3}$$

This term gives information about the densities $f_J(x)$ of the covariate. Since we intend to model these densities nonparametrically, we initially concentrate on (2).

By differencing (2) with respect to N we easily show that the maximum likelihood estimator of N equals the integer part of n/p. However, the three probability parameters (p_1, p_2, p) are not all identifiable in (2). To proceed further we reparameterise the p_J in terms of the positive quantity

$$\alpha = \int \{f_1(x)f_2(x)/f(x)\}\,dx = E\left\{\frac{f_1(x)}{f(x)}\,\frac{f_2(x)}{f(x)}\right\}, \tag{4}$$

which we call the heterogeneity index. Throughout, all expectations are with respect to f, and the integration area is $A(f) = \{x \in \Re^d \,|\, f(x) > 0\}$. This α is an extension of the definition of Chen (1999, 2000), who assumed $f(x)$ was uniform with respect to the distance x in the independent-observer-line-transect survey context. A related quantity $\mathrm{IG} = \int g_1(x)g_2(x)\,dx$ has appeared in the line transect literature for univariate line transect surveys when f is uniform on $[0, w]$; see Butterworth & Borchers (1988), Buckland (1987), Buckland & Turnock (1992) and D. L. Borchers' 1996 University of Cape Town Ph.D. thesis. Under these assumptions, it is easily seen that $\mathrm{IG} = p_1 p_2 w\alpha$.

Probability parameters may now be expressed in terms of (p_1, p_2, α). For instance

$p_{11} = p_1 p_2 \alpha$ and $p = p_1 + p_2 - \alpha p_1 p_2$. The likelihood (2) now becomes

$$\frac{N!}{(N-n)!}(p_1 - \alpha p_1 p_2)^{n_{10}}(p_2 - \alpha p_1 p_2)^{n_{01}}(\alpha p_1 p_2)^{n_{11}}(1 - p_1 - p_2 + \alpha p_1 p_2)^{N-n}. \tag{5}$$

Reparameterisation does not make the parameters identifiable. However, we will see later that α is estimable externally. We therefore consider estimation of (N, p_1, p_2) in (5) for fixed α, and it is easy to verify that

$$\hat{p}_1(\alpha) = \frac{n_{11}}{\alpha n_2}, \quad \hat{p}_2(\alpha) = \frac{n_{11}}{\alpha n_1}, \quad \hat{p}(\alpha) = \frac{n n_{11}}{\alpha n_1 n_2}, \quad \hat{N}(\alpha) = \frac{n_1 n_2}{n_{11}}\alpha. \tag{6}$$

4. INTERPRETATION OF α

From (4), we see that α is a measure of how far the density ratios $f_1(x)/f(x)$ and $f_2(x)/f(x)$ deviate from one. Under homogeneity, these ratios are both identically one and thus $\alpha = 1$. In this section, we give three further interpretations of α.

Interpretation 1: *α is a correction factor for bias.* From (6) the maximum likelihood estimator of N for fixed α is $\hat{N}(\alpha) = \hat{N}_P \alpha$, where $\hat{N}_P = n_1 n_2 / n_{11}$ is the well-known Petersen estimator. While the exact distribution of $\hat{N}(\alpha)$ is complicated, we show in Appendix 1 that

$$\text{bias}\{\hat{N}(\alpha)\} = \frac{(1-p_1)(1-p_2)}{p_1 p_2} + \frac{\alpha - 1}{\alpha}\left(\alpha - \frac{1}{p_1 p_2}\right) + O\left(\frac{1}{N}\right), \tag{7}$$

$$\text{var}\{\hat{N}(\alpha)\}/N = \frac{(1-p_1)(1-p_2)}{p_1 p_2} + \frac{\alpha - 1}{\alpha}\left(2\alpha - \frac{1}{p_1 p_2}\right) + O\left(\frac{1}{N}\right). \tag{8}$$

Thus, $\hat{N}(\alpha)$ is consistent for N whereas $\hat{N}_P$ is consistent not for N but for N/α. Only in the homogeneous case is $\hat{N}_P$ consistent with bias $(1-p_1)(1-p_2)/(p_1 p_2)$ in agreement with Darroch (1958, p. 352) and Lloyd (1994, eqn (7)).

Interpretation 2: *α is a measure of catch dependence.* Even though we assume detections are independent at the individual level, heterogeneity induces detection dependence at the population level. We saw that $p_{11} = \alpha p_1 p_2$ so α directly measures departure from independence. Only when $\alpha = 1$ does $p_{11} = p_1 p_2$ and detections are independent. Within a wider framework of spatial heterogeneity, Darroch (1961, § 4.4) has also noted that $\hat{N}_P$ is inconsistent if the correlation between $I_{i \in S_1}$ and $I_{i \in S_2}$ is nonzero, calling such correlation 'catch dependence'.

Interpretation 3: *α is a measure of detection variability.* Let γ_i^2 be the coefficient of variation of $g_i(X)$ and ρ_{12} be the correlation between $g_1(X)$ and $g_2(X)$. Then it is easy to show that $\alpha = 1 + \rho_{12}\gamma_1\gamma_2$. In the special case that $g_1 = g_2 = g$, α is simply the coefficient of variation of $g(X)$ plus 1 and so cannot be less than 1. Thus $\hat{N}_P$ has nonpositive bias in this context, as is well known in the mark–recapture experiment literature. On the other hand, if either g_1 or g_2 is constant then $\alpha = 1$ and $\hat{N}_P$ is consistent. This is easily explained: any method, biased or not, may be used to mark a cohort of individuals. So long as the second sampling is random it follows that $n_{11} \sim \text{Bi}(n_2, n_1/N)$ conditional on n_2. The Petersen estimator then follows simply. It is perhaps less intuitive that, if the initial markings are done randomly, then the second sampling may be biased, so long as the probability

of capture does not depend on the tags. It is only when detection heterogeneity is present for both samples simultaneously that the Petersen estimator may become invalid.

Equation (4) simplifies when the components of the covariate are independent conditional on their history. Let $f_{J_i}(x)$ be the density of the ith component conditional on history J. It then follows that

$$\alpha = \alpha_1 \alpha_2 \ldots \alpha_d, \tag{9}$$

where $\alpha_i = \int f_{1i} f_{2i} / f_{11i}$ is a measure of capture heterogeneity due to component x_i. The overall α is a product of these component specific values. Each covariate affecting detection probabilities that we fail to identify will lead to a typically downward bias in our estimation. We can mitigate this bias by identifying as many covariates as possible with the largest values of α_i. When components are not independent then the absolute value of α is generally less than the product of marginal values. This is to be expected since, when components are dependent, information about unseen components is carried in the measured components.

We finally give an alternative expression for α which suggests an estimation procedure. Using $\alpha = p_{11}/(p_1 p_2)$, $g_{11}(x) = g_1(x) g_2(x)$ and (1) we have

$$\alpha = \frac{p_{11}}{p_1 p_2} \int \frac{g_1(x) g_2(x)}{g_{11}(x)} f(x)\, dx = \int \frac{f_1(x) f_2(x)}{f_{11}(x)}\, dx. \tag{10}$$

Unlike (4), this alternative expression has $f_{11}(x)$ instead of $f(x)$ in the denominator. This is useful since $f_{11}(x)$ is directly estimable from the covariate values of the individuals in S_{11}, whereas $f(x)$ is not directly estimable.

5. Kernel estimation of α

To estimate α it suffices to estimate f_1, f_2 and f_{11}. The covariate values for the individuals indexed by S_1, S_2 and S_{11} are direct samples from these distributions. One could certainly propose parametric forms for the density functions and then maximise (3). Here we investigate a nonparametric approach.

Let K be a d-dimensional kernel that is a probability density itself and satisfies moment conditions

$$\int u K(u)\, du = 0, \quad \int u u^{\mathrm{T}} K(u)\, du = I_d,$$

for $u \in \Re^d$, where I_d is the d-dimensional identity matrix. Let H_J, for $J = \{1\}$, $\{2\}$ and $\{11\}$, be d-dimensional matrices that contain the smoothing bandwidths and such that $H_J = h_J A_J$, where h_J is a positive scalar quantity and $|A_J| = 1$. To simplify the notation, we assume that $H_J = h_J I_d$, as otherwise the data can be transformed linearly by A_J^{-1}. We assume that $N h_J^2 \to \infty$ as $N \to \infty$.

A kernel estimator of $f_J(x)$ based on the sample S_J is

$$\hat{f}_J(x) = \frac{1}{n_J h_J^d} \sum_{i \in S_J} K\left(\frac{x - X_i}{h_J}\right).$$

The last expression for α given in (10) suggests the estimator

$$\hat{\alpha} = \int_{-\infty}^{\infty} \frac{\hat{f}_1(t) \hat{f}_2(t)}{\hat{f}_{11}(t)}\, dt, \tag{11}$$

which depends on smoothing parameters h_1, h_2 and h_{11}. Substituting in (6) we have

$$\hat{N}(\hat{\alpha}) = \int_{-\infty}^{\infty} \frac{n_1 \hat{f}_1(t) n_2 \hat{f}_2(t)}{n_{11} \hat{f}_{11}(t)}\, dt. \tag{12}$$

We denote the integrand on the right-hand side by $\hat{N}(x)$ which estimates $Nf(x)$. This describes the frequency distribution of the covariate across the population and may be of interest in its own right.

6. Asymptotic properties of $\hat{N}(\hat{\alpha})$

When we write $O(h^4)$ we mean a term which behaves like an order-4 product of h_1, h_2 and h_{11}, such as $h_1 h_2 h_{11}^2$. We first give expressions for the asymptotic accuracy of the integrand $\hat{N}(x)$ of (12). In Appendix 2 we show that

$$\begin{aligned}\operatorname{bias}\{\hat{N}(x)\} = {} & \frac{1}{2} Nf(x) \sum_{l=1}^{d} \left\{ h_1^2 \frac{f''_{1l}(x)}{f_1(x)} + h_2^2 \frac{f''_{2l}(x)}{f_2(x)} - h_{11}^2 \frac{f''_{11l}(x)}{f_{11}(x)} \right\} \\ & + h_1^{-d} R(h_2/h_1) + h_{11}^{-d} \left\{ \frac{R(1)}{g_1(x) g_2(x)} - \frac{R(h_2/h_{11})}{g_2(x)} - \frac{R(h_1/h_{11})}{g_1(x)} \right\} \\ & + f(x) \left\{ \frac{(1-p_1)(1-p_2)}{p_1 p_2} + \frac{\alpha - 1}{\alpha} \left(\alpha - \frac{1}{p_1 p_2} \right) \right\} \\ & + O(Nh^4 + h^{-d+1}),\end{aligned} \tag{13}$$

where f''_{Jl}, for $l = 1, \ldots, d$, is the second-order partial derivative of f_J with respect to the lth component of x and $R(\theta) = \int K(t) K(\theta t)\, dt$ for a scalar θ. When $d = 1$, $R(\theta) = \{2\pi(1+\theta^2)\}^{-\frac{1}{2}}$ for the normal kernel. The first term in (13) is $O(Nh^2)$ and is a typical bias term from kernel smoothing of the densities. The second and third terms in (13) are $O(h^{-\mathrm{d}})$ and come from the asymptotic variance of the kernel density estimators. The fourth term in (13) is an $O(1)$ parametric term from estimating Np_1p_2/p_{11} by $n_1 n_2 / n_{11}$ or equivalently from estimating N by $\hat{N}_P \alpha$ with α known; see (7). The bias of $\hat{N}$ is obtained by integrating the above bias expression for $\hat{N}(x)$. The terms in this asymptotic bias are of the same orders as for $\hat{N}(x)$.

Appendix 2 derives the asymptotic expression

$$\begin{aligned}\{Nf(x)\}^{-1} \operatorname{var}\{\hat{N}(x)\} = {} & \frac{R(1)}{h_1^d g_1(x)} + \frac{R(1)}{h_2^d g_2(x)} + \frac{R(1)}{h_{11}^d g_1(x) g_2(x)} - 2 \frac{R(h_1/h_{11})}{h_{11}^d g_1(x)} - 2 \frac{R(h_2/h_{11})}{h_{11}^d g_2(x)} \\ & + 2 \frac{R(h_1/h_2)}{h_2^d} - 1 + O(h^2).\end{aligned}$$

The estimator $\hat{N}(x)/N(x)$ as an estimator of 1 has bias $O\{h^2 + 1/(Nh^d)\}$ and variance $O\{1/(Nh^d)\}$. From these orders it is a routine calculation to show that, provided $h \to 0$ and $Nh^d \to \infty$ as $N \to \infty$, the bandwidth minimising mean squared error is of the classical order $h = O\{N^{-1/(4+d)}\}$.

To derive an expression for $\operatorname{var}\{\hat{N}(\hat{\alpha})\}$, we integrate the covariance of $\hat{N}(x)$ and $\hat{N}(y)$ with respect to x and y. Appendix 2 derives the expression

$$\operatorname{var}\{\hat{N}(\hat{\alpha})\} = N \int \frac{\{1 - g_1(s)\}\{1 - g_2(s)\}}{g_1(s) g_2(s)} f(s)\, ds + O(Nh^2). \tag{14}$$

Note in particular that the variance of $\hat{N}(\hat{\alpha})$ is $O(N)$ whereas the variance of $\hat{N}(x)$ was $O(N/h^d)$. This implies that optimal bandwidths for estimating $\hat{N}(x)$ and for estimating N are not of the same order.

We have earlier identified the effect of heterogeneity on bias. The new estimator $\hat{N}(\hat{\alpha})$ tries to adjust for this bias. What is the effect on variance? An alternative expression for the leading term of (14) is

$$N \operatorname{cov}\{g_1^{-1}(X), g_2^{-1}(X)\} + NE\left\{\frac{1-g_1(X)}{g_1(X)}\right\} E\left\{\frac{1-g_2(X)}{g_2(X)}\right\}, \tag{15}$$

and the second term is what the variance of the Petersen estimator would be if the detection functions were constant at the values $E\{g_i(X)\}$. Note that the covariance term is positive if and only if ρ_{12} as defined earlier is positive. This suggests that positive catch or detection dependence not only causes negative bias shown in (7) but also inflates the achievable asymptotic variance. When either g_1 or g_2 is constant, the covariance term is zero and the asymptotic variance is identical to the well-known asymptotic variance of the Petersen estimator. This means that allowance for possible detection heterogeneity, when there is none, comes at no cost in asymptotic variance. This observation is largely borne out in later simulations.

The freedom of the leading term of (14) from h_J inspires confidence that the choice of bandwidths will have relatively little influence on our estimator of N. Let us assume that $h_2 = \kappa_2 h_1$ and $h_{11} = \kappa_{11} h_1$, and define the constants

$$b_1 = \frac{1}{2}\int f(t) \sum_{l=1}^{d} \left\{\frac{f_{1l}''(t)}{f_1(t)} + \kappa_2^2 \frac{f_{2l}''(t)}{f_2(t)} - \kappa_{11}^2 \frac{f_{11l}''(t)}{f_{11}(t)}\right\} dt,$$

$$b_2 = \int \left[R(\kappa_2) + \kappa_{11}^{-1} \left\{\frac{R(1)}{g_1(t)g_2(t)} - \frac{R(\kappa_2/\kappa_{11})}{g_2(t)} - \frac{R(1/\kappa_{11})}{g_1(t)}\right\}\right] dt.$$

Then the mean squared error of $\hat{N}(\hat{\alpha})$ is

$$\text{MSE}\{\hat{N}(\hat{\alpha})\} = N \int \frac{\{1-g_1(s)\}\{1-g_2(s)\}}{g_1(s)g_2(s)} f(s)\, ds + (Nh_1^2 b_1 + b_2/h_1^d)^2,$$

with error $o\{(Nh^2 + 1/h^d)^2\}$. Minimising mean squared error with respect to h_1, we derive the asymptotically optimal h_1 as

$$h_1^* = \left[\frac{\sqrt{\{8+(d-2)^2\}}|b_1 b_2| + (d-2)b_1 b_2}{4b_1^2}\right]^{-1/(d+2)} N^{-1/(d+2)}.$$

This implies that the optimal rate for h_1, h_2 and h_{11} is $N^{-1/(d+2)}$. With this rate, the bias of $\hat{N}(\hat{\alpha})$ is $O(N^{d/(d+2)})$ and the variance is $O(N) + O(N^{d/(d+2)})$. Note that the optimal bandwidth actually minimises the asymptotic bias and does not affect the leading term of the asymptotic variance. This is in contrast to other smoothing problems where there is a bias-variance trade-off in the leading terms.

7. An example

The number of aboriginal peoples in the area of Vancouver-Richmond in British Columbia, Canada, is of interest to government agencies, as is the age distribution of this population. The previous census counted around 10 000 people but census is notoriously

unreliable for such social groups and some local government bodies believe the population could be as high as 30 000. It is likely that the chance of recording an individual will depend on their age so both the count and the age distribution from the census are likely to be biased.

There are at least two alternative sources of data on this population. The first is a questionnaire distributed at aboriginal gatherings. The second source is data from those applying for or waiting for government housing. Both these sources covered the period December 1998 to February 1999. From the survey protocols, there is good reason to believe that the surveys are statistically independent. For each individual surveyed, the birth date was recorded, allowing calculation of the person's age on 1 March 1999, which will be our covariate x. The two surveys had similar penetration, with $n_1 = 1358$ and $n_2 = 1285$, though there was no reason beforehand to expect that this would be the case. Enough information was recorded to match the $n_{11} = 93$ individuals appearing in both surveys.

Figure 1(a) shows estimates of the detection functions $g_1(x)$ and $g_2(x)$, where x is measured in years. These are local kernel smooths of the binary event of detection in both surveys conditional on detection in one. The bandwidth has been chosen according to a standard plug-in bandwidth rule and we will not dwell on the details here. The 'dips' in the detection function near 18 and 65 had already been observed in previous studies and at least the first of these was anticipated from the sociology of that age-group. Individuals in these age ranges are under-represented in the surveys and the purpose of our method is to correct for such bias.

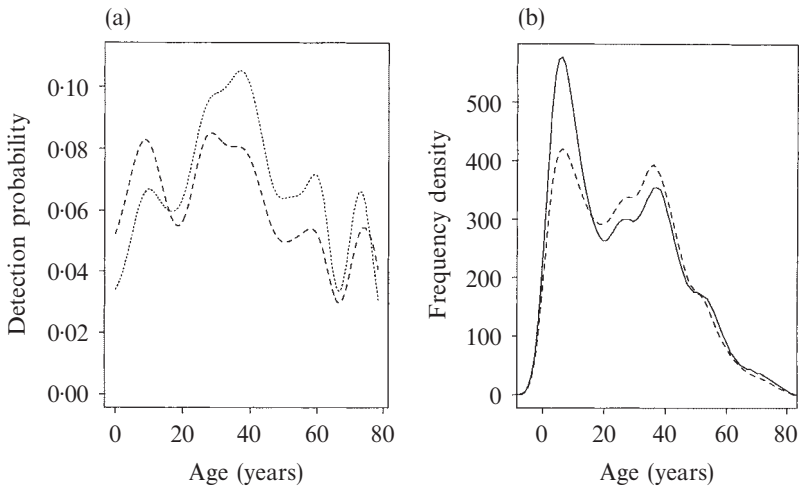

Fig. 1. Aboriginal data. (a) Nonparametric estimates of the detection functions $g_1(x)$ (dotted line) and $g_2(x)$ (dashed line). (b) Estimated frequency curve $\hat{N}(x)$ allowing for heterogeneity (solid) and assuming homogeneity (dotted).

The Petersen estimate is 18 590 with standard error 1820, where we have employed the bias correction of Chapman (1951). Using the plug-in bandwidth estimator mentioned in § 8·2, we obtain an estimate of $\hat{N}(\hat{x}) = 20\,990$ with standard error 2780. The bandwidths used are $h_1 = 11{\cdot}8$ and $h_2 = 9{\cdot}9$. After substracting off the 2550 known individuals, our estimate of 18 440 undetected individuals is 15% higher than the 16 040 undetected individuals estimated assuming homogeneity. The estimate $\hat{N}(x)$ is displayed as a solid line in Fig. 1(b), which also shows the naive estimate obtained by scaling up an ordinary

density estimator, with bandwidth 9·6, by the naive population estimate 18 590. The new method predicts many more individuals in the age-group around 18 years old. There are also more individuals estimated near 65 but these numbers are much smaller.

Bearing in mind the shape of Fig. 1(a), it is unlikely that parametric logistic regression, such as Huggins (1989) and Alho (1990) describe, will be successful in describing the detection functions $g_1(x)$ and $g_2(x)$. As a quick check, we attempted to fit the simple linear logistic regression in age. Buckland, Breiwick et al. (1993) have suggested a simple algorithm for maximising the conditional likelihood but this algorithm is not always successful and fails at the first iteration in this case, as well as for a quadratic logistic regression model. One could certainly fit logistic regression models to these data, but this would require direct numerical maximisation of the conditional likelihood and we have not pursued this.

8. Simulation results on the performance of $\hat{N}(x)$ and $\hat{N}$

8·1. *Modification of estimator*

Numerical problems in the computation of $\hat{N}(x)$ and $\hat{N}$ can be expected in the extreme tails where both numerator and denominator will be close to zero. To address this problem, we take the view that in covariate ranges where there is hardly any redetection it will be next to impossible to account for detection bias. We will say that there is 'little redetection data' at x if

$$n_{11}\hat{f}_{11} = \sum_{i \in S_{11}} h_{11}^{-1} K\{h_{11}^{-1}(x - X_i)\} < h_{11}^{-1} K(0).$$

The left-hand side is the effective number of redetections at x while the right-hand side is the value of the left-hand side if there was a single redetection at x. In regions where there is little redetection data $\hat{N}(x)$ is defined to equal $\hat{N}\{\hat{f}_1(x) + \hat{f}_2(x)\}/2$. Provided that $f_{11}(x) > 0$ for all x, this modified estimator is still consistent since for large enough N all regions will contain redetection data. We have experimented with other rules for handling regions of sparse data. Provided that there is no region with high $f(x)$ and low $g_i(x)$, where no estimator will perform well, we have found little to choose between the different methods.

8·2. *Bandwidth selection*

For estimating N we expect that the choice of h_1, h_2 and h_{11} will have a moderate effect. We have not pursued formal optimisation of the mean squared error of $\hat{N}(x)$ but have rather used existing plug-in bandwidths for the density estimators of f_1, f_2 and f_{11}, namely the two-stage selector of Wand & Jones (1995, p. 72) which involves estimating the fourth derivative of the curve assuming that it is normal.

8·3. *Range of simulation conditions*

We used $N = 500$, $N = 1000$ and two density functions $f(x)$, one the standard normal and the other a mixture $0{\cdot}7N(0, 1) + 0{\cdot}3N(3, 0{\cdot}75)$, displayed in Fig. 2(a). Figure 2(b) displays four detection functions of our own devising, labelled $d_0, \ldots, d_3$. For $N = 1000$ we used the same four detection functions divided by 1·25 so as to reduce the total proportion of detections to more realistic levels. Each combination of density and detection functions implies a value of α and these are listed in Table 1.

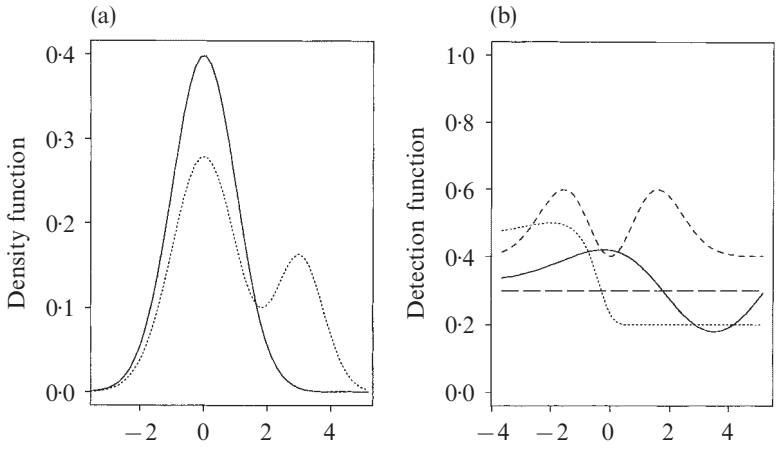

Fig. 2. Density and detection functions chosen for the simulations. (a) The standard (solid line) and mixed (dotted line) normal densities. (b) Four bias functions: d_0 (long dashed line), d_1 (solid line), d_2 (dotted line), d_3 (short dashed line).

Table 1. *Implied heterogeneity parameter* α *for each combination of detection functions*

	Second detection function			
Label	d_0	d_1	d_2	d_3
d_0	1·00, 1·00	1·00, 1·00	1·00, 1·00	1·00, 1·00
d_1	1·00, 1·00	1·01, 1·10	1·01, 1·08	0·99, 1·03
d_2	1·00, 1·00	1·01, 1·08	1·14, 1·19	0·99, 1·03
d_3	1·00, 1·00	0·99, 1·03	0·99, 1·03	1·02, 1·05

The first value in each pair corresponds to the normal density; the second to the normal mixture density.

8·4. *Results for estimating* $N(x)$

Figure 3 summarises the simulation results for two of the 40 simulation scenarios covered. Results under other scenarios are similar and available from the authors in a technical report. The upper three plots describe results with detection functions d_0 and d_2, $f(x)$ normal and $N = 1000$. The lower three plots describe results with detection functions d_3 and d_3, $f(x)$ a normal mixture and $N = 500$. Each set of three plots summarises the variability of 500 simulated curve estimates $\hat{N}(x)$.

We summarise the samples of curves in three ways. First, Figs 3(a), (d) display pointwise 5%, 50% and 95% quantiles. The plotted points summarise the true population $\mathscr{X}$ by a kernel density estimate based on the automatic bandwidth selector discussed earlier. In order to describe better the range of shapes in the 500 sample curves we have used ideas of Jones & Rice (1992) of treating each curve as a multivariate observation and performing principle components analysis on these data. Figures 3(b), (e) display curves whose scores in the first principle component were ranked 25th, 250th and 475th out of 500. Thirdly, Figs 3(c), (f) display the results for the second principal component. For both sets of

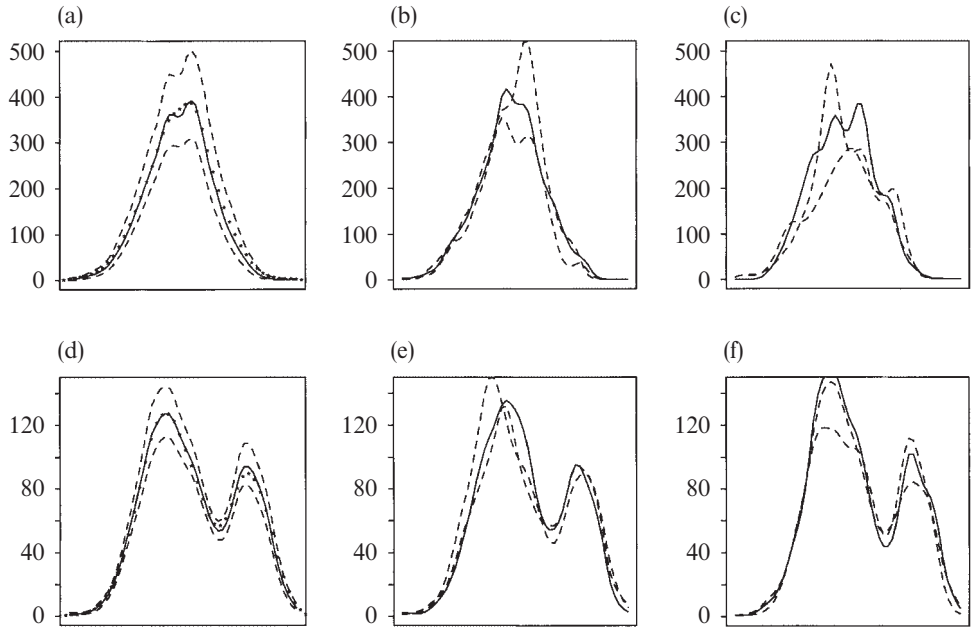

Fig. 3. Variability of 500 simulated curve estimates $\hat{N}(x)$. Plots (a), (b) and (c) are for $N = 1000$, f standard normal and detection functions d_0 and d_2 ($\alpha = 1$). Plots (d), (e) and (f) are for $N = 500$, f a normal mixture and detection functions d_3 and d_3 ($\alpha = 1{\cdot}05$). Plots (a) and (d) give pointwise 5%, 50% and 95% quantiles; the points are a kernel summary of the actual population $\mathscr{X}$. Plots (b) and (e) give curves whose first principal component ranked 25, 250 and 475 out of 500. Plots (c) and (f) repeat this for the second principal component.

simulations, the first two principal components accounted for just under 80% of the total variation.

8·5. *Results for estimating N*

Table 2 compares the accuracy of $\hat{N}(\hat{\alpha})$ and the Petersen estimator $\hat{N}_P = \hat{N}(1)$. In view of the well-known high skewness of population size estimators, it is not appropriate to look at the mean squared error of these estimators directly. Rather, the row marked 'RMSE(%)' measures the root mean square of

$$T = \log(\hat{N} - n_1 - n_2 + n_{11}) - \log(N - n_1 - n_2 + n_{11}),$$

which compares the estimated number of undetected individuals with the actual number of undetected individuals, on the log scale; see Chao (1989) for justification. The row labelled 'Wgts' gives the combination of detection functions. For instance 1/2 means one detection function is of the form d_1 and the other of the form d_2 as detailed in Table 1. As expected from equation (15) the variance of the bias-adjusted estimator tends to become larger than that of the Petersen estimator as the heterogeneity index α increases. However, the actual standard deviation is often much less than the estimated standard error; see the row marked 'Ave(SE)'. This suggests that the leading neglected term in (14) is negative under all these models. Under negative catch dependence scenarios the reverse would be true, but we have not presented results for this less common situation.

Table 2. *Simulation comparison of new estimator with Petersen estimator. Summaries of* 500 *simulations of* $\hat{N}(\hat{\alpha})$ (*upper figures*) *and* $\hat{N}_P$ (*lower figures*)

	f standard normal				f normal mixture			
	Wgts				Wgts			
	0/0	1/0	2/0	3/0	0/0	1/0	2/0	3/0
				$N = 500$				
Mean(SE)	500(55)	501(42)	493(58)	500(34)	490(50)	497(52)	486(54)	502(38)
	502(55)	500(40)	498(52)	499(34)	499(53)	505(52)	499(55)	502(37)
RMSE(%)	22	19	23	19	22	22	23	21
	22	18	21	19	22	21	22	21
Ave(SE)	62	52	67	38	59	64	72	38
	54	43	53	35	53	50	58	37
Captures	255	288	254	322	255	270	244	320
				$N = 1000$				
Mean(SE)	997(103)	992(87)	989(111)	1000(69)	981(99)	987(91)	975(111)	999(68)
	1000(99)	994(86)	999(102)	1000(69)	998(101)	1010(93)	1010(114)	999(68)
RMSE(%)	18	21	19	7	18	22	19	14
	17	21	17	7	17	21	18	14
Ave(SE)	118	107	127	78	115	92	138	80
	103	91	103	69	102	96	112	72
Captures	422	288	419	445	422	446	403	533

	f standard normal				f normal mixture			
	Wgts				Wgts			
	1/1	2/2	3/3	2/1	1/1	2/2	3/3	2/1
				$N = 500$				
Mean(SE)	502(38)	505(81)	505(25)	501(50)	494(53)	486(75)	503(25)	491(61)
	496(35)	441(41)	494(23)	499(42)	466(36)	434(46)	492(24)	478(49)
RMSE(%)	19	28	7	22	18	28	5	25
	19	36	7	19	18	37	5	24
Ave(SE)	43	97	38	58	74	98	27	80
	35	40	33	45	38	47	24	48
Captures	315	246	366	286	278	226	365	256
				$N = 1000$				
Mean(SE)	1000(73)	1000(139)	995(49)	1000(91)	983(101)	978(139)	997(48)	982(112)
	990(67)	883(80)	975(46)	995(81)	927(71)	874(91)	984(47)	951(92)
RMSE(%)	15	22	13	9	16	19	13	19
	15	29	15	8	22	31	14	21
Ave(SE)	85	174	53	107	139	190	54	149
	68	82	47	82	74	92	48	92
Captures	531	410	625	483	463	377	625	427

8·6. *Bivariate case*

We simulated from a population of $N = 1000$ with the covariate (x_1, x_2) having a bivariate standard normal distribution. The detection functions were chosen to be multiplicative, that is $g_J(x_1, x_2) = g_J(x_1)g_J(x_2)$, and so $\alpha = \alpha_1\alpha_2$; see equation (9). The independence of the covariates was not used in our estimation procedure though it would be advantageous to do so. We chose g_1 and g_2 to take the shape of detection function d_3 but rescaled to

vary from 0·4 to 0·8. The implied heterogeneity index was $\alpha = (1{\cdot}073)^2 = 1{\cdot}151$ so that the mean of the Petersen estimator should be around $1000/1{\cdot}151 = 868$. The expected number of distinct detections is 463 and of redetections is 88. In Fig. 4(a), for a typical simulation, the dotted line bounds the region within which the effective number of redetections exceeds 1. Outside this region the modified estimator is used but the estimated frequencies here are quite small. Figure 4(b) compares the Petersen estimator with $\hat{N}(\hat{\alpha})$ for 250 simulations. The histograms are of the quantity T defined above, and the horizontal scale essentially measures proportional error in the estimated and actual number of unseen individuals. Apparently, the negative bias is reduced without excessive inflation of variance. The root mean squared error for the two estimators on this scale are 33% for $\hat{N}_P$ and 20% for $\hat{N}(\hat{\alpha})$. We note again that the estimated variance from the leading term of (14) tends to overestimate the true variability. It is also noteworthy that the occasional occurrences of rather large values of $\hat{N}(\hat{\alpha})$ are always flagged by a large standard error.

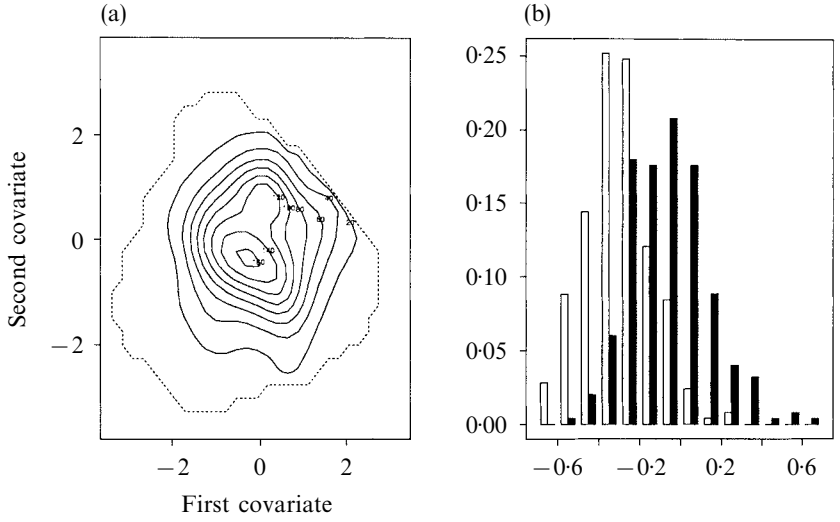

Fig. 4. Simulation from bivariate standard normal population with $N = 1000$ and multiplicative detection functions described in the text. (a) Contour plot of realisation of $\hat{N}(x)$; contour lines are in steps of 20, ranging from 160 in centre to 20 on outside; dotted line bounds the area where the data are not sparse. (b) Summary of 250 simulations; unfilled bars are for $\hat{N}_P$, filled bars are for $\hat{N}(\hat{\alpha})$.

9. Discussion

Heterogeneity is a major problem for estimating population size. Each extra source of heterogeneity leads to a bias in the Petersen estimator. Multiple sources would typically lead to extreme downward bias, for instance through the multiplicative relation in (9). Both our methods and the methods of Huggins (1989) and Alho (1990) attempt to adjust for this bias. One problem common to both approaches is that when there are many measured sources of heterogeneity then modelling $g_j(x)$ or $f_J(x)$ is subject to the 'curse of dimensionality'. Some readers might consider it natural to assume an additive predictor in the Huggins and Alho approach, but this assumption is no more plausible than assuming independence of the sources, in which case our method is simply implemented by estimating α_i separately for each component. In summary, there is no particular advantage to either method with respect to the curse of dimensionality. Another problem is that, when

there are unmeasured sources of heterogeneity, then accounting for only the measured ones will not eliminate all sources of bias. Both methods are subject to this problem. Accounting for heterogeneity from unmeasured sources can be done in principle but is difficult in practice; see for instance Lloyd & Yip (1991).

The full data comprise $\{x_i, J_i\}$ for detected individuals $i = 1, \ldots, n$. The collection of histories $\{J_i\}$ implies values for the history totals $\{n_J\}$. Thus, the distribution of $\{x_i\}$ conditional on $\{J_i\}$ and $\{n_J\}$ is the same as that conditional on $\{J_i\}$ only. Note also that the distribution of $\{n_J\}$ given $\{J_i\}$ is a known multinomial and so contributes nothing to the likelihood. Our method is based on factorising the full likelihood as

$$\begin{aligned} L(\{x_i, J_i\}) &= L_1(\{n_J\})L_2(\{J_i\}\,|\,\{n_J\})L_3(\{x_i\}\,|\,\{J_i\}, \{n_J\}) \\ &\propto L_1(\{n_J\})L_3(\{x_i\}\,|\,\{J_i\}, \{n_J\}). \end{aligned}$$

For fixed α, we used maximum likelihood estimation of the unknown parameters in the first factor $L_1(\{n_J\})$. The last factor, which is simply a product of the conditional densities f_1, f_2 and f_{11}, motivates direct estimation of these densities from the observed $\{x_i\}$. We are thus combining efficient maximum likelihood estimation with kernel density estimation. For this reason we anticipate that our estimator is fully efficient asymptotically, and have demonstrated this explicitly for the case of homogeneity.

Borchers, Zucchini & Fewster et al. (1998) have factorised the full likelihood as

$$L(\{x_i, J_i\}) = L_1(n)L_2(\{x_i\}\,|\,n)L_3(\{J_i\}\,|\,\{x_i\}, n).$$

The first term is a $\mathrm{Bi}(N, p)$ likelihood. The second term is a product of the densities, conditional on being captured at all, in other words of the densities

$$p^{-1}\{g_1(x) + g_2(x) - g_1(x)g_2(x)\}\, f(x).$$

They comment that 'no attempt seems to have been made in the mark–recapture literature to model the density of the explanatory variables affecting detectability'. We have done this, but through the factor $L(\{x_i\}\,|\,\{J_i\}, \{n_J\})$ rather than the term $L(\{x_i\}\,|\,n)$. The third term in the above factorisation is a product of binomial terms with probabilities depending on detection functions. For instance the observation $(x_i, J_i = 11)$ gives a contribution

$$g_1(x_i)g_2(x_i)\{1 - g_1(x_i)g_2(x_i)\}.$$

Especially in the independent-observer-line-transect survey context, it may be quite reasonable to propose parametric forms for the $g_i(x)$, such as linear models on the log-odds scale.

Finally we consider the factorisation underlying the methods of Huggins and Alho:

$$L(\{x_i\}, \{J_i\}) = L_1(n)L_2(\{C_i\}\,|\,n)L_3(\{x_i\}\,|\,\{C_i\}, n)L_4(\{J_i\}\,|\,\{x_i\}, \{C_i\}),$$

where $C_i = 1$ if individual i is captured at all. The last factor L_4 involves products of detection probabilities conditional on the event of capture at all. Unfortunately, assuming a logistic model for the g_i does not imply a simple model for the conditional detection functions; the conditional likelihood is not straightforward to maximise and may not be well behaved. The first three terms of the factorisation are not used in the Huggins approach. For instance the term L_3, which is identical to the last factor in our methods, is ignored. Estimation of N is based on a Horvitz–Thompson-type method rather than the estimator $n/\hat{p}$ implied by the first likelihood factor, because an estimator of p does not follow from the conditional approach.

Appendix 1

Asymptotics of $\hat{N}(\alpha)$

Define $r_1 = n_1/(Np_1)$, $r_2 = n_2/(Np_2)$ and $r_{11} = n_{11}/(Np_{11})$, which are all consistent estimators of 1 with error $O_p(1/\sqrt{N})$. Further define

$$r := \frac{\hat{N}_P}{N}\,\alpha = \frac{n_1 n_2}{N n_{11}}\,\frac{p_{11}}{p_1 p_2} = \frac{r_1 r_2}{r_{11}}.$$

By Taylor expansion of the r_i about their limiting value of 1 we have

$$\begin{aligned} r \doteq\, & 1 + (r_1 - 1) + (r_2 - 1) - (r_{11} - 1) + (r_1 - 1)(r_2 - 1) \\ & - (r_1 - 1)(r_{11} - 1) - (r_2 - 1)(r_{11} - 1) + (r_{11} - 1)^2. \end{aligned}$$

Thus

$$E(r) \doteq 1 + \operatorname{cov}(\hat{r}_1 - \hat{r}_{11}, \hat{r}_2 - \hat{r}_{11}).$$

Since $(n_{00}, n_{10}, n_{01}, n_{11})$ is multinomially distributed, $\operatorname{cov}(\hat{n}_1, \hat{n}_2) = N(p_{11} - p_1 p_2)$ and so

$$\operatorname{cov}(\hat{r}_1, \hat{r}_2) = \frac{p_{11} - p_1 p_2}{N p_1 p_2} = \frac{\alpha - 1}{N}.$$

Similarly

$$\operatorname{cov}(\hat{r}_1, \hat{r}_{11}) = \frac{1 - p_1}{N p_1}, \quad \operatorname{cov}(\hat{r}_2, \hat{r}_{11}) = \frac{1 - p_2}{N p_2}, \quad \operatorname{var}(\hat{r}_{11}) = \frac{1 - p_{11}}{N p_{11}},$$

and collecting terms we find that

$$E(\hat{N}_p \alpha) \doteq N + \alpha - 1 + \frac{1 - p_{11}}{p_{11}} - \frac{1 - p_1}{p_1} - \frac{1 - p_2}{p_2},$$

which reduces to (7). From (A1) below we also obtain an expression

$$\begin{aligned} \operatorname{var}(r) &\doteq \operatorname{var}(\hat{r}_1) + \operatorname{var}(\hat{r}_2) + \operatorname{var}(\hat{r}_{11}) - 2\operatorname{cov}(\hat{r}_1, \hat{r}_{11}) - 2\operatorname{cov}(\hat{r}_2, \hat{r}_{11}) + 2\operatorname{cov}(\hat{r}_1, \hat{r}_2) \\ &= \frac{1 - p_1}{N p_1} + \frac{1 - p_2}{N p_2} + \frac{1 - p_{11}}{N p_{11}} - \frac{2(1 - p_1)}{N p_1} - \frac{2(1 - p_2)}{N p_2} + \frac{2(\alpha - 1)}{N}, \end{aligned}$$

which reduces to (8).

Appendix 2

Asymptotics of $\hat{N}(x)$

If we condition on all sampling histories $\mathscr{S}$ then the density estimator $\hat{f}_J(x)$ has mean

$$\mu_J(x) = \int \frac{1}{h_J^d} K\left(\frac{x - t}{h_J}\right) f_J(t)\, dt = f_J(x) + \frac{1}{2} h_J^2 \sum_{l=1}^{d} f''_{Jl}(x) + O(h_J^4).$$

For deriving the bias and variance of $\hat{N}(x)$ using $f_1 f_2 / f_{11} = \alpha f$ we have

$$\begin{aligned} \frac{\hat{N}(x)}{\hat{N}_P \alpha f(x)} = 1 &+ \frac{\hat{f}_1 - f_1}{f_1} + \frac{\hat{f}_2 - f_2}{f_2} - \frac{\hat{f}_{11} - f_{11}}{f_{11}} - \frac{(\hat{f}_1 - f_1)(\hat{f}_{11} - f_{11})}{f_1 f_{11}} \\ & - \frac{(\hat{f}_2 - f_2)(\hat{f}_{11} - f_{11})}{f_2 f_{11}} + \frac{(\hat{f}_1 - f_1)(\hat{f}_2 - f_2)}{f_1 f_2} + \frac{(\hat{f}_{11} - f_{11})^2}{f_{11}^2}, \end{aligned}$$

with cubic and smaller terms neglected. Conditional on $\mathscr{S}$,

$$E\left(\frac{\hat{N}(x)}{\hat{N}_P \alpha f(x)}\middle|\mathscr{S}\right) = 1 + \frac{1}{2}\sum_{l=1}^{d}\left\{h_1^2\frac{f_{1l}''(x)}{f_1(x)} + h_2^2\frac{f_{2l}''(x)}{f_2(x)} - h_{11}^2\frac{f_{11l}''(x)}{f_{11}(x)}\right\}$$
$$- \frac{R(h_1/h_{11})}{n_1 h_{11}^d f_1(x)} - \frac{R(h_2/h_{11})}{n_2 h_{11}^d f_2(x)} + \frac{R(1)}{n_{11} h_{11}^d f_{11}(x)}$$
$$+ \frac{n_{11} R(h_1/h_2) f_{11}(x)}{n_1 n_2 h_2^d f_1(x) f_2(x)} + O(Nh^4 + h^{-d+1}).$$

Multiplying this by $\hat{N}_p \alpha f(x)$ and taking expectations with respect to $\mathscr{S}$ we obtain

$$E\{\hat{N}(x)\} = E(\hat{N}_p \alpha) f(x)\left\{1 + \frac{1}{2}\left(h_1^2\frac{f_1''}{f_1} + h_2^2\frac{f_2''}{f_2} - h_{11}^2\frac{f_{11}''}{f_{11}}\right)\right\}$$
$$+ \frac{R(h_1/h_2)}{h_2^d} + h_{11}^{-d}\left\{\frac{R(1)}{g_1 g_2} - \frac{R(h_2/h_{11})}{g_2} - \frac{R(h_1/h_{11})}{g_1}\right\}, \tag{A1}$$

and use of the earlier expression for the mean gives the expressions (13).

Define $\hat{\eta}_J(x) = n_J \hat{f}_J(x)$ so that $\hat{N}(x) = \hat{\eta}_1(x)\hat{\eta}_2(x)/\hat{\eta}_{11}(x)$. Computing the variance of $\hat{N}$ requires an expression for $\operatorname{cov}\{\hat{\eta}_I(x), \hat{\eta}_J(y)\}$. To obtain this we first condition on $\mathscr{S}$. As $E\{\hat{\eta}_J(x)|\mathscr{S}\} = n_J \mu_J(x)$, we have

$$\operatorname{cov}\{E(\hat{\eta}_I(x)|\mathscr{S}), E(\hat{\eta}_J(y)|\mathscr{S})\} = N(p_{I\cap J} - p_I p_J)\mu_I(x)\mu_J(y). \tag{A2}$$

The conditional covariance is

$$\operatorname{cov}\{\hat{\eta}_I(x), \hat{\eta}_J(y)|\mathscr{S}\} = \sum_{i\in S_I}\sum_{j\in S_J}\operatorname{cov}\left\{\frac{1}{h_I^d}K\left(\frac{x-X_i}{h_I}\right), \frac{1}{h_J^d}K\left(\frac{x-X_j}{h_J}\right)\middle|\mathscr{S}\right\}$$
$$= n_{I\cap J}\{C(x, y, I, J) - \mu_I(x)\mu_J(x)\}, \tag{A3}$$

where $C(x, y, I, J) = f_{I\cap J}(x)K^{(2)}(x-y, h_I, h_J) + O(h_I)$. Here

$$K^{(2)}(z, h_1, h_2) = \int K(t)h_2^{-d}K\left(\frac{z-h_1 t}{h_2}\right)dt$$

is the density of $h_1 Z_1 + h_2 Z_2$, where $Z_1, Z_2 \in R^d$ are independent random vectors with density K. There is an alternative expression for $C(x, y, I, J)$ reversing the roles of (x, h_I) and (y, h_J). Combining (A2) and (A3), with error $O(h_I^2 + h_J^2)$ we have

$$\operatorname{cov}\{\hat{\eta}_I(x), \hat{\eta}_J(y)\} = \eta_{I\cap J}(x)K^{(2)}(x-y, h_I, h_J) - Nf(x)f(y)g_I(x)g_J(y). \tag{A4}$$

Based on the expansion

$$\frac{\hat{N}(x)}{Nf(x)} = 1 + \left\{\frac{\hat{\eta}_1(x)}{\eta_1(x)} - 1\right\} + \left\{\frac{\hat{\eta}_2(x)}{\eta_2(x)} - 1\right\} - \left\{\frac{\hat{\eta}_{11}(x)}{\eta_{11}(x)} - 1\right\},$$

and from (A4), we have

$$\frac{\operatorname{cov}\{\hat{N}(x), \hat{N}(y)\}}{Nf(x)} = \frac{K^{(2)}(x-y, h_1, h_1)}{g_1(y)} + \frac{K^{(2)}(x-y, h_2, h_2)}{g_2(y)} + \frac{K^{(2)}(x-y, h_{11}, h_{11})}{g_1(y)g_2(y)}$$
$$- \frac{K^{(2)}(x-y, h_1, h_{11})}{g_1(y)}[2] - \frac{K^{(2)}(x-y, h_2, h_{11})}{g_2(y)}[2]$$
$$+ K^{(2)}(x-y, h_1, h_2)[2] - 1 + O(h^2). \tag{A5}$$

Now we are ready to derive $\text{var}(\hat{N})$. Note that $\text{var}(\hat{N}) = \int\int \text{cov}\{\hat{N}(x), \hat{N}(y)\}\, dx\, dy$. Substituting (A5) and noting that $f(x)K^{(2)}(x-y, h_i, h_j)$ is the density function of $X + h_i Z_1 + h_j Z_2$, where Z_1, Z_2 are independent variables with density k, we have

$$\begin{aligned} N^{-1}\,\text{var}(\hat{N}) &= E\{g_1^{-1}(X + h_1 Z_1 + h_1 Z_2)\} + \{g_2^{-1}(X + h_2 Z_1 + h_2 Z_2)\} \\ &\quad + E\{g_1^{-1}(X + h_{11} Z_1 + h_{11} Z_2) g_2^{-1}(X + h_{11} Z_1 + h_{11} Z_2)\} \\ &\quad - 2E\{g_1^{-1}(X + h_1 Z_1 + h_{11} Z_2)\} - 2E\{g_2^{-1}(X + h_2 Z_1 + h_{11} Z_2)\} + 1 + O(h^2) \\ &= E\{g_1^{-1}(X)\} + E\{g_2^{-1}(X)\} + E\{g_1^{-1}(X) g_2^{-1}(X)\} - 2E\{g_1^{-1}(X)\} \\ &\quad - 2E\{g_2^{-1}(X)\} + 1 + O(h^2), \end{aligned}$$

which gives (14).

References

Alho, J. M. (1990). Logistic regression in capture–recapture models. *Biometrics* **46**, 623–35.

Borchers, D. L., Buckland, S. T., Clark, E. D. & Cumerworth, S. L. (1998). Horvitz–Thompson estimators for double-platform line transect surveys. *Biometrics* **54**, 1207–20.

Borchers, D. L., Zucchini, W. & Fewster, R. M. (1998). Mark–recapture models for line transect surveys. *Biometrics* **54**, 1221–37.

Buckland, S. T. (1987). Estimation of minke whale numbers from 1984/85 IWC/IDCR Antarctic sightings data. *Rep. Int. Whaling Commis.* **37**, 263–8.

Buckland, S. T., Anderson, D. R., Burnham, K. P. & Laake, J. L. (1993). *Distance Sampling*. London: Chapman and Hall.

Buckland, S. T., Breiwick, J. M., Cattanach, K. L. & Laake, J. L. (1993). Estimating population size of the Californian gray whale. *Marine Mammal Sci.* **9**, 235–49.

Buckland, S. T. & Turnock, B. J. (1992). A robust line transect method. *Biometrics* **48**, 901–9.

Butterworth, D. S. & Borchers, D. L. (1988). Estimation of $g(0)$ for minke schools from results of the independent observer experiments of the 1985/86 and 1986/87 IWC/IDCR Antarctic assessment cruise, 1978/79. *Rep. Int. Whaling Commis.* **38**, 301–13.

Chao, A. (1989). Estimating population size from sparse data in capture–recapture experiments. *Biometrics* **45**, 427–38.

Chapman, D. G. (1951). Some properties of the hypergeometric distribution with applications to zoological censuses. *Univ. Calif. Pub. Statist.* **1**, 131–60.

Chen, S. X. (1999). Estimation in independent observer line transect surveys for clustered populations. *Biometrics* **55**, 754–9.

Chen, S. X. (2000). Animal abundance estimation for independent line transect surveys. *Envir. Ecol. Statist.* To appear.

Darroch, J. (1958). The multiple recapture census I. Estimation in a closed population. *Biometrika* **45**, 343–59.

Darroch, J. (1961). The two-sample capture–recapture census when tagging and sampling are stratified. *Biometrika* **48**, 241–60.

Huggins, R. M. (1989). On the statistical analysis of capture experiments. *Biometrika* **76**, 133–40.

Jones, M. C. & Rice, J. A. (1992). Displaying the important features of large collections of similar curves. *Am. Statist.* **46**, 140–5.

Lloyd, C. J. (1994). Efficiency of martingale methods in recapture studies. *Biometrika* **81**, 305–15.

Lloyd, C. J. & Yip, S. F. P. (1991). A unification of inference on capture–recapture studies through martingale estimating functions. In *Estimating Functions*, Ed. V. P. Godambe, pp. 65–88. Oxford University Press.

Manly, B. F. J., McDonald, L. L. & Garner, G. W. (1996). Maximum likelihood estimation for the double count method with independent observers. *J. Agric. Biol. Envir. Statist.* **1**, 176–89.

Seber, G. A. F. (1982). *Estimation of Animal Abundance and Related Parameters*, 2nd ed. New York: Hafner.

Wand, M. P. & Jones, M. C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

[*Received October* 1998. *Revised October* 1999]