Statistica Sinica **12**(2002), 505-518

ESTIMATION OF POPULATION SIZE FROM BIASED SAMPLES USING NON-PARAMETRIC BINARY REGRESSION

Song Xi Chen and C. J. Lloyd

National University of Singapore and Australian Graduate School of Management

Abstract: We develop a new estimator of population size when data come from an independent double sampling experiment and at least one continuous covariate for each detection is measured. The new estimator has two features: (i) detection probabilities are estimated by non-parametric smoothing of redetections; (ii) population size is estimated with a Horvitz-Thompson type estimator. Expressions for asymptotic bias and variance are developed. The estimators are shown to be efficient when sampling is unbiased. We provide an illustration on two-stage recapture data on aboriginals in Canada.

Key words and phrases: Biased sampling, kernel regression, local linear estimator, Nadaraya-Watson estimator, wildlife abundance estimation.

1. Introduction

This paper concerns the analysis of population survey data where individuals are heterogeneous and sampling is biased. We begin then with a two-stage model describing these main features. Imagine a finite population of N individuals, each carrying a measurable covariate x of dimension $d \geq 1$. At the first stage, prior to experimental observation, the covariate values $x_1, \ldots, x_N$ are drawn by nature from an unknown density f, sometimes called a 'super-population'. At the second stage, the experimenter draws an independent sample of size n from the population but the probability $g(x)$ of being sampled depends on the covariate x. Let $S \subset \{1, \ldots, N\}$ denote the indices of those individuals who are sampled. The observed values $\{x_k : k \in S\}$ are a random sample on the weighted density $f_w(x) = g(x)f(x)/\int g(x)f(x)dx$. The number sampled, n, follows a binomial distribution with parameter N and $p = \int g(x)f(x)dx = pr(k \in S)$. When g is not a constant function, f_w and f are distinct and the sampling is called biased. Biased sampling is especially common in ecological and public health studies, where being sampled requires some degree of volition on the part of the individual.

The three quantities, N, $f_w(x)$ and p, determine the distribution of the data. However, they are not all estimable from a single sample. This is clear if we consider the effect of dividing the detection function $g(x)$ by a positive constant

c and multiplying N by the same constant c. Then the weighted density $f_w(x)$ for the distribution of the covariates is unchanged. The distribution of the observed number n changes from $\mathrm{Bi}(N, p)$ to $\mathrm{Bi}(Nc, p/c)$, however N and p cannot both be estimated from the single binomial observation. Some information about p, or better still the entire detection function $g(x)$, is required to make N estimable. This requires either further assumptions or further data.

When $g(x)$ is fully specified, the Horvitz-Thompson estimator

$$\hat{N} = \sum_{j=1}^{N} \frac{I_{\mathrm{j}\in\mathrm{S}}}{g(x_j)}. \tag{1}$$

has optimal properties as an estimator of N. In a conventional line-transect survey (LTS) lack of identifiability is overcome by assuming that $f(x)$ is uniform and that $g(0) = 1$, where the covariate x is distance from the transect line to a detected animal. There are also methods available for estimating $f(x)$ under the assumptions that $g(x) \propto x$, so-called length-biased sampling (see Vardi (1982) and Jones (1991)), but these methods do not give an estimator for N. Parametric models for $f(x)$ and $g(x)$ could be assumed, but bearing in mind that these functions are not identifiable, results would be highly model sensitive. Moreover, plausible parametric models for $g(x)$ and $f(x)$ are seldom available in wildlife or public health contexts.

In this paper we consider a sampling design which we call independent double sampling (IDS). This design includes the classical mark recapture experiment (MRE) (Seber(1982)) and the independent observer line transect survey (IOLTS) as introduced by Butterworth and Borchers (1988) and Schweder (1990), who have already pointed out the mathematical equivalence of the two designs. The defining feature of the IDS design is that the same population is sampled twice independently, and that there is some cross-validation mechanism by which population units that appear on both samples can be identified. In MRE cross-validation is done by tagging; in IOLTS by a third judge who cross-validates sightings by two independent observers. Why does taking two samples solve the problem of identifiability? Because, each sample can then be treated as a population *of known size* and the remaining unknown parameters estimated from data on individuals that reappear in the other sample.

We use the term 'detection' to mean that an individual and its covariate value are observed. While the IDS design makes $g(x)$ theoretically identifiable (see Section 2), its estimation remains difficult. Indeed, modeling the detection function in terms of observable covariates remains the central statistical issue. Huggins (1989) and Alho (1990) have advocated logistic linear models for g. Parameters are then estimated by maximizing the likelihood, conditional on detection

at all. Buckland and Turnock (1992) modified the Huggins-Alho approach by replacing the conditional by an unconditional logistic regression of redetections of animals already detected by observer 1. A weakness of this approach is that results depend on which observer is designated "observer 1". Once detection probabilities are known or estimated, population size can be estimated in several ways, for instance by maximum likelihood, using optimally weighted estimating equations (Lloyd and Yip (1990)), or using Horvitz-Thompson type estimators, as suggested by Huggins/Alho. Confidence intervals are best constructed on the log-scale, see Chao (1989).

Huggins/Alho methods have seen little application in the MRE context but have been applied in the IOLTS context. Borchers, Buckland, Goedhart, Clark and Hedley (1998) applied them to an aerial survey of porpoise. One of their findings was that a 'shoulder' in the sighting distances was not easily accounted for by logistic models. They concluded that "it seems likely that, in general, a more flexible form than the logistic will be required for the analysis of double platform LT surveys" and went on to describe some non-linear parametric forms. Kernel smoothing methods have been considered in Chen (1999,2000), but under the extra assumption that $f(x)$ is uniform.

In this paper we introduce a Horvitz-Thompson(HT) type estimator based on non-parametric kernel regression estimators of the probability function g assuming data from an IDS design. The method enjoys several advantages over other approaches: (i) no assumptions about $f(x)$ or $g(x)$; (ii) invariance to relabeling the detection occasions; (iii) flexibility because of its non-parametric nature; (iv) computational simplicity, in contrast to conditional logistic regression methods which require custom maximization; (v) availability of asymptotic expressions for bias and variance. Estimation of f is not covered in this paper, however we note that once $g(x)$ is estimated the methods considered in Jones (1991) may be directly applied.

The plan of the paper is as follows. Details of our notation and estimator are given in Sections 2–3 and asymptotics of the estimator are established in Section 4. Sections 5 and 6 address the important issues of standard errors and bandwidth selection. Sections 7 and 8 present results for both real and simulated data sets and we make some closing conclusions in Section 9.

2. Framework and Notation

In an IDS design, two independent samples are drawn from a population. Appearance in these samples is governed by two possibly different detection functions $g_j(x)$, being the probability that an individual with covariate x appears in sample j. We use lower case subscripts $j = 1, 2$ to denote the two samples and

we equate 3 with 1 and 0 with 2 so that if g_j is one detection function then g_{j+1} is the other.

The IDS design makes the detection functions $g_j(x)$ identifiable. An estimate of $g_1(x)$ can be constructed by considering those individuals who appear in the second sample as a fixed known population, and then observing those who appear in the first sample. Conversely, g_2 is estimated by restricting attention to those who appeared in the first sample and then counting redetections in the second sample.

Each individual sampling unit realizes one of four detection histories. We use an indicator notation for full detection histories, for instance '01' denotes detection in sample 2 only. We also use a single index '1' and '2' to denote detection in sample 1 and 2 respectively. Each history J has five associated quantities: the index set S_J of those individuals with history J; the number n_J of individuals with history J; the probability $g_J(x) = \mathrm{pr}(i \in S_J | x_i = x)$ of history J given x; the unconditional probability $p_J = \mathrm{pr}(i \in S_J)$ of history J; the density $f_J(x)$ of covariate x given history J. The last three quantities are connected according to

$$f_J(x) = g_J(x)f(x)/p_J,\ p_J = \int g_J(x)f(x)dx.$$

The estimator we propose requires several assumptions. First, let $A = \{x \in R^d | f(x) > 0\}$. We assume throughout the paper that

A1: all detections are conditionally independent;
A2: redetections are identified without error;
A3: population size N is constant over the experiment;
A4: $\exists c_0 > 0$ such that $g_j(x) \geq c_0\ \forall x \in A$;
A5: g_j have continuous second derivatives in A;
A6: f is bounded in A and $f'(x)$ is continuous.

Conditions A5, A6 are technical conditions required for deriving asymptotics of our estimator. Condition A4 means that no individual is essentially undetectable. Conditions A1-A3 together imply that n_J has binomial distribution with parameters N and p_J. Condition A1 also implies that $g_{10}(x) = g_1(x)\{1 - g_2(x)\}$, etc. We will use no subscript at all to denote the union of the histories 10, 11, 01, in other words being detected at all. Thus $n = n_1 + n_2 - n_{11}$ is the number of distinct individuals detected, p is the probability of being detected at all and the conditional probability of being detected at all is

$$g(x) = g_1(x) + g_2(x) - g_1(x)g_2(x). \tag{2}$$

We now describe a natural family of estimators of the unknown population size N. If the detection functions g_j can be estimated, say by $\hat{g}_j$, then an estimator of g is $\hat{g}(x) = \hat{g}_1(x) + \hat{g}_2(x) - \hat{g}_1(x)\hat{g}_2(x)$, and an HT type estimator for N

is given by

$$\hat{N} = \sum_{j=1}^{N} \frac{I_{j \in S_1 \cup S_2}}{\hat{g}(x_j)}. \tag{3}$$

In the MRE context, Huggins (1989) and Alho (1990) proposed parametric logistic models for g_j, estimated these from the likelihood conditional on detection at all and then suggested (3) as the estimator of N. A more flexible non-parametric and unconditional approach to the estimation of g_j is the focus of this paper.

An alternative approach to the present problem has been considered by Chen and Lloyd (2000a). In the present notation, they define an index of heterogeneity

$$\alpha = (p_1 p_2)^{-1} \int f(x) g_1(x) g_2(x) dx.$$

They show that the well known Petersen estimator $n_1 n_2 / n_{11}$ is consistent for N/α and go on to estimate α using kernel density estimation of the densities f_1, f_2 and f_{11}. Chao and Tsay (1998) have given a similar expression for α, interpreted it in terms of the coefficient of covariation of the detection functions g_1 and g_2, and pursued estimation using the idea of sample coverage. The approach here is to estimate the detection functions directly and α is simply used as a measure of the amount of heterogeneity present. Neither the density estimation approach nor the detection estimation approach will be superior in general. Our attitude is that good statistical practice involves modeling those process features that are simply and accurately estimated. When the underlying distribution $f(x)$ is highly irregular, as it could be in biological applications, we may still hope that the detection function $g(x)$ would vary regularly with x. In these circumstances the approach we present here would be preferrable.

3. Kernel Estimation of the Detection Functions

In this section, we give details of our proposed non-parametric estimators of $g_j(x)$. We take x to be d-dimensional and $g_j(x)$ to be smooth over all dimensions. Our data comprise a list of n observed covariate values and n associated histories. To define our estimators explicitly, it is convenient to index the observed covariate values by two subscripts, the history J and a simple counting index. Thus $x_{11,7}$ is the covariate value for the 7th individual with history $J = 11$, and $x_{J1}, \ldots, x_{Jn_J}$ is the entire sample of observed covariates with history J. These are a random sample from $f_J(x)$.

Restrict attention to individuals appearing in S_j and label them $1, \ldots, n_j$. Define the n_j binary variables $y_{jk} = I_{k \in S_{11}}$ and denote their collection by $Y_j^T = (y_{j1}, \ldots, y_{jn_j})$. By assumption A1, $\mathrm{pr}(y_{jk} = 1 | k \in S_j, x_k = x) = g_{j+1}(x)$. Therefore, $g_{j+1}(x)$ is the mean function of the binary variables y_{jk} conditional on x_k

and S_j and can be estimated by standard parametric or nonparametric binary regression methods. We investigate the non-parametric approach.

Let K be a d-dimensional bounded probability density function, with compact support on the d-dimensional cube $[-1,1]^d$, that satisfies moment conditions:
A7: $\int uK(u)du = 0,\ \int uu^T K(u)du = \sigma_K^2 I_d$,
I_d being the $d \times d$ identity matrix and σ_K^2 a positive constant. Let h_1, h_2 be two smoothing bandwidths, implying that the same amount of smoothing is used in all directions (when the scales of the covariates are different, they can be standardized by their standard deviations). Let $K_h(u) = h^{-d}K(h^{-1}u)$. Define the matrix of smoothing weights, $\mathbf{W}_j = diag\{K_{h_{j+1}}(x_{ji} - x)\}$ for $i = 1, \ldots, n_j$ and the $n_j \times (d+1)$ linear design matrices

$$\mathbf{X}_j = \begin{pmatrix} 1\, x_{j1} - x \\ \vdots \\ 1\, x_{jn_j} - x \end{pmatrix}.$$

We consider two types of nonparametric regression estimators for g_j: the Nadaraya-Watson (NW) estimator

$$\hat{g}_{NW,j+1}(x) = \frac{\sum_{k=1}^{n_j} y_{jk} K_{h_{j+1}}(x_{jk} - x)}{\sum_{k=1}^{n_j} K_{h_{j+1}}(x_{jk} - x)}, \tag{4}$$

and the local linear (LL) estimator (Ruppert and Wand (1994))

$$\hat{g}_{LL,j+1}(x) = e_1^T (\mathbf{X}_j^T W_j \mathbf{X}_j)^{-1} \mathbf{X}_j^T \mathbf{W}_j \mathbf{Y}_j, \tag{5}$$

where $e_1^T = (1, 0, \ldots, 0) \in R^{d+1}$. A simpler expression is available when $d = 1$, see Wand and Jones ((1995),p119). Substituting these estimated detection functions into (3) defines our estimators $\hat{N}_{NW}$ and $\hat{N}_{LL}$ of the population size.

Both NW and LL estimators are members of a class of non-parametric estimators, called local polynomial estimators, which have been well-studied in the literature; see Härdle (1990) and Fan and Gijbels (1996) for comprehensive reviews.

4. Bias and Variance of the Estimators

The results given in this section rely on standard asymptotic techniques. Further details are available in Chen and Lloyd (2000b). The boundary bias problems depend largely on the boundary region of A which is defined as $\mathcal{B} = \{x \in A : \text{at least one } x_i \text{ is within } h \text{ distance from the boundary}\}$. Without complicating the main scheme of the paper, we assume that

A8: $Pr(x \in \mathcal{B}) = O(h^m), \quad m \geq 2;$
A9: $h_j \to 0$ and $Nh_j^d \to \infty$ as $N \to \infty$.

When $m = \infty$, all covariates are unbounded and $\mathcal{B}$ is the empty set. The case $m \geq 2$ happens when the marginal density of each component of the covariate decays to zero at the boundary. Condition A8 essentially means that all the x can be regarded as interior points without affecting the first two orders of magnitude for bias and variance in the expressions below. These asymptotic expressions hold under the classical assumption about bandwidth decay rates given in A9.

In describing error terms we let h^d denote order d products of h_1 and h_2. For instance terms like h_1^2, h_2^2 or $h_1 h_2$ are all $O(h^2)$. Define $R(t) = \int K(u)K(tu)du$,

$$B_{NWj}(x) = \tfrac{1}{2}\sigma_K^2 tr\{2 \bigtriangledown g_j(x) \bigtriangledown^T f_{j+1}(x) + \bigtriangledown^2 g_j(x) f_{j+1}(x)\} f_{j+1}^{-1}(x),$$

and $B_{LLj}(x) = \frac{1}{2}\sigma_K^2 tr\{\bigtriangledown^2 g_j(x)\}$. In the above expression, $\bigtriangledown$ denotes the vector of first derivatives, $\bigtriangledown^2$ the (Hessian) matrix of second derivatives of a function, and $tr(\cdot)$ the trace of a matrix.

Under Assumptions A1-A9 we have the following expansions for the asymptotic mean and variance of the estimators $\hat{N}_{NW}$ and $\hat{N}_{LL}$. To clarify the formulas, the dummy variables of all integrals are suppressed.

$$\begin{aligned}
\text{Bias}(\hat{N}_{NW}) \;=\; & -N \sum_{j=1}^{2} h_j^2 \int B_{NWj} \frac{\{1 - g_{j+1}\} f}{g} \\
& + R(1) \sum_{j=1}^{2} h_j^{-d} \int \frac{g_j \{1 - g_j\} \{1 - g_{j+1}\}^2}{g^2 g_{j+1}} \\
& + h_2^{-d} R(h_1/h_2) \int 2(1 - g^{-1})^2 + (1 - g)g + o(h^{-d} + Nh^2), \qquad (6) \\
\text{Bias}(\hat{N}_{LL}) \;=\; & -N \sum_{j=1}^{2} h_j^2 \int B_{LLj} \frac{\{1 - g_{j+1}\} f}{g} \\
& + R(1) \sum_{j=1}^{2} h_j^{-d} \int \frac{g_j \{1 - g_j\} \{1 - g_{j+1}\}^2}{g^2 g_{j+1}} \\
& + h_2^{-d} R(h_1/h_2) \int 2(1 - g^{-1})^2 + (1 - g)g + o(h^{-d} + Nh^2) \qquad (7) \\
\text{Var}(\hat{N}_{LL}) \;=\; & \text{Var}(\hat{N}_{NW}) + O(h^{-d} + Nh^2) = N \int \frac{\{1 - g\} f}{g_1 g_2} + O(h^{-d} + Nh^2). \qquad (8)
\end{aligned}$$

Both estimators have very similar bias in the first order. It is also not surprising to see that both estimators have the same leading variance term. This is because the underlying estimators for g_j have different first order variance terms only in $\mathcal{B}$, and the "size" of $\mathcal{B}$ is assumed small under A8. The variance

expression in (8) is identical to the expression given by Chen & Lloyd (2000a) for an estimator based on kernel density estimation. Furthermore, when the g_j are constant functions, so that there is no heterogeneity of sampling, the variance reduces to $N(1-p_1)(1-p_2)/(p_1p_2)$, the asymptotic variance of the Petersen estimator. This means that using the new estimators comes at no cost in asymptotic variance when sampling is random. When sampling is biased, the Petersen estimator is biased while the new estimators are consistent.

When $m = 1$ in A8, the bias of $\hat{g}_{NWj}(x)$ is of order $O(h_j)$ if $x \in B$ which is a larger order than the bias in the interior. However, the order of the bias of $\hat{N}_{NW}$ is unaffected by this boundary bias and still maintains the order of $Nh^2 + h^{-d}$. Continuing with the case $m = 1$, the remainder terms of the variance for both $\hat{N}_{NW}$ and $\hat{N}_{LL}$ are no longer $O(h^{-d} + Nh^2)$, but rather $O(h^{-d} + Nh)$, while the leading term is maintained. This means that when the boundary is not negligible, both estimators are affected in their asymptotic variance.

5. Computing Standard Errors

The asymptotic variance (8) depends on the unknown functions g_1, g_2 and f. There are many ways in which the integral could be estimated but we prefer a completely non-parametric estimator which we derive by observing that

$$\int \frac{(1-g)f}{g_1 g_2} = p_1 \mathrm{E}_1 \left\{ \frac{(1-g(X))}{g_1^2(X) g_2(X)} \right\}, \tag{9}$$

where E_1 denotes expectation with respect to $f_1(x)$. There is a complementary expression involving E_2 instead of E_1. An estimate of (9) is

$$\frac{n_{11}}{n_2} \frac{1}{n_1} \sum_{k \in S_1} \frac{(1-\hat{g}(x_k))}{\hat{g}_1^2(x_k)\hat{g}_2(x_k)}.$$

The leading factor $n_{11}/(n_1 n_2)$ is the reciprocal of $\hat{N}_P$. We use the geometric mean of the complementary estimates as our final estimate of standard error.

6. Choosing the Smoothing Bandwidths

In this section we address the issue of bandwidth selection for the estimator $\hat{N}_{NW}$ and $\hat{N}_{LL}$. For the purpose of establishing asymptotic rates of convergence, we assume that h_1 and h_2 decrease as N increases at the same asymptotic rate. Explicitly we assume that $h_2 = \gamma h_1 = h$ for a positive constant γ. Denote either of the estimators $\hat{N}_{NW}$ or $\hat{N}_{LL}$ by $\hat{N}$. Both (6) and (7) are of the form $\beta_1 Nh^2 + \beta_2 h^{-d}$ where β_1, β_2 are bias coefficients but also involve $\gamma = h_2/h_1$. Combining with the respective variance expression (8), the mean squared error of $\hat{N}$ is

$$MSE(\hat{N}) = N \int \frac{\{1-g\}f}{g_1 g_2} dx + (\beta_1 Nh^2 + \beta_2 h^{-d})^2 + O(Nh^2 + h^{-d}).$$

Differentiating the above we find that the $O(Nh^2 + h^{-d})$ may be ignored, and solving the resulting quadratic for N gives the optimal bandwidth

$$h^\star = \left(\frac{(2+d)|\beta_1\beta_2| - (2-d)\beta_1\beta_2}{4N\beta_1^2} \right)^{1/(d+2)}. \tag{10}$$

For kernel estimation of regression or density functions optimal bandwidths are of order $N^{-1/(d+4)}$. The optimal bandwidth here is of the smaller order $N^{-1/(d+2)}$, essentially because N is an integrated quantity. It is particularly noteworthy that the optimal bandwidth depends only on the bias coefficients β_1, β_2 and not on the leading, or on the next $O(Nh^2)$ term of the variance. Substituting the optimal bandwidth we have $\text{bias}(\hat{N}) = O(N^{d/(d+2)})$ and $\text{Var}(\hat{N}) = O(N) + O(N^{d/(d+2)})$. So

$$\text{MSE}(\hat{N}) = O(N) + O(N^{2d/(d+2)}). \tag{11}$$

The second $O(N^{2d/(d+2)})$ term is larger than $O(N)$ if $d > 2$.

There are difficulties in using (10) to choose the bandwidths h_1 and h_2. For given γ, the plug-in method may be used to derive h_2 according to (10) by estimating f_i'', f and g_i by either fully nonparametric methods or referencing to certain parametric families. This can be quite involved in computation. We may choose γ to be the ratio of the standard deviations of the two design samples, or the ratio of two bandwidths obtained by conventional bandwidth selectors such as the those described below.

Because $\hat{N}$ is an integrated quantity, we expect that results will be quite insensitive to bandwidth choice. This is confirmed in the earlier variance expressions whose leading terms do not depend on h. A simple way of prescribing the smoothing bandwidths is to use bandwidths h_j optimized for estimation of $g_j(x)$. We are particularly attracted to the cross validation or the penalized function approaches (Härdle (1990)) in nonparametric regression as their computation is quite standard and software is readily available. There will be some loss of efficiency as the goal is estimation of N, but the loss should be moderate as accurate estimation of N and g are so intimately bound, see (3).

7. First Nations members in Canada

There is interest in the number of First Nations (i.e. aboriginal) people in a region of Canada. We are limited in the details we can provide because of confidentiality issues. In the previous census the population was estimated at around 10,000 but some local government bodies believe the population could be as high as 30,000. Two surveys were taken. The first, known as the "normal questionnaire", was administered over the period December 1998 to January 1999

by distributing surveys and conducting interviews at First Nations gatherings. The second, known as the "special housing questionnaire", was administered during February 1999 and comprised records of those applying for or currently residing in public housing. 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 March 1, 1999, our covariate x. The two surveys had similar penetration - $n_1 = 1358$, $n_2 = 1285$ - 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. From these figures we calculate the Petersen estimate $\hat{N}_P = 18,764$. However, it is more usual to use the estimator of Chapman (1951):$\hat{N} = (n_1+1)(n_2+1)/(n_{11}+1)\ -1 = 18591$ with standard error 1810.

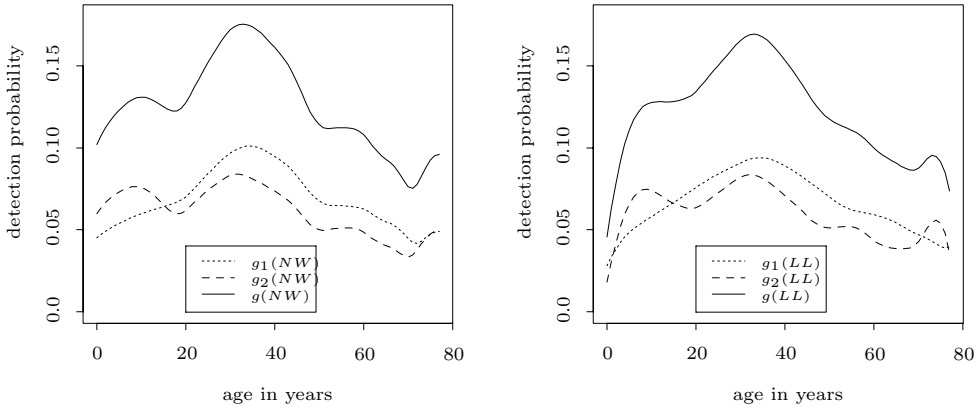

Figure 1. **Estimated detection functions for aboriginal data**. Dotted and dashed lines are respectively the estimated detection functions for the "normal questionaire" g_1 and the "special housing questionaire" g_2. The solid line is for overall detection. The left plot gives estimates based on local averaging (Nadaraya-Watson) and the right plot based on local linear regression. Chosen bandwidths are described in the text.

Figure 1 displays estimates of the detection functions $g_1(x)$ and $g_2(x)$ using both NW and LL smoothers. Further details are given below. The estimate of $g(x)$ is plotted as a solid curve in each case. The estimates of $g_1(x)$ and $g_2(x)$ are highly correlated since they both depend largely on the intersection sample of x-values. Their similar shape does not therefore indicate that $g_1(x)$ and $g_2(x)$ are close. To test the equality of g_1 and g_2 one can test the equivalent hypothesis that $f_{10} = f_{01}$ using a non-parametric test such as the Kolmogorov-Schmirnov two-sample test. For these data, there is indeed strong evidence that the two

detection functions differ (P=0.000). 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 methods is to correct for such bias.

We now list some estimates using the methods recommended in this paper. Using the biweight kernel with selected bandwidths $h_1 = 13.8$ and $h_2 = 12.5$ the estimate $\hat{N}_{NW} = 18912$ with a standard error of 1917.7 is obtained. The corresponding estimates of $g_1(x), g_2(x)$ and $g(x)$ are plotted in the left panel of Figure 1. Typically, increasing both bandwidths will shrink the estimator towards the homogeneous Petersen estimator. For example when $h_1 = h_2 = 15$ the estimate decreases by less than 1%. Further evidence for the insensitivity of the NW estimator to bandwidth is that doubling h_1 and halving h_2 gives the estimator $\hat{N}_{NW} = 19176$. Using selected bandwidths $h_1 = 21.8, h_2 = 13.6$ and the biweight kernel, the local linear estimator is $\hat{N}_{LL} = 20135$ with standard error of 2338.2. The corresponding estimates of $g_1(x), g_2(x)$ and $g(x)$ are plotted in the right panel of Figure 1. Doubling the first bandwidth and halving the second causes a smallish change ($\hat{N}_{LL} = 20294$). As expected from the theory of local linear smoothers, the main difference between the estimated detection functions $g(x)$ using NW and LL smoothers occurs at the boundary, in this context at very young ages, and in regions of high curvature.

8. Simulation Results

We present some simulation results designed to compare empirical outcomes with the theory developed in this paper.

The first study simulated univariate IOLTS with the sighting distance being the only covariate that influences the detection. The real underlying density f for the animal objects was the uniform distribution within $[-w, w]$ where $w = 7$ was the maximum detection distance. We chose $g_1(x) = 0.6Exp\{-|bx|^a\}$, $g_2(x) = 0.7Exp\{-|bx|^a\}$, which are modifications of the usual generalised exponential power series detection functions used in a conventional LTS. We fixed $b = 0.2$ and chose the shape parameters $a = 2.0$, 2.5 and 3.0. Corresponding values of α appear in Table 1.

For our simulation study, the HT estimator in (3) is modified as

$$\hat{N} = \sum_{\hat{g}(x_j)>0.01} \frac{I_{\mathrm{j}\in \mathrm{S}_1 \cup \mathrm{S}_2}}{\hat{g}(x_j)}$$

where 0.01 is a truncation value. The reason for truncating is to avoid the volatility that occurs when estimating the reciprocal of a small probability $g_1(x)$.

When such erratic estimates occur they are always flagged by an appropriately large standard error, however in our simulations we need to control the small number of volatile outcomes. Bandwidths for the simulations have been chosen to optimize estimation of g_1 and g_2 using the penalizing function approach given in Härdle (1990). This is a well-established and easily implemented technology. In each simulation, we generated $N = 500$ or 1000 uniform distributed points, which simulated the positions of a biological population, within a rectangular area with length L and width $2w$. The detection functions g_i given early are used to detect the points to generate the samples. For comparison purposes, the Petersen estimator, $\hat{N}_p$, was also included.

Table 1. Average estimates for N, their standard errors (S.E.) and root mean square error (RMSE): $\hat{N}_p$ - the Petersen estimator, and $\hat{N}_{NW}$ and $\hat{N}_{LL}$, together with average sample size (ave. n) and the average number of data points per simulation (π_{NW} and π_{LL}) where the estimated detection values are less than 0.01.

	$N = 500$			$N = 1000$		
a	2.0	2.5	3.0	2.0	2.5	3.0
α	1.44	1.50	1.55	1.44	1.50	1.55
$\hat{N}_p$	350.1	334.1	321.1	700.0	667.3	642.4
S.E.	22.8	21.8	19.9	33.6	30.3	28.5
RMSE	151.6	167.3	180.0	301.8	334.1	358.7
$\hat{N}_{NW}$	535.2	488.0	443.2	1016.5	924.0	862.2
S.E.	133.2	118.1	104.3	184.9	144.6	134.9
RMSE	137.7	118.7	118.8	185.6	163.4	192.9
$\hat{N}_{LL}$	538.4	494.5	460.2	1050.4	950.6	893.2
S.E.	121.5	106.8	96.1	176.9	134.3	132.5
RMSE	127.4	106.9	104.0	183.9	143.1	170.2
ave. n	150	152	153	300	304	306
π_{NW}	4.0	3.7	3.2	8.9	6.6	6.3
π_{LL}	7.3	6.9	5.8	15.0	13.0	12.4

Table 1 summarizes 500 simulations under each set of conditions. Presented are the simulated average, standard error and root mean square error (RMSE) for each of the estimators considered. To indicate the amount of heterogeneity in the simulation, the α values are listed as well as the average total number of captures/detections. To gauge the effect of truncation on estimation, the average number of data points per simulation where the estimated detection values are below the truncation value of 0.01 is also reported.

The Petersen estimator $\hat{N}_p =: n_1 n_2/n_{11}$ performs poorly, and more poorly for larger α as anticipated by theory; the two nonparametric Horvitz-Thompson estimators have quite similar performance, reflecting the similar theoretical behaviour revealed in Section 4; the local linear based estimates tend to be slighter larger in mean and smaller in variability. When the population size N increases from 500 to 1000, the RMSE of $\hat{N}_p$ doubles in all the cases, whereas those of $\hat{N}_{NW}$ and $\hat{N}_{LL}$ increase by only about 50%. This indicates the performance of $\hat{N}_{NW}$ and $\hat{N}_{LL}$ improves as N increases.

In summary: (i) the effect of α in the performance of the Petersen estimator is confirmed by the simulation; (ii) the Petersen estimator cannot be used when the amount of heterogeneity is severe (the proposed estimators should be used instead); (iii) the new detection-based estimators seem to be relatively more effective when heterogeneity is large.

9. Discussion

Our non-parametric detection function approach is a direct competitor with the Petersen estimator. Theoretical analysis has shown that the new method has MSE $O(N + N^{2d/(d+2)})$ and is therefore consistent for N. The Petersen estimator, on the other hand, has asymptotic mean N/α and is inconsistent. For sufficiently large sample size and heterogeneity α, the new estimator will dominate the Petersen estimator in MSE. Our simulation study has shown superior performance for $\alpha > 1.1$ and $N = 500, 1000$.

Our method is also a direct competitor with Huggins (1989). Both methods are consistent. The new method differs from Huggins' in two respects. First, the detection functions g_1, g_2 are modeled non-parametrically. Second, the detection functions are estimated directly from binary regression of capture on one sample conditional on detection in the other sample, rather than by maximizing a conditional likelihood.

For the aboriginal data that we analyzed, the gender covariate was also available. We have analyzed genders separately and aggregated the results, producing virtually identical estimates to those reported here where we have ignored gender. However, in general, all covariates that affect the detection functions should be included. Logistic regression methods are attractive mainly because they avoid the curse of dimensionality by making an additivity assumption. We have described a fully non-parametric method which makes no additivity assumption. Of course, for dimensions larger than 2 or 3 we may be forced to assume additivity and replace the multivariate non-parametric estimators by an additive non-parametric estimator. Alternatively pre-testing of the data, and background knowledge, may allow us to reduce the number of covariates involved before applying the nonparametric estimators proposed in this paper. These are topics for future research.

References

Alho, J. M. (1990). Logistic regression in capture-recapture models. *Biometrics* **46**, 623-635.

Buckland, S. T. and Turnock, B. J. (1992). A robust line transect method. *Biometrics* **48**, 901-909.

Borchers, D. L., Buckland, S. T., Goedhart, P. W., Clark, E. D. and Hedley, S. L. (1998). Horvitz-Thompson estimators for double-platform line transect surveys. *Biometrics* **54**, 1207-1220.

Butterworth, D. S. and Borchers, D. L. (1988). Estimation of $g(0)$ for Minke schools from results of the independent observer experiments on the 1985/86 and 1986/87 IWC/IDCR Antarctic assessment cruise, 1978/79. *Report of the IWC.*

Chao, A. (1989). Estimating population size from sparse data in capture-recapture experiments. *Biometrics* **45**, 427-438.

Chao, A. and Tsay, P. K. (1998). A sample coverage approach to multiple system estimation with application to census undercount. *J. Amer. Statist. Assoc.* **93**, 283-293.

Chen, S. X. (1999). Estimation in independent observer line transect surveys for clustered populations. *Biometrics* **55**, 754-759.

Chen, S. X. (2000). Animal abundance estimation for independent line transect surveys. *Environmental and Ecological Statistics* **7**, 285-299.

Chen, S. X. and Lloyd, C. J. (2000a). A non-parametric approach to the analysis of two stage mark-recapture experiments. *Biometrika* **87**, 633-649.

Chen, S. X. and Lloyd, C. J. (2000b). Estimation of population size from biased samples using non-parametric binary regression. AGSM Working paper, downloadable from www.agsm.unsw.edu.au/~chrisl/papers.html.

Chapman, D. G. (1951). Some properties of the hypergeometric distribution with applications to zoological censuses. *University of California Publications in Statistics* **1**, 131-60.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications.* Chapman and Hall, London.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press. Cambridge.

Huggins, R. M. (1989). On the statistical analysis of capture experiments. *Biometrika* **76**, 133-140.

Jones, M. C. (1991). Kernel density estimation for length biased data. *Biometrika* **78**, 511-519.

Lloyd, C. J. and Yip, P. (1990). A unification of inference on capture recapture studies through martingale estimating functions. In *Estimating Functions* (Edited by V. P. Godambe), 65-88.

Ruppert, D. and Wand, M. P. (1994). Multivariate Locally Weighted Least Squares Regression. *Ann. Statist.* **22**, 1346-1370.

Seber, G. A. F.(1982). *The Estimation of Animal Abundance.* Griffin, London.

Schweder, T. (1990). Independent observer experiments to estimate detection functions in line transect surveys of whales. *Report of the International Whaling Commission* **40**, 349-355.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *Ann. Statist.* **10**, 616-620.

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

Department of Statistics and Applied Probability, National University of Singapore, Singapore 117543.

E-mail: stacsx@nus.edu.sg

Australian Graduate School of Management, NSW 2052, Australia.

E-mail: chrisl@agsm.unsw.edu.au

(Received June 2000; accepted May 2001)