

Capture-recapture models with matching uncertainty

Andrea Tancredi

Abstract In this paper we illustrate a Bayesian hierarchical modelling approach for estimating the size of a closed population by capture-recapture models when the number of recaptured individuals is unknown but unit level data are available. We will consider both the case of categorical data and the case of multivariate normal data. We will show how these models can be profitably adopted also in record linkage problems providing a new modelling approach in this context.

Key words: Gibbs sampler, population size estimation, record linkage.

1 Introduction

This paper is a contribution to the methodology of capture-recapture models when the size of a finite population is the main object of interest and the number of "recaptured" individuals is unknown because a perfect matching between the units observed across the samples is not possible. Most of the current approaches to population size estimation with matching uncertainty, especially in official statistics and epidemiological applications, see for example [2], consider the matching and the size estimation as two logically well separated steps. Loosely speaking, firstly a preliminary probabilistic record linkage step is used to estimate the number of common individuals observed in different samples drawn from the same population. Then, such estimates are plugged-in inside capture-recapture models to estimate the size of the population to which data refer.

The framework outlined above, in our opinion, may be criticized for the following reasons. Firstly, the matching uncertainty associated to the record linkage

Andrea Tancredi

Dipartimento di Studi geoeconomici, Linguistici, Statistici e Storici per l'Analisi Regionale, Sapienza Università di Roma, Via del castro Laurenziano 9 00161, Roma, e-mail: andrea.tancredi@uniroma1.it

process is not accounted for both for estimating the population size and for estimating the parameters of whatever statistical model fitted to the linked data. At this regard, [4] discusses the general problem of inference with integrated data and [6] face the problem of regression with linked data. Other remarkably recent exceptions, proposed in the field of ecological statistics, are [7] and [11] where genotype misidentification is embedded into multiple mark-recapture models for estimating animal abundance using DNA samples. Secondly, the definition itself of the record linkage statistical procedure by which matching among units should be performed is troublesome. The standard practice, see for example [3] and [5], is to consider, for each pair of different samples units, a comparison vector measuring the agreement for each observed variable. The information provided by these comparison vectors is used to perform the record linkage step. Anyway, these comparison vectors are considered independent and identically distributed random variables without checking the validity of such assumption and the possible misspecification effects when the original observations are drawn from a finite population or are extracted from census or administrative data.

In this paper we illustrate a general alternative Bayesian strategy which may overcome the drawbacks discussed above. In the next section we illustrate the standard two-stage capture-recapture model which can be considered the starting point for our modelling approach. In Section 3, information about the number of recaptures is substituted by unit level categorical data and a Gibbs sampler algorithm to simulate the posterior distribution of the model parameters is illustrated. A more complex model accounting for measurement error and matching estimation is also briefly discussed. Finally, in Section 4 we consider the case of multivariate normal data.

2 The two-stage capture-recapture model with known and unknown recaptures

The standard two-stage capture-recapture model, which considers two capture periods and a marking stage, can be formulated as follows. Let n^A and n^B be the number of individuals sampled in two different occasions and let T be the number of individuals captured in both the experiments. Moreover let N be the population size and let p be the capture probability for all the population individuals. Considering a closed population and independence between individuals and between captures, the two stage capture-recapture model assumes

$$\begin{aligned} n^A | N, p &\sim \text{Binomial}(N, p) \\ T | n^A, N, p &\sim \text{Binomial}(n^A, p) \\ n^B | n^A, T, N, p &\sim T + \text{Binomial}(N - n^A, p) \end{aligned}$$

see, for example, [9]. An alternative formulation, which will be helpful in the following Section, is

$$\begin{aligned}
n^B | n^A, T, N, p &\sim T + \text{Binomial}(N - n^A, p) \\
n^A | T, N, p &\sim T + \text{Binomial}(N - T, p(1 - p)/(1 - p^2)) \\
T | N, p &\sim \text{Binomial}(N, p^2)
\end{aligned}$$

Note also that n^A and n^B given N and p are independent binomial random variables both with size N and probability p . If we choose the prior distribution $\pi(N, p) \propto 1/N$, that is N and p prior independent with p uniform and $p(N) \propto 1/N$, the posterior distribution is

$$\pi(N, p | n^A, n^B, T) \propto \binom{N}{n^A + n^B - T} p^{n^A + n^B} (1 - p)^{2N - n^A - n^B} N^{-1} \quad N \geq n^A + n^B - T. \quad (1)$$

which can be easily simulated by a Gibbs sampler algorithm. When T , the number of recaptured individuals is unknown, the same model produces the following posterior distribution for N and p

$$\pi(N, p | n^A, n^B) \propto \binom{N}{n^A} \binom{N}{n^B} p^{n^A + n^B} (1 - p)^{2N - n^A - n^B} N^{-1} \quad N \geq \max\{n^A, n^B\}. \quad (2)$$

Note that the posterior distribution (2) can be simulated via a Gibbs sampler algorithm with data augmentation where the missing value T is simulated at each step of the algorithm. Specifically, in this case, the Gibbs algorithm proceeds by simulating the following conditional distributions

$$\begin{aligned}
T | N, p, n^A, n^B &\sim \text{Hypergeometric}(n^A, N - n^A, n^B) \\
p | T, N, n^A, n^B &\sim \text{Beta}(n^A + n^B + 1, 2N - n^A - n^B + 1) \\
N | T, p, n^A, n^B &\sim n^A + n^B - T + \text{Negative Binomial}(n^A + n^B - T, 1 - (1 - p)^2)
\end{aligned}$$

The first two lines of Figure 1 show the posterior distributions (1) and (2) when $n^A = 78$, $n^B = 69$ and $T = 54$ and the true values for N and p are respectively 100 and 0.75. Note that assuming T unknown produces, as expected, posterior distributions with a greater variability.

3 Unknown recaptures and categorical variables

Now suppose that, in addition to the sample sizes n^A and n^B , in the two experiments we observe also data at the individual level. In this Section we show how these data can be used to reduce the uncertainty caused by the lack of information about the number of recaptures. Specifically we assume that the data comprise two record configurations x^A and x^B of different sizes n^A and n^B with

$$x^A = (x_1^A, \dots, x_a^A, \dots, x_{n^A}^A)' \quad \text{and} \quad x^B = (x_1^B, \dots, x_b^B, \dots, x_{n^B}^B)'$$

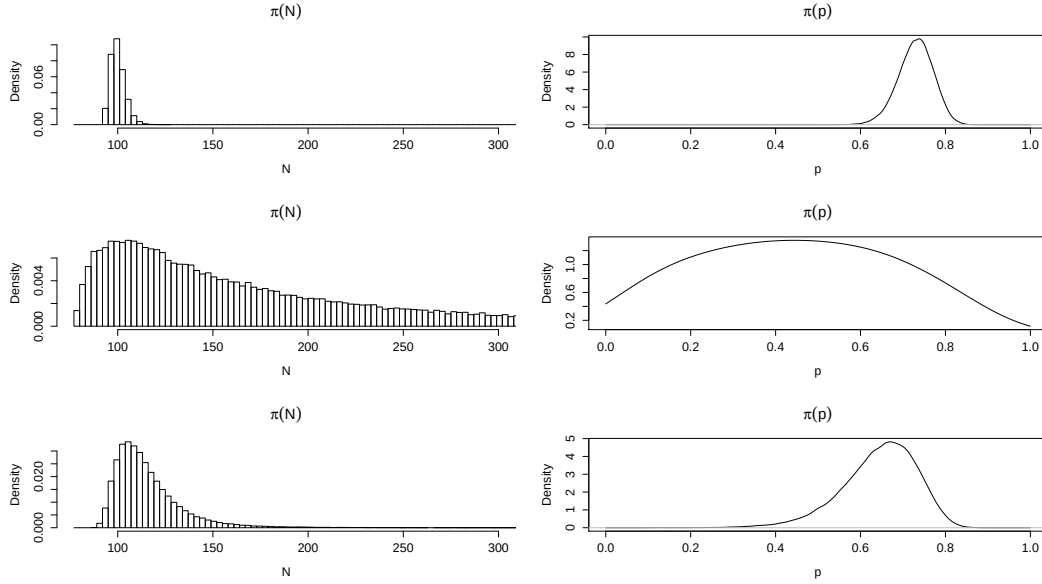


Fig. 1 Marginal posterior distributions for N and p . The true values are $N = 100$ and $p = 0.75$. The first line corresponds to the posterior (1) with $n^A = 78$, $n^B = 69$ and $T = 54$; the second line to the posterior (2) with $n^A = 78$, $n^B = 69$; the third line to the posterior of the model discussed in Section 3 with $n^A = 78$, $n^B = 69$, $k = 50$ and $\theta_j \propto j$ for $j = 1, \dots, k$

where $x_a^A = (x_a^{A_1}, \dots, x_a^{A_h})$ and $x_b^B = (x_b^{B_1}, \dots, x_b^{B_h})$ are the observed values of a categorical random vector $x = (x^1, \dots, x^h)$ whose support is the set

$$V = \left\{ v_{j_1 j_2, \dots, j_h} = (v_{j_1}^1, v_{j_2}^2, \dots, v_{j_h}^h) \quad j_1 = 1, \dots, k_1; \dots; j_h = 1, \dots, k_h \right\}.$$

The elements of the set V arranged in lexicographic order will be indicated with v_j for $j = 1, \dots, k = k_1 \cdot k_2 \cdot \dots \cdot k_h$. Moreover, let $f^S = (f_1^S, \dots, f_j^S, \dots, f_k^S)$, $S = A, B$, be the observed true sample counts for each element $v_j \in V$ and let $F = (F_1, \dots, F_j, \dots, F_k)$ be the vector with the corresponding counts at the population level. Note that $\sum_{j=1}^k F_j = N$ and $\sum_{j=1}^k f_j^S = n^S$

In principle, we can model the observed quantities f^A, n^A, f^B, n^B by assuming that given F and p , the quantities f^A, n^A observed in the first experiment are independent of the quantities f^B, n^B observed in the other experiment. Moreover we may assume

$$\begin{aligned} f^B | n^B, F &\sim \text{Multivariate Hypergeometric}(F_1, \dots, F_k; n^B) \\ n^B | N, p &\sim \text{Binomial}(N, p) \\ f^A | n^A, F &\sim \text{Multivariate Hypergeometric}(F_1, \dots, F_k; n^A) \\ n^A | N, p &\sim \text{Binomial}(N, p), \end{aligned}$$

where the probability mass function of a multivariate hypergeometric random variable $Y = (Y_1, \dots, Y_k)$ with parameters $N_1, \dots, N_k$ and n is

$$p(y_1, \dots, y_k) = \frac{\prod_{j=1}^k \binom{N_j}{y_j}}{\binom{\sum_{j=1}^k N_j}{n}}$$

with $\max\{0, n + N_j - \sum N_l\} \leq y_j \leq \min\{n, N_j\}$ and $\sum y_j = n \leq \sum N_j$. Finally, at the prior level, we may take

$$F|N, \theta \sim \text{Multinomial}(N, \theta)$$

where $\theta = (\theta_1, \dots, \theta_k)$, with $0 \leq \theta_i \leq 1$ and $\sum_{i=1}^k \theta_i = 1$, gives the probabilities for the entries of the contingency table V at the super-population level. For this vector we assume a Dirichlet distribution with parameters $\alpha_j = 1$ for $j = 1, \dots, k$ and, as in the previous section, we take $\pi(N, p) \propto 1/N$.

An alternative way to write the model for the observed quantities f^A, n^A, f^B, n^B outlined above is based on the use of the unknown vector $t = (t_1, \dots, t_j, \dots, t_k)$ denoting, for each element of V , the number of recaptured individuals having v_j as the observed record. Note that $\sum_{j=1}^k t_j = T$ where T is again the unknown number of individuals sampled in both the occasions. By means of the vector t we can write the model in the following way

$$\begin{aligned} f^B|n^B, f^A, n^A, t, p, F &\sim t + \text{Multivariate Hypergeometric}(F_1 - f_1^A, \dots, F_k - f_k^A; n^B - T) \\ n^B|f^A, n^A, t, p, F &\sim T + \text{Binomial}(N - n^A, p) \\ f^A|n^A, t, p, F &\sim t + \text{Multivariate Hypergeometric}(F_1 - t_1, \dots, F_k - t_k; n^A - T) \\ n^A|t, p, F &\sim T + \text{Binomial}(N - T, p(1 - p)/(1 - p^2)) \\ t|T, p, F &\sim \text{Multivariate Hypergeometric}(F_1, \dots, F_k; T) \\ T|p, F &\sim \text{Binomial}(N, p^2). \end{aligned}$$

It is easy to see that, by averaging out over t , one reobtains the model given above directly on the quantities f^A, n^A, f^B, n^B . However, the specification of the model via the vector t is helpful because it offers the opportunity to simulate the posterior distribution by a Gibbs sampler algorithm. In fact, straightforward calculations show that

$$\begin{aligned} t_j|f^B, n^B, f^A, n^A, p, F, \theta &\sim \text{Hypergeometric}(f_j^A, F_j - f_j^A, f_j^B) \quad j = 1, \dots, k \\ p|f^B, n^B, f^A, n^A, t, F, \theta &\sim \text{Beta}(n^A + n^B + 1, 2N - n^A - n^B + 1) \\ \theta|f^B, n^B, f^A, n^A, p, t, F &\sim \text{Dirichlet}(\alpha_1 + F_1, \dots, \alpha_k + F_k) \end{aligned}$$

Moreover, the distribution of F given $N, f^B, n^B, f^A, n^A, t, p, \theta$ is

$$f^A + f^B - t + \text{Multinomial}(\theta, N - n^A - n^B + T)$$

while $N - n^A - n^B + T | f^B, n^B, f^A, n^A, t, p, \theta$ is Negative Binomial with parameters $n^A + n^B - T$ and $1 - (1 - p^2)$. This way, random draw from the conditional distribution of F can be obtained by first simulating N and then simulating F given N .

In the last line of Figure 1 we show the marginal posterior distributions for N and p obtained with the new model and the same samples giving the posterior distributions displayed in the upper part of the Figure. In particular, we have considered a categorical variable with $k = 50$ and $\theta_j \propto j$ for $j = 1, \dots, k$. Note that the posterior distributions are now less concentrated with respect to the case with known T but more concentrated with respect to the case with unknown T . In Figure 2 we show the marginal posteriors for N and p obtained with three different pairs of samples. In the first row the categorical variable has $k = 10$ categories while in the second and third row the categories are respectively $k = 100$ and $k = 500$. It is evident that increasing the number of categories we reduce the posterior uncertainty.

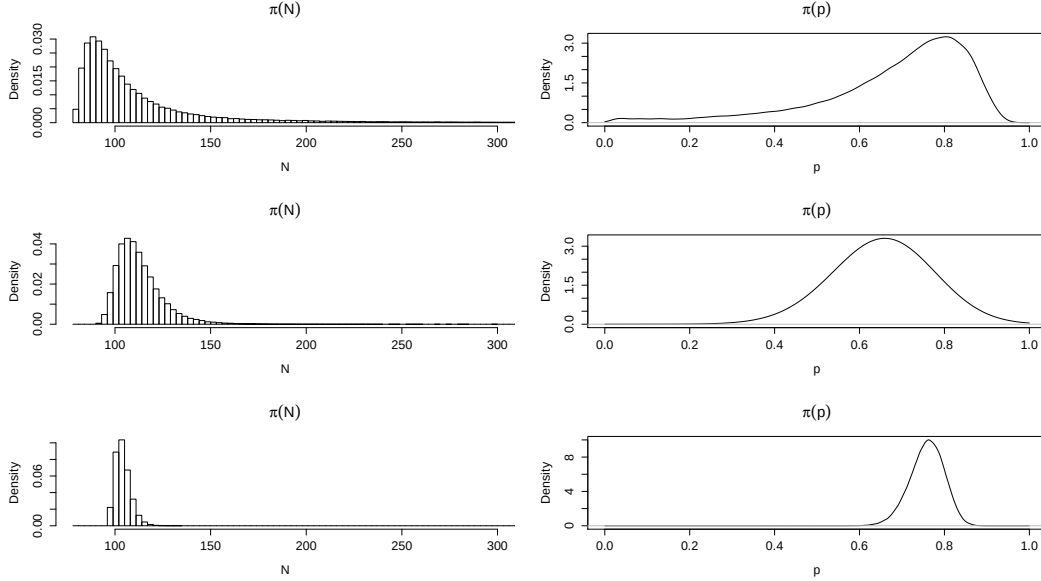


Fig. 2 Posterior distributions for N and p with the model proposed in Section 3. In the first row the categorical variable has $k = 10$ categories while in the second and third row the categories are respectively $k = 100$ and $k = 500$. The true values of the parameters are $N = 100$, $p = 0.75$ and $\theta_j \propto j$

A slight modification of the model proposed in this Section has been discussed by [10]. In particular they fix the samples size n^A and n^B and introduce a measurement error model for the observed values x^S , $S = A, B$. Specifically they assume that each observed variable x_s^S ($s = 1, \dots, n^S$) can be modelled as a mixture of two components: the first component is concentrated on the true value μ_s^S while the second one is uniformly distributed on the set $v^i = \{v_i^1, \dots, v_k^i\}$. The weights β_i for $i = 1, \dots, h$

of each mixture variable are assumed unknown and represent the probability of no measurement error for the observed variables. In [10] it is also explicitly introduced the matching matrix C . This is a $n^A \times n^B$ matrix whose generic element C_{ab} indicates whether or not unit a in sample A and unit b in sample B are the same population unit, that is

$$C_{ab} = \begin{cases} 1, & \text{if } a \text{ and } b \text{ refer to the same population unit} \\ 0, & \text{otherwise.} \end{cases}$$

The matrix C is the actual quantity of interest in record linkage problems. Note also that assuming that multiple matches are not possible gives the following constraints: $\sum_a C_{ab} = C_{.b} \leq 1, \forall b = 1, \dots, n^B$, $\sum_b C_{ab} = C_{a.} \leq 1, \forall a = 1, \dots, n^A$. Also, note that there are $\binom{n^A}{T} \binom{n^B}{T} T!$ different C matrices with exactly $T = \sum_{ab} C_{ab}$ matches, for all $T \leq \min(n^A, n^B)$. Inference on the matching C can be obtained by embedding it in a hierarchical model similar to that described in this Section and assuming that conditionally on $T = \sum_{ab} C_{ab}$, the matrix C has a uniform distribution on the set of all possible matching matrices with exactly T matches. However, see [10] for the complete details regarding the model, the simulation of the posterior distribution for the unknown quantities including the matching matrix C and a small application with real data.

4 Unknown recaptures and multivariate normal variables

In this Section we show how the model proposed by [8] can handle the case of continuous variables observed at the unit level. In particular they assume that a multivariate normal sample is observed in both the occasions. In fact, given the unobserved h -dimensional vectors $\mu_a^A = (\mu_{a1}^A, \dots, \mu_{ah}^A)^t$ and $\mu_b^B = (\mu_{b1}^B, \dots, \mu_{bh}^B)$ and a diagonal matrix $\Gamma = \text{diag}\{\gamma_1, \dots, \gamma_h\}$ it is assumed that

$$x_a^A \sim N_h(\mu_a, \Gamma) \quad \text{and} \quad x_b^B \sim N_h(\mu_b, \Gamma) \quad (3)$$

independently for $a = 1, \dots, n^A$ and $b = 1, \dots, n^B$. The unknown quantities μ_a^A and μ_b^B represent the true values of a multivariate continuous variable $\mu = (\mu^1, \dots, \mu^h)$ assumed by the units of a finite populations which occupy respectively the positions a and b in the two samples. Thus, the vectors x_a^A and x_b^B can be considered measurements subject to recording error of the population variable μ . Note also that, given the unobserved true values of μ , the recording errors are independent across the samples, across the observations of each sample and also between the components of each observation, since Γ is diagonal,

At the population level, the unknown values of the variable μ are assumed independently and identically distributed as a multivariate Normal $N(\theta, \Sigma)$, in the spirit of the Bayesian approach via superpopulation model proposed by [1]. This way we have that $\mu_a^A \sim N(\theta, \Sigma)$ independently for $a = 1, \dots, n^A$ given n^A and

$\mu_b^B \sim N(\theta, \Sigma)$ independently for $b = 1, \dots, n^B$ given n^B . Note, however, that it is not possible to state that μ^A and μ^B are independent. In fact, if we know that the values $(\mu_1^A, \dots, \mu_{n^A}^A)$ are observed in sample A , these values are part of the population values and have a positive probability of being observed also in sample B .

To model the dependence structure between μ^A and μ^B it is possible to make use of the matching matrix C introduced in the previous Section. Given the matching matrix C , the vectors μ_a^A and μ_b^B corresponding to the same unit, that is $C_{ab} = 1$, assume identical values. Moreover, marginally, both μ_a^A and μ_b^B are multivariate Normal with mean θ and covariance matrix Σ . Setting

$$\mu_{ab}^{AB} = \begin{bmatrix} \mu_a^A \\ \mu_b^B \end{bmatrix} \quad \theta_2 = \begin{bmatrix} \theta \\ \theta \end{bmatrix} \quad \text{and} \quad \Sigma_2 = \begin{bmatrix} \Sigma & \Sigma \\ \Sigma & \Sigma \end{bmatrix},$$

it follows that, when $C_{ab} = 1$, we have $\mu_{ab}^{AB} \sim N(\theta_2, \Sigma_2)$. Moreover, note that the dependencies between the components of μ^A and μ^B given C are limited to the matched pairs. Specifically we have that

$$\begin{aligned} p(\mu^A, \mu^B | C, \theta, \Sigma) &= \prod_{a: C_{a, \cdot} = 0} \phi_h(\mu_a^A | \theta, \Sigma) \times \prod_{b: C_{\cdot, b} = 0} \phi_h(\mu_b^B | \theta, \Sigma) \\ &\times \prod_{ab: C_{ab} = 1} \phi_{2h}(\mu_{ab}^{AB} | \theta_2, \Sigma_2) \end{aligned}$$

where $\phi_k(\cdot | \xi, \Omega)$ is the density of a k dimensional multivariate Normal distribution with mean ξ and covariance matrix Ω .

An important feature of the model is the possibility to easily obtain the distribution of x^A and x^B given C, θ, Σ , that is to marginalize the model respect to μ^A and μ^B . In this way, by fixing the values of the sample sizes n^A and n^B , assuming $C|T$ uniformly distributed on the set of all possible matching matrices with T matches, taking T Hypergeometric $[n^A, N_n^A, n^B]$ and finally considering for θ, Σ and Γ respectively a multivariate normal, an inverse Wishart and a product of Gamma variates, [8] are able to simulate the posterior distribution $p(C, N, \theta, \Sigma, \Gamma | x^A, x^B)$ via a Gibbs sampler algorithm with Metropolis-Hastings steps for the matching matrix C and the elements of Γ .

Also in this case, see [8] for the complete details regarding the model, the simulation of the posterior distribution and some large scale simulation results.

References

1. Ericson, W.: Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, Series B*, **31**, 195224 (1969).
2. Herzog, T., Scheuren, F., Winkler, W.: *Data Quality and Record Linkage Techniques*. Springer, New York (2007)
3. Jaro, M.: Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, **84**:414-420, (1989).

4. Judson, D.H.: Information integration for constructing social statistics: history, theory and ideas towards a research program. *Journal of the Royal Statistical Society, Series A*, **170**, 483–501 (2007).
5. Larsen, M.D., Rubin, D.: Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, **96**:32–41 (2001).
6. Lahiri, P., Larsen, M.D.: Regression analysis with linked data. *Journal of The American Statistical Association*, **100**, 222–230 (2005).
7. Link, W.A., Yoshizaki, J., Bailey, L.L., Pollock, K.H.: Uncovering a Latent multinomial: Analysis of Mark-Recapture Data with Misidentification. *Biometrics*. **66**, 178–185 (2009).
8. Liseo, B., Tancredi, A.: Bayesian estimation of population size via linkage of multivariate Normal data sets. Working paper 66. Dipartimento di Studi Geoeconomici Linguistici Statistici e Storici per l'Analisi Regionale (2009)
9. Marin, J. M., Robert, C.P.: *Bayesian Core. A Practical Approach to Computational Bayesian Statistics*. Springer, New York (2007).
10. Tancredi, A., Liseo, B.: A hierarchical Bayesian approach for record linkage and population size estimation. Working paper 65. Dipartimento di Studi Geoeconomici Linguistici Statistici e Storici per l'Analisi Regionale (2009)
11. Wright, J. A., Baker, R. J., Schofield, M. R., Frantz, A. C., Byrom, A. E., Gleeson, D. M.: Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using dna samples. *Biometrics*. **65** 833–840, (2009).