Abundance Estimation from Multiple Photo Surveys: Confidence Distributions and Reduced Likelihoods for Bowhead Whales off Alaska

Tore Schweder
Department of Economics, University of Oslo, Norway

email: tore.schweder@econ.uio.no

SUMMARY. Maximum likelihood estimates of abundance are obtained from repeated photographic surveys of a closed stratified population with naturally marked and unmarked individuals. Capture intensities are assumed log-linear in stratum, year, and season. In the chosen model, an approximate confidence distribution for total abundance of bowhead whales, with an accompanying likelihood reduced of nuisance parameters, is found from a parametric bootstrap experiment. The confidence distribution depends on the assumed study protocol. A confidence distribution that is exact (except for the effect of discreteness) is found by conditioning in the unstratified case without unmarked individuals.

KEY WORDS: Abundance estimation; Bootstrap; Bowhead whales; Confidence distribution; Mark-recapture; Reduced likelihood; Study protocol.

1. Introduction

Commercial whaling on bowhead whales ceased nearly a century ago. The species was severely depleted in all its previous habitats in the Arctic ocean. Except for the population in the Bering-Chukchi-Beaufort Seas off Alaska, called the bowhead population in this paper, the species has not recovered appreciably. The bowhead population off Alaska has been subject to harvesting by Native Alaskans as “aboriginal subsistence whaling.” The International Whaling Commission decided in May 2002 to discontinue the Alaskan hunt, but a Special Meeting in October 2002 reversed the decision and the Native Alaskans continues their whaling, with a maximum of 67 strikes per year (http://www.iwcoffice.org).

The bowhead population was visually and acoustically surveyed in 1993 (Anonymous, 1999) and again in 2001. The survey in 1993 lead to an abundance estimate of 8200 animals (SE = 564), while George et al. (2002) give a preliminary abundance estimate of 9869 (SE = 1222) based on the last survey. The population was also subject to four photo-surveys in the spring migration off Point Barrow, and in the Beaufort Sea in the autumn, in the two years 1985 and 1986. The data are described in da Silva et al. (2000), who use them to estimate population size based on a stratified multinominal multiple capture model which is an extension of the model of Darroch (1958). Due to the great interest in the status of this population, as reflected by the “in-depth assessment” scheduled for 2004 in the Scientific Committee of the International Whaling Commission, older data like those from the photo-survey deserve continued attention. In the present article I reanalyze these data. This provides an opportunity to present some methodological material on the modeling and analysis of multiple capture data, with special emphasis on confidence distributions and reduced likelihoods. These latter concepts are, in my view, not sufficiently familiar to biologists and statisticians.

A confidence distribution for a scalar parameter, such as the size of a given population of animals, provides confidence intervals by its quantiles. In clear-cut cases, the confidence distribution is identical to the fiducial distribution of Fisher (1930), see Efron (1998) and Schweder and Hjort (2002). Approximate confidence distributions are often found by parametric bootstrapping. A study of such simulation results can reveal a construct which might be used as a pivot leading to a confidence distribution, and also to a reduced likelihood. A pivot is a function of the data and the parameter, which is increasing in the parameter for all possible data, and which has a fixed distribution for all values of the parameter (Barndorff-Nielsen and Cox, 1994).

The confidence distribution and the reduced likelihood serve two different purposes. The confidence distribution is interpreted in the same way as confidence intervals. From the duality between testing and confidence interval estimation, the cumulative confidence distribution function for the parameter θ evaluated at θ₀, C(θ₀), is the p-value of testing H₀: θ ≤ θ₀ against its one-sided alternative. It is thus a compact format of representing the information regarding θ contained in the data and given the model. Before the data have been observed, the confidence distribution is a stochastic element with quantile intervals (C⁻¹(α), C⁻¹(1 - β)) which covers the unknown parameter with probability 1 - α - β. After having observed the data, the realized confidence distribution is not a distribution of probabilities in the frequentist sense, but of confidence attached to interval statements concerning θ. The practitioner is likely to interpret the confidence
distribution in much the same way as a posterior distribution from a Bayesian analysis. The confidence distribution is however a frequentist concept and does not rely on a prior distribution.

The reduced likelihood might not be as readily interpreted as the confidence distribution, but it is useful in meta-analyses and other integrative analyses. The reduced likelihood is often the marginal or the conditional likelihood of the statistic on which the confidence distribution is based, or it might be an approximate likelihood in the interest parameter only, with all nuisance parameters reduced away. Schweder and Hjort (2002) propose that both confidence distributions and reduced likelihoods for parameters of primary interest are routinely reported.

As an example take the interest parameter to be the number of marked mature bowhead whales, $N$. The confidence distribution is calculated from the data in Table 1 by the method developed below. The result is shown in Figure 1 as the confidence histogram in the left panel and a funnel plot of the two-sided confidence $2|C(N) - 0.5|$ in the right panel. Here $C(N)$ is the cumulative confidence at abundance $N$. The funnel points at the confidence median, $\hat{N} = 719$ which is a point estimate of abundance within this stratum. At any degree of two-sided confidence it shows the confidence interval as the horizontal stretch across the funnel. An approximate confidence density based on the normal distribution is added to the left panel. The reduced likelihood, which in this case is a conditional likelihood, is shown in Figure 2.

Figure 1. Confidence distribution for the abundance of marked mature bowheads. Confidence density with dashed approximation (see text) in left panel, and two-sided confidence, $2|C(N) - 0.5|$, in right panel.
2. Data and Notation

The population is stratified by maturity category $s = 0$ (immature), 1 (mature), and mark category $m = 0$ (unmarked), 1 (marked). The number of individuals in maturity category $s$ and mark category $m$ is $N_{sm}$. The goal is to obtain a confidence distribution and a reduced likelihood for the total abundance, $N_{00} + \cdots + N_{11}$. The population is assumed closed, and so are the strata.

The data are described in da Silva et al. (2000). Calves are disregarded in the present study. The basic data unit is a capture. The photo surveys were primarily conducted to study reproduction in bowhead whales rather than abundance. The survey was carried out by aeroplane. At each flight, several pictures were taken of the encountered whales. The encounters were random, with no attempts to revisit areas where encounters had been made (Zeh, personal communication). Some whales have individual marks that mostly are scars from contact with ice or other objects, and are assumed permanent. Each capture is given an identifier, and is represented by a set of photos of acceptable quality. The exact time is recorded for each photo. Captures of marked whales are matched, and the identifiers are modified to represent unique individuals. Based on estimated body length, a capture is coded as being either of an immature or a mature bowhead whale.

In the saturated two-stage Poisson model to be discussed below, there is a separate intensity of initial capture, and also a separate intensity for recaptures within each survey. This calls for the following notation. For maturity stratum $s$ and survey $t$, let $n_t$ be the number of unique captures of marked individuals, $r_t$ the number of recaptures, and $u_t$ the number of captures of unmarked whales. $X_s$ is the number of unique marked individuals of maturity $s$ captured over the four surveys pooled.

The photo survey data are presented in the next section. The confidence distribution and the reduced likelihood discussed in the previous paragraph are developed in Section 3. A model for multiple captures of marked and unmarked individuals in a stratified population is developed and fitted to the data in Section 4. The confidence distribution and the reduced likelihood for total abundance will typically depend on assumptions regarding the nuisance parameters. These assumptions should be based on the study protocol, which has not been available in the present case. Results for two hypothetical protocols are given, and they do indeed differ. The final section provides some discussion, and also sketches a model that allows for individual heterogeneity in capture intensity. A short appendix presents the essential of the multiple capture model of Darroch (1958).
Assuming a common encounter process for marked and unmarked whales within stratum and survey, the multiplicity distribution is estimated from capture histories of marked whales. To find the number of captures for each of the marked whales, a maximal capture period of 3 hours was identified as follows. There are 986 unmarked captures with multiple photos and with time span less than 10 hours from the first to the last photo. The remaining 39 unmarked captures have much longer time span, and are identified by other means than visual features defining proper marks. Two of the 986 unmarked captures had photographs taken more than 3 hours apart, while the remaining 984 show a time span distribution that tapers off with a long but contiguous tail. From this, I conclude that photos taken more than 3 hours apart represent different captures. Applying the three hour rule to photos of marked whales, the numbers of recaptures within surveys are calculated. Only one marked whale (number 85009112) was captured more than twice in a survey. It was actually captured 3 times in the spring of 1985.

Let \( V_{ts} \) be the time of the 3\textsuperscript{rd} initial capture of a marked individual of maturity \( s \) in survey \( t \). \( V_{ts} = \sum_{j=1}^{n_{ts}} v_{ts,j} \). Time is measured in abstract operational time, in which the initial capture intensity is constant. Data on effort are not available. Since survey conditions must have varied over time it is hardly realistic to assume that observed capture times represent operational times. The observed times are therefore disregarded, and the exposure times for recaptures within survey are treated as a latent variables with distribution depending on initial capture intensity, \( \lambda_{ts} \). If, however, \( \{V_{ts}\}_{ts} \) were observed, the statistics \( n_{ts}, V_{ts}, r_{ts}, u_{ts}, t = 1, \ldots, 4, \) and \( X \) form a sufficient set for the intensities and abundances within maturity stratum \( s \). My analysis is therefore based on \( n_{ts}, r_{ts}, u_{ts}, t = 1, \ldots, 4, \) and \( X, s = 0, 1 \). Theses data are presented in Table 1.

3. Confidence Distribution and Reduced Likelihood for a Single Population of Marked Individuals

We are now in the multiple recapture model of Darroch (1958) that is sketched in the Appendix. Suppressing the indices representing mark category and maturity, let \( N \) be the number of individuals, and let the number of unique individuals captured over the four surveys be \( X \). From the correspondence between confidence intervals and hypothesis testing, the cumulative confidence \( C(N_0) \) is for each value of \( N_0 \) obtained as the p-value when testing the hypothesis \( H_0: N \leq N_0 \) versus \( N > N_0 \). The test is conditional given \( n = (n_1, \ldots, n_k) \), as argued below.

Due to the discreteness of the statistic \( X \), half-correction is used in the p-value:

\[
C(N) = P_N(X > x_{obs} \mid n) + \frac{1}{2} P_N(X = x_{obs} \mid n). \tag{1}
\]

For observed data such as those given in Table 1 for marked mature whales, \( C(N) \) is calculated for a number of values of \( N \) by (5) in the Appendix.

For a given value of \( N, n \) is a sufficient statistic for \( p = (p_1, \ldots, p_k) \). Since \( n \) is binomially distributed and independent over the four surveys, the family of distributions of \( (X, n) \) is complete in the parameter \( (N, p) \) (Cox and Hinkley, 1974), see the Appendix. The likelihood ratio \( L(N_1, p) / L(N_2, p) = (N_1^N) / (N_2^N) \) is increasing in \( X \) when \( N_1 > N_2 \). Thus, the Neyman-Pearson lemma can be invoked and the confidence distribution for \( N \) based on the conditional distribution of \( X \) given \( n \) is (almost, due to discreteness) uniformly optimal, e.g., the confidence distribution is less dispersed in a stochastic sense than confidence distributions based on other statistics (Schweder and Hjort, 2002).

By some trial, I found the normal confidence score, \( \Phi^{-1}(C(N)) \), to be practically linear in \( 1/(N)^{1/2} \) for mature whales as well as for immatures. An empirical equation \( \Phi^{-1}(C(N)) \approx a + bN^{-1/2} \) is thus obtained by simply regressing \( \Phi^{-1}(C(N)) \) on \( N^{-1/2} \). By writing \( \hat{\tau} = b^{-1} \) and \( \hat{N} = (b/a)^2 \),

\[
C(N) \approx \Phi((\hat{N}^{-1/2} - N^{-1/2})/\hat{\tau}), \tag{2}
\]

where \( \Phi \) is the cumulative normal distribution function. \( \hat{N} \) is here the median unbiased estimate of \( N \), and \( \hat{\tau} \) is a scale estimate.

The pivot behind (1) that leads to (2) is piv = \((\hat{N}^{-1/2} - N^{-1/2})/\hat{\tau}\). It is nearly exact, with a standard normal pivotal distribution. Disregarding the sampling variation in \( \hat{\tau} \), it does also lead to the reduced log-likelihood \( l(N) = -1/2[\Phi^{-1}(C(N))]^2 \), called the normal scores log-likelihood. It provides an excellent approximation, as shown by Figure 2. See Schweder and Hjort (2002) for more discussion of pivots and reduced likelihoods.

Note that the reduced log-likelihood is different from the log of the confidence density. The difference is \(-3/2 \log(N)\), and is due to the Jacobian that arises when the confidence density is obtained by differentiating the cumulative confidence distribution function. A Bayesian learning about the confidence distribution of \( N \), would presumably use the confidence density as his prior density to be combined with some other independent data at hand. He would then multiply this density with the likelihood of the new data, and normalize. A Fisherian, however, would treat the new and the old data on the same footing, and obtain a joint likelihood by multiplying the reduced likelihood of \( N \) with the likelihood of the new data. Since the confidence density is different from the reduced likelihood (because of the Jacobian) the Bayesian and the Fisherian will arrive at different conclusions.

4. Multiple Captures of Marked and Unmarked Individuals in Stratified Populations

4.1 A Model for the Encounter Process

Each individual in stratified population \( s \) is initially captured in survey \( t \) according to a Poisson process with intensity \( \lambda_{ts} \). The time for the Poisson process will typically be an unknown nonlinear transform of clock time, according to unobserved variation in survey effort and observational conditions. Each survey is assumed to last 1 unit of observational time. The capture probabilities are thus

\[
p_{ts} = 1 - \exp(-\lambda_{ts}).
\]

For individuals that are initially captured in a survey, there is a subsequent Poisson process of recaptures with intensity \( \sigma_{ts} \lambda_{ts} \) in the remaining operational time of that survey. The times of initial captures \( V_{ts} \) have density \( f_{obs}(v) = \lambda_{ts} \exp(-\lambda_{ts} v) / p_{ts} \quad 0 < v < 1 \). Given that an initial capture
occurred at $V_{ts}$, the number of recaptures of this individual within this survey is Poisson distributed with mean $\sigma_{ts}\lambda_{ts}(1 - V_{ts})$.

The parameter $\sigma_{ts}$ is introduced to allow the recapture intensity to differ from the initial capture intensity. Log-linear models are suggested for $\sigma_{ts}$ and $\lambda_{ts}$. Since recapture intensity is nearly proportional to initial capture intensity, a simple model is sufficient for $\sigma_{ts}$. Model selection is done by a standard deviance analysis, and models are fitted by maximizing the likelihood.

Individuals are initially captured and recaptured within surveys and between surveys independently, and independently of each other. Consequently, the number of marked captures, $n_{ts}$, is binomially distributed with parameters $N_{ts}$ and $p_{ts}$. Given $n_{ts}$, the total (unobserved) exposure time for marked recaptures, $W_{ts} = \sum_{j=1}^{n_{ts}}(1 - V_{jts})$, has mean $EW_{ts} = n_{ts}(p_{ts}^{-1} - \lambda_{ts}^{-1})$ and variance $\text{var}(W_{ts}) = n_{ts}(\lambda_{ts}^{-2} - (1 - p_{ts})p_{ts}^{-2})$, and the number of marked recaptures, $r_{ts}$, is conditionally Poisson distributed with mean $\sigma_{ts}\lambda_{ts}W_{ts}$. Approximating the conditional distribution of $W_{ts}$ by a gamma distribution with the two first moments given above, the conditional distribution of $r_{ts}$ given $n_{ts}$, but with $W_{ts}$ integrated out, is approximated by the negative binomial distribution with density

$$f(r; \alpha_{ts}, \beta_{ts}, \sigma_{ts}, \lambda_{ts}) = \left(\frac{\alpha_{ts} + r - 1}{r}\right)\left(\frac{\beta_{ts}}{\sigma_{ts}\lambda_{ts} + \beta_{ts}}\right)^{\alpha_{ts}} \\ \times \left(\frac{\sigma_{ts}\lambda_{ts}}{\sigma_{ts}\lambda_{ts} + \beta_{ts}}\right)^{r} \quad r = 0, 1, \ldots, \\
\alpha_{ts} = n_{ts}(\lambda_{ts} - p_{ts})^2(p_{ts}^2 - \lambda_{ts}^2(1 - p_{ts}))^{-1}, \\
\beta_{ts} = \lambda_{ts}p_{ts}(\lambda_{ts} - p_{ts})(p_{ts}^2 - \lambda_{ts}^2(1 - p_{ts}))^{-1}.$$

The joint likelihood for $(X_s, n_s)$ is given in the Appendix.

It remains to find the likelihood of the numbers of captures of unmarked individuals, $n_{0ts}$, Let $n_{0ts}$ and $r_{0ts}$ be the unobserved counterparts to $n_{ts}$ and $r_{ts}$ for unmarked individuals, and let $n_{ts} = n_{0ts} + r_{0ts}$. By repeated use of conditional expectation and variance,

$$\mu_{ts} = EU_{ts} = EE(u_{ts} | n_{0ts}) = N_{0ts}(\sigma_{ts}\lambda_{ts} + (1 - \sigma_{ts})p_{ts}), \\
\tau_{ts} = \text{var}(u_{ts}) = N_{0ts}\left(\sigma_{ts}\lambda_{ts} + (1 - \sigma_{ts})p_{ts}(1 - p_{ts})\right) \\
\times \left[\left(1 - \sigma_{ts}\right) + 2\sigma_{ts}\lambda_{ts}p_{ts}^{-1} - \sigma_{ts}(1 - p_{ts})^{-1}\right].$$

Since $u_{ts}$ is the sum of $N_{0ts}$ independent components, I approximate its distribution by the normal distribution with moments as given above.

Finally, the (approximate) log likelihood used in fitting the two-stage Poisson model is

$$l = \sum_{s=0,1} \sum_{t=1}^{4} \left[ -\log(\tau_{ts}) - \frac{1}{2} \left(\frac{u_{ts} - \mu_{ts}}{\tau_{ts}}\right)^{2} \right] \\
+ \sum_{s=0,1} \sum_{t=1}^{4} \log f(r_{ts}; \alpha_{ts}, \beta_{ts}, \sigma_{ts}) \\
+ \sum_{s=0,1} \left\{ \log\left(N_{ts}\right) \right\} \\
+ \frac{1}{4} \sum_{t=1}^{4} n_{ts} \log(p_{ts}) + (N_{ts} - n_{ts}) \log(1 - p_{ts}) \right\}.$$

Log-linear models in the intensity parameters $\lambda$ and $\sigma$ are fitted to the data in Table 1. The results for the more interesting models are shown in Table 2. With $A$ representing maturity status, $C$ year, and $D$ season, denote by $\lambda \sim A \times C \times D$ the full factorial model for $\lambda$, $\sigma \sim A + D$ the model with main effects in both $A$ and $D$ for $\sigma$ etc. From Table 2, model 6 is chosen. Note that the abundance estimates are not wildly dependent on the model, except when recapture intensity within surveys are equal to initial capture intensities, $\sigma \sim 0$. Recapture intensities are however significantly different from initial capture intensities within survey.

Table 3 shows maximum likelihood estimates for abundances and encounter intensities, with standard errors obtained from a simple parametric bootstrapping for the latter. Standard errors are not given for abundances due to bias in estimates and skewness in confidence distributions. The recapture intensities differ significantly from the initial capture

<table>
<thead>
<tr>
<th>Model</th>
<th>Log-linear model structure</th>
<th>$\lambda$</th>
<th>$\sigma$</th>
<th>Immatures</th>
<th>$M$</th>
<th>$U$</th>
<th>Matures</th>
<th>$M$</th>
<th>$U$</th>
<th>Dev</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$A \times C \times D$</td>
<td>$A \times D + C$</td>
<td>284</td>
<td>2776</td>
<td>676</td>
<td>1533</td>
<td>(340.3)</td>
<td>(17)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>$A \times C \times D$</td>
<td>$A + D + C$</td>
<td>283</td>
<td>2757</td>
<td>675</td>
<td>1529</td>
<td>0.8</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>$A \times C \times D$</td>
<td>$A + D$</td>
<td>282</td>
<td>2755</td>
<td>675</td>
<td>1532</td>
<td>2.8</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>$A \times C \times D$</td>
<td>1</td>
<td>342</td>
<td>3322</td>
<td>605</td>
<td>1406</td>
<td>9.8</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>$A \times C \times D$</td>
<td>0</td>
<td>214</td>
<td>2077</td>
<td>293</td>
<td>910</td>
<td>19.0</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>$A \times C + A \times D + C \times D$</td>
<td>$A + D$</td>
<td>282</td>
<td>2753</td>
<td>675</td>
<td>1531</td>
<td>2.8</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>$A \times D + C \times D$</td>
<td>$A + D$</td>
<td>331</td>
<td>3215</td>
<td>652</td>
<td>1518</td>
<td>26.8</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
intensities. The estimated ratio is for immatures in autumn surveys $13.5 = \exp(0.98 + 1.62)$.

4.2 Study Protocols and Resulting Confidence

Distributions and Reduced Likelihoods

Relative to the interest parameter total abundance, $N = N_{00} + \cdots + N_{11}$, there are a number of nuisance parameters in the model. A recurring problem is that exact or good approximate pivots might not be available in situations with nuisance parameters. This is the case in the present situation, where data are rather weak and no recourse can be made to asymptotic theory.

Instead of searching in vain for a construct that is an approximate pivot with respect to all values of the nuisance parameters, I will reduce the ambition and look for an approximate pivot over a subset of the parameter space. It will be based on the maximum likelihood estimator $\hat{N}$. The idea is to let the nuisance parameter vary with the interest parameter, and look for a pivot over an appropriate curve in the parameter space. Let the vector of nuisance parameters be $(\nu, \theta)$. Here, $\nu = (N_{010}, N_{011}, N_{10}/N)$ represents the population structure and $\theta$ represents the sampling process. Assumptions regarding the likely variation in the nuisance parameters when hypothetically varying $N$ should ideally be based on the study protocol or other external information.

For the multiple capture experiment, many protocols are conceivable. The “Petersen” protocol would, for example, be that the numbers of captures are fixed in advance, independent of population size. The “fixed budget” protocol would, on the other hand, be that the survey effort is fixed in advance, and that the numbers of captures thus would increase linearly in $N$. An intermediate protocol specifying the number of captures to grow as $(N)^{1/2}$ leads approximately to constant coefficient of variation in the resulting abundance estimator, and is called the “fixed CV protocol.”

With the actual study protocol unavailable, consider now the fixed cv protocol, and the Petersen protocol. Under both, assume $\sigma_{ts} = \sigma_{ts0}$ and $\nu = \nu^0$. In the fixed CV protocol, assume $\lambda_{ts} = \lambda_{ts0}/(N)^{1/2}$, while under the Petersen protocol, $\lambda_{ts} = \lambda_{ts0}/N$. Here, $\sigma^0_{ts}, \lambda^0_{ts}$ and $\nu^0$ are the observed maximum likelihood estimates, as is $N^0 = \hat{N}_{obs}$. Let $N^*$ be the parametric bootstrap value for the maximum likelihood estimator $\hat{N}$ for given abundance $N$. By varying $N$ and studying the bootstrap distributions discussed below I found

$$piv(N, N^*) = \frac{N}{N^2} - \frac{b_0 + b_1 \frac{N}{N^2}}{a_0 + a_1 \frac{N}{N^2}}$$

(4)
to be an approximate normal pivot for both protocols. The bias coefficients $b_0$ and $b_1$, and the scale- and acceleration coefficients $a_0$ and $a_1$, depend on the protocol, and they influence the resulting confidence distribution and the related reduced likelihood. The coefficients are determined empirically by linear regression applied to the parametric bootstrap results to be presented next. In both cases the estimated coefficients satisfy $b_0 a_1 - b_1 a_0 > 0$, ensuring that the pivot in terms of the estimator, $piv(N, \hat{N})$, is increasing in $N$ for all values of $\hat{N}$.

The approximate pivot (4) leads to the following confidence distribution, confidence quantile, confidence density, and reduced log likelihood (Schweer and Hjort, 2002).

$$C(N) = \Phi \left( \frac{(1-b_1)N - b_0 \hat{N}_{obs}}{a_1 N + a_0 \hat{N}_{obs}} \right)$$

$$C^{-1}(p) = \frac{b_0 + a_0 \Phi^{-1}(p)}{1 - b_1 - a_1 \Phi^{-1}(p)} \hat{N}_{obs}$$

$$c(N) = \phi \left( \frac{(1-b_1)N - b_0 \hat{N}_{obs}}{a_1 N + a_0 \hat{N}_{obs}} \right) \frac{(a_0(1-b_1) - b_0 a_1) \hat{N}_{obs}}{(a_1 N + a_0 \hat{N}_{obs})^2}$$

$$I(N) = \left( \frac{(1-b_1)N - b_0 \hat{N}_{obs}}{a_1 N + a_0 \hat{N}_{obs}} \right)^2 2 + \log(N) - \log(a_1 N + a_0 \hat{N}_{obs})$$

The confidence density and the reduced log likelihood for total abundance are plotted in Figure 3 for each of the two protocols, as obtained from the bootstrap experiment. The fixed CV protocol leads to a nearly symmetric confidence distribution. The Petersen protocol, on the other hand, leads to

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate (SE)</th>
<th>Estimate (SE)</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>1985S 0.076 (0.030)</td>
<td>0.069 (0.032)</td>
<td></td>
</tr>
<tr>
<td>$\lambda$</td>
<td>1985A 0.033 (0.013)</td>
<td>0.119 (0.056)</td>
<td></td>
</tr>
<tr>
<td>$\lambda$</td>
<td>1986S 0.055 (0.021)</td>
<td>0.025 (0.012)</td>
<td></td>
</tr>
<tr>
<td>$\lambda$</td>
<td>1986A 0.019 (0.008)</td>
<td>0.035 (0.016)</td>
<td></td>
</tr>
<tr>
<td>$\theta^0_{intercept}$</td>
<td></td>
<td>0.98 (0.87)</td>
<td></td>
</tr>
<tr>
<td>$\theta^0_{mature}$</td>
<td></td>
<td>-1.70 (1.56)</td>
<td></td>
</tr>
<tr>
<td>$\theta^0_{autumn}$</td>
<td></td>
<td>1.62 (0.89)</td>
<td></td>
</tr>
<tr>
<td>Marked</td>
<td>282</td>
<td>675</td>
<td></td>
</tr>
<tr>
<td>Unmarked</td>
<td>2753</td>
<td>1531</td>
<td></td>
</tr>
<tr>
<td>Total abundance</td>
<td>3035</td>
<td>2206</td>
<td>5241</td>
</tr>
</tbody>
</table>
Figure 3. Approximate confidence density, left panel, and reduced log likelihood for the fixed CV protocol (solid line), and for the Petersen protocol (broken line). The vertical dotted line marks the maximum likelihood estimate, and the vertical solid line marks the median confidence for the fixed CV protocol.

a slightly skewed confidence distribution, and a more skewed reduced likelihood. Confidence quantiles are given in Table 5.

The confidence distribution for total abundance differs in structure from that for number of marked mature whales. The inclusion of unmarked whales does in fact introduce substantial uncertainty, particularly towards high abundance, while summing over strata tends to normalize the confidence distribution.

4.3 Bootstrap Experiment

A bootstrap experiment was carried out to identify approximate pivots and their distributions for the two hypothetical study protocols, each simulated under 9 different settings of the 14 parameters of model 6, and in about 1000 replicates. Total population size is varied systematically, and set at \( N = \text{scale} \cdot 5241 \), where the scale factor runs over scale = 0.6, 0.7, ..., 1.4 (Table 4). For each of the 18000 simulated data sets, maximum likelihood estimates were calculated according to (3).

The bootstrap distribution of the maximum likelihood estimate of total abundance is highly skewed, with an extreme right tail. The distributions of \( N/N^* \) are shown as q-q plots against the standard normal distribution in Figure 4 for the fixed CV protocol. The same picture emerge for the Petersen protocol. Except for the extreme left tail, \( N/N^* \) is approximately normally distributed with mean and standard deviation being linear in \( N \). The left tail results from some very high abundance estimates in data sets with very few recaptures. Table 4 summarizes the bootstrap experiment. Although the residual standard deviations in the linear regressions of \( b \) and \( \tau \) on \( N \) (or rather on scale = \( N/\hat{N}_{\text{obs}} \)) are higher than the standard deviations of the measurements, I settle on linear models for these parameters.

5. Discussion

My results are based on the assumption of independent and identically distributed encounters within survey and maturity class. All capture-recapture studies are vulnerable to such assumptions of homogeneity. Schwarz and Seber (1999), Buckland, Goudie, and Borchers (2000) and others stress the importance of homogeneity assumption. da Silva et al. (2000) found that moderate heterogeneity within strata is of minor concern in the present situation. One might, however, wonder whether intensities for recaptures within surveys being significantly different from the initial capture intensities imply unobserved heterogeneity between individuals. Unobserved heterogeneity implies such differences, but is not necessarily implied. If all individuals are present in the spring surveys, but some are absent in the autumn surveys, the model might hold if absence is independent between individuals and seasons. In this case, recapture intensities will be larger than initial capture intensities the more absence there is in the autumns.

The model can be modified to accommodate unobserved heterogeneity leading to positive dependency in the capture process for individuals over the surveys, and to seemingly higher recapture intensities than initial capture intensities.
Figure 4. Normal probability plots of $N/N_i$ (on the $y$ axis) for 9 sets of simulated data under the fixed CV protocol. The plots are, from upper left corner to bottom right, based on $N = \hat{N}_\text{obs}$, scale = 0.6, 0.7, ..., 1.4. The straight lines represent fitted normal distributions, with both mean and standard deviation linear in assumed $N$.

Focus on one closed stratum with marked individuals. Assume multiplicative heterogeneity for individual $j$, $\lambda_j = g_j \lambda_1$, where $g_j$ represents i.i.d. random effects. Let $y_{jt}$ be the number of captures in survey $t$, $y_j = (y_{j1}, \ldots, y_{j4})$ the capture history, and $y_j$ the total number of captures of individual $j$. When recapture intensities are set equal to initial capture intensities, $\sigma_t = 1$, the conditional distribution of $y_j$ given $y_j$ is multinomial with probability parameter $\lambda/\lambda_1$ regardless of

Table 4
Summary table for simulation results. In each case, 1000 data sets were simulated. $n$ is the number of these that lead to successful maximum likelihood estimation, and is the sample size used in the robust regressions to estimate $b$ and $\tau$.

<table>
<thead>
<tr>
<th>scale = $N/N_i$</th>
<th>Petersen</th>
<th>Fixed CV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$b$</td>
<td>$\tau$</td>
</tr>
<tr>
<td>0.6</td>
<td>.999</td>
<td>.223</td>
</tr>
<tr>
<td>0.7</td>
<td>.975</td>
<td>.254</td>
</tr>
<tr>
<td>0.8</td>
<td>.965</td>
<td>.273</td>
</tr>
<tr>
<td>0.9</td>
<td>.964</td>
<td>.288</td>
</tr>
<tr>
<td>1</td>
<td>.963</td>
<td>.291</td>
</tr>
<tr>
<td>1.1</td>
<td>.931</td>
<td>.303</td>
</tr>
<tr>
<td>1.2</td>
<td>.936</td>
<td>.339</td>
</tr>
<tr>
<td>1.3</td>
<td>.937</td>
<td>.345</td>
</tr>
<tr>
<td>1.4</td>
<td>.948</td>
<td>.358</td>
</tr>
<tr>
<td>Intercept ($a_0$ or $b_0$)</td>
<td>1.025</td>
<td>.307</td>
</tr>
<tr>
<td>Slope* ($a_1$ or $b_1$)</td>
<td>−.068</td>
<td>.160</td>
</tr>
<tr>
<td>rsd$^b$</td>
<td>.012</td>
<td>.008</td>
</tr>
<tr>
<td>p-value$^c$</td>
<td>.004</td>
<td>.000</td>
</tr>
</tbody>
</table>

$a$ Lower panels summarizes upper panel.

$b$ rsd is residual standard deviation, and slope is regression estimate in linear model of coefficient versus scale.

$c$ The p-value refers to testing for slope in the linear model for the coefficient.
the random effect (\( \lambda = \sum_{t=1}^{4} \lambda_t \)) etc. The number of unique captures over the surveys, \( X \), is binomial with number parameter \( N \) and success parameter \( 1 - \exp(-g\lambda) \). The likelihood for observed data \( (X, \{y_j\}) \) is therefore available, and it is explicit when \( g \) is gamma distributed. For a stratified population with both marked and unmarked individuals, a likelihood component is developed for each stratum for marked individuals and combined with a likelihood component for the numbers of unmarked captures, exactly as in the two-stage Poisson model.

The assumption of a closed population is also crucial, but should be reasonably satisfied since bowhead whales have very low mortality and fertility, and they grow slowly. The individual marks are assumed permanent. Miller et al. (1992) found large scars to remain over years. One cannot, however, exclude the possibility of some unmarked whales to become marked at some point during the study. This is also a source of error. Another source of error is the identification of photographs to individuals within encounters, between encounters and between surveys for marked whales. Matching errors might occur, and a small number of mismatches will impact the estimates considerably. I have made no attempt to investigate the impact of unobserved heterogeneity, population dynamics, dynamics in marking over the study period, nor of matching errors.

da Silva et al. (2000) use data for the same 175 marked whales as I use. They do, however, estimate multiplicity from the number of photos and not from the number of captures. Table 5 shows confidence intervals for two versions of the method of da Silva et al. (2000). Their confidence intervals are obtained from the bootstrap distribution by the percentile method and their results are consistent with mine. The Bering-Chukchi-Beaufort Seas population of bowhead whales was assessed in the International Whaling Commission in 1998 (Anon, 1999), and will be reassessed in 2004. My abundance estimate is consistent with the abundance estimates from the surveys in 1993 and 2001 combined with the estimated annual rate of population growth, 2.45%.

It is important to note that the two hypothetical study protocols I have considered are represented by one-dimensional curves in the 14-dimensional parameter space, indexed by total abundance and going through the maximum likelihood point. The bootstrap experiments allowed me to identify approximate pivots for each protocol. The standard bootstrap procedure is to construct a confidence distribution from one bootstrap experiment by the bias-corrected percentile method (Efron and Tibshirani, 1993), or by using a more sophisticated method taking account of acceleration. These standard methods are based on certain pivots, which seldom are explicitly presented and discussed. As we have seen, the pivot depends on the protocol of the study, and it thus seems advisable to discuss whether the pivot underlying the standard method indeed is appropriate in the particular case. The percentile method and its variants with names such as ABC, BCα, abc, etc. (Efron and Tibshirani, 1993; Davison and Hinkley, 1997; Schweder and Hjort, 2002) are based on asymptotic arguments. In our case, and in many other cases with meager data, extra caution is in place. Reference should then be made to the protocol of the study when choosing the pivot with consequential confidence distribution and reduced likelihood.

**Acknowledgements**

Dr Judy Zeh, University of Washington, has kindly allowed me access to the photo-identification data. I also acknowledge the resistance from the editors that has led to improvements in the presentation.

**Résumé**

À partir de relevés photographiques répétés, nous calculons l’estimateur du maximum de vraisemblance de l’effectif d’une population fermée et stratifiée d’animaux porteurs ou non d’un marquage individuel naturel, les baleines franches. Nous supposons pour l’observabilité un modèle log-linéaire en fonction de la strate, de l’année et de la saison. Pour le modèle sélectionné, nous obtenons par une expérience de bootstrap paramétrique une distribution de confiance approchée pour l’effectif total de la population, ainsi que la vraisemblance réduite des paramètres de nuisance. La distribution de confiance dépend du protocole d’étude supposé. Une distribution de confiance exacte (à l’effet de la discrétisation près) peut être obtenue en conditionnant dans le cas non stratifié sans individus non marqués.

**References**


Received April 2001. Revised May 2003.

Accepted May 2003.

**APPENDIX**

The Multiple Capture Model of Darroch

Consider a closed population of $N$ differently marked individuals, say the number of marked mature whales $N_{11}$. Captures are made in four surveys and recaptures within surveys are now disregarded. Capture histories are characterized by $\omega = (\omega_1, \ldots, \omega_t)$ consisting of zeros and ones, where $\omega_t = 1$ indicates capture in survey $t$. Let $X_{\omega}$ be the number of individuals with capture history $\omega$, $n_t$ the number of captures in survey $t$, and let $X$ be the number of unique individuals captured over the four surveys.

The Poisson model of Section 4.1 agrees with the model of Darroch (1958). The probability of being captured in survey $t$ is $p_t = 1 - \exp(\lambda_t)$. The probability of an individual having capture history $\omega$ is

$$p_{\omega} = \prod_{t=1}^{4} p_t^{n_t} (1 - p_t)^{1 - \omega_t}.$$ 

The vector $(N - X, \{X_{\omega}\})$ is then multinomially distributed with number parameter $N$. The multinomial likelihood simplifies to

$$L(N, p) = \binom{N}{X} \prod_{t=1}^{4} p_t^{n_t} (1 - p_t)^{N - n_t}.$$ 

The sufficient statistics are $X$ and $n$, where the latter component is ancillary for $N$ (from $n$ no additional information is available on $N$). Conditional inference is thus available from the conditional distribution of $X$ given $n$. Given these sample sizes, the four samples are independent simple random samples from the population. The sampling is done in sequence. For each survey $t$, the number of new captures, $y_t$, has a hypergeometric distribution, given the number of uniquely sampled individuals in previous surveys. Here, $y_1 = n_1$ and $X = \sum y_t$.

The conditional distribution is thus

$$P_N(X = x \mid n) = \sum_{y_{t-1}=0}^{x - y_t} \prod_{t=1}^{4} h\left(y_t; n_t, N - \sum_{j=1}^{t-1} y_j, N\right),$$

where $h(r; n, M, N) = \binom{M}{r} \binom{N - M}{n - r} / \binom{N}{n}$ is the hypergeometric probability.