Population size estimated from a random sample of DNA-profiles

Hans Julius Skaug
Institute of Marine Research
Box 1870 Nordnes
5817 Bergen, Norway
skaug@imr.no

1. Introduction

DNA-profiles contain information about biological relationship between individuals. They are, among other things, used to identify putative fathers in parent dispute cases. Skaug (2001) presented an application of DNA-profiles to ecology; estimation of population size N. Assume that n individuals are sampled randomly from a population of size N. Our data are DNA-profiles of these individuals. If sampling of individuals is done with replacement one can, at a later stage, draw a second sample and apply mark-recapture methods. The idea of using DNA-profiles as "tags" is not new. However, when sampling is done without replacement, traditional mark-recapture methods cannot be applied. This problem arises in the analysis of commercial catch data where the sampling method is lethal. The application considered by Skaug (2001) was a DNA-register containing 10 microsatellite markers from minke whales caught by Norwegian whalers. The rationale underlying the "single sample mark-recapture method" of Skaug (2001) is that the degree of relatedness among individuals in the sample is a function of N, and that DNA-profiles provide a measure of relatedness.

The laboratory costs of producing DNA-profiles is currently preventing the method of Skaug (2001) from being applied to large populations, where large sample sizes are required. Costs are also proportional to the number of makers (loci) analyzed per individual. The current paper explores the possibility of using "confirmatory loci" for a subsample of individuals in order to improve the precision of the method. The idea is to first assess all n individuals at S_0 loci, and on basis of these loci, to select pairs of individuals which appear to be related. (In the case of the minke whale DNA-register $S_0 = 10$). The selected individuals are then assessed at S_1 additional loci. We explore by simulation the optimal criterion for selecting the subsample.

2. DNA-profiles and allele-sharing

Consider S loci and denote by $(A_{i,s}^{(1)}, A_{i,s}^{(2)})$ the two alleles individual i holds at locus s. The data on individual i are

$$D_i = \{(A_{i,s}^{(1)}, A_{i,s}^{(2)}), 1 \leq s \leq S\}.$$

For natural populations the joint distribution of $D_1, \ldots, D_n$ is complicated and estimation of N by maximum likelihood does not seem feasible. Instead, Skaug (2001) used a pseudo-likelihood approach involving only the distribution of pairs (D_i, D_j) of DNA-profiles. The pseudo-loglikelihood function is given by

$$l(N) = \sum_{i<j} \log \{p_N(D_i, D_j)\}, \tag{1}$$

where p_N denotes the probability distribution of (D_i, D_j). In the present paper we consider only parent-offspring relationships. Skaug (2001) imposed no such restriction, and allowed individual i and j to have an arbitrary relationship. We define the likelihood ratio statistic

$$L_{ij} = \frac{P(D_i, D_j|\text{parent-offspring})}{P(D_i, D_j|\text{unrelated})}. \tag{2}$$

Using Mendel's law Skaug (2001) derived both an expression for L_{ij} and the approximation

$$l(N) \approx \sum_{i<j} \log \left[1 + N^{-1} 2(\mu_{\mathrm{m}} + \mu_{\mathrm{f}}) \{L_{ij} - 1\}\right].$$

Here μ_{m} and μ_{f} denote the probabilities that the mother and father, respectively, of a randomly sampled individual are alive. In practice, information about the demographic structure of the population must be used to estimate μ_{m} and μ_{f}, but this problem is not addressed in the present paper.

3. Confirmatory loci

Denote by $L_{ij}^{(S)}$ the likelihood ratio (2) based on S loci. The extended sampling scheme is as follows:

1. A sample G_0 of size n is drawn from the population. All individuals in G_0 are assessed at S_0 loci.
2. Pairs (i, j) of individuals with $L_{ij}^{(S_0)} > c$ are assessed at S_1 additional loci. Here c is a threshold that must be specified.

The set of individuals that have been assessed at $S_0 + S_1$ loci is denoted by G_1 $(G_1 \subset G_0)$. How should the new data be incorporated in the pseudo-likelihood? The following lemma shows that we can use the data collected in step 2) as if the set G_1 was determined prior to looking at $\{L_{ij}^{(S_0)}\}$.

Lemma 1 *The pseudo-score function based on (1) remains unbiased (has expectation zero) under the following modification:*

$$p_N(D_i, D_j) = \begin{cases} p_N(D_i^{(S_0+S_1)}, D_j^{(S_0+S_1)}) & \text{if } \{i, j\} \subset G_1, \\ p_N(D_i^{(S_0)}, D_j^{(S_0)}) & \text{otherwise,} \end{cases}$$

where $D^{(S)}$ denotes a DNA-profile based on S loci.

Proof. Let I_{ij} be the indicator function taking the value 1 if $\{i,j\} \subset G_1$, 0 otherwise. Then, the contribution from (i,j) to the pseudo-loglikelihood (1) is

$$I_{ij} \log\left\{p_N(D_i^{(S_0+S_1)}, D_j^{(S_0+S_1)})\right\} + (1 - I_{ij}) \log\left\{p_N(D_i^{(S_0)}, D_j^{(S_0)})\right\}. \qquad (3)$$

We must show that the derivative (w.r.t. N) of this expression has expectation zero. Note that

$$E_N\left[I_{ij}\frac{d}{dN}\log\left\{p_N(D_i^{(S_0+S_1)}, D_j^{(S_0+S_1)})\right\}\right] = \frac{d}{dN}E_N\left(I_{ij}\right), \qquad (4)$$

where E_N denotes expectation taken under the parameter value N. Equation (4) follows from changing the order of differentiation and integration, and if $I_{ij} \equiv 1$ the equation states the well known fact that the score function is an unbiased estimating equation. It follows similarly that the derivative of the last part of (3) has expectation $-d/dN\ E_N\left(I_{ij}\right)$, canceling with (4).

4. Simulations

A simulation study with $N = 100000$, $n = 3000$, $S_0 = 10$ and $S_1 = 15$ were conducted using the simulation model of Skaug (2001). Table 1 shows the mean and the standard deviation of the estimator $\widehat{N}$ for different values of c based on 400 simulation replicas. The last row in the table shows the average number of individuals in G_1. The column '$c = 0$' corresponds to the situation were all n individuals are assessed at $S = 25$ loci (and no confirmatory loci), whereas '$c = \infty$' corresponds to all n individuals assessed at $S = 10$ (no confirmatory loci). It is seen that, by assessing 240 individuals for 15 more loci we are able to reduce the standard deviation of $\widehat{N}$ by 50%.

	$c = 0$	$c = 100$	$c = 1,100$	$c = 5,000$	$c = 10,000$	$c = \infty$
$E(\widehat{N})$	100,000	100,000	101,000	105,000	106,000	108,000
$SD(\widehat{N})$	11,000	11,000	13,000	18,000	23,000	40,000
size(G_1)	3,000	2,700	1,100	240	120	0

5. Discussion

With the current cost level of doing genetic analyses there are not many animal species for which the method of Skaug (2001) is applicable. However, as laboratory costs are likely to go down due to technological improvements, the scope of the method will increase. It has been shown in the present paper that it may be cost efficient to assess a subsample of individuals at a larger number of loci, but in many situations the costs of collecting samples may be much larger than the costs of doing the genetic analysis.

REFERENCES

Skaug, H.J. (2001). *Allele-sharing methods for estimation of population size. Accepted for Biometrics*

RESUME

Les données genétiques devient de plus en plus importantes dans les études écologiques.On présente ici une méthode pour l'estimation d'abondance d'une population, basée sur l'utilisation d'empreintes de DNA obtenues d'un échantillon d'individus pris au hazard. L'idée de base est que les empreintes de DNA peuvent fournir information sur les relations biologiques et que le degrée de relation entre individus dépand de la taille de la population. Les estimations sont basées sur la méthode de pseudo-vraissemblance appliquée sur des comparaisons deux à deux d' individus. Des simulations sont presentées qui montre quelle taille d'échantillon et quel nombre de marqueurs génetiques sont necessaires pour l'utilisation pratique de cette méthode. On présente des resultats préliminaires obtenues de l'application de cette méthode sur des données de petit rorqual de l'Atlantique du Nord.