BIOMETRICS 55, 294–301
March 1999

CONSULTANT'S FORUM

The Use of Mixed Logit Models to Reflect Heterogeneity in Capture-Recapture Studies

Brent A. Coull

Department of Biostatistics, Harvard School of Public Health,
655 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.
email: bcoull@hsph.harvard.edu

and

Alan Agresti

Department of Statistics, University of Florida,
Gainesville, Florida 32611-8545, U.S.A.

SUMMARY. We examine issues in estimating population size N with capture-recapture models when there is variable catchability among subjects. We focus on a logistic-normal mixed model, for which the logit of the probability of capture is an additive function of a random subject and a fixed sampling occasion parameter. When the probability of capture is small or the degree of heterogeneity is large, the log-likelihood surface is relatively flat and it is difficult to obtain much information about N. We also discuss a latent class model and a log-linear model that account for heterogeneity and show that the log-linear model has greater scope. Models assuming homogeneity provide much narrower intervals for N but are usually highly overly optimistic, the actual coverage probability being much lower than the nominal level.

KEY WORDS: Generalized linear mixed model; Latent class model; Log-linear model; Profile likelihood; Quasi symmetry; Rasch model.

1. Introduction

This article discusses the use of mixture models in capture-recapture experiments devised to estimate the size N of a closed population. These models address the estimation of N when heterogeneity exists in the population with respect to catchability.

Norris and Pollock (1996) listed papers that consider models allowing for heterogeneous catchability. Relatively few models have been developed in the class, denoted by M_{th}, that allows heterogeneous capture probabilities to vary also across sampling occasions. Sanathanan (1972a,b) considered this setting in the context of visual scanning experiments and used a mixed logit model. Darroch et al. (1993) and Agresti (1994) discussed log-linear models motivated by a fixed-effects formulation of Sanathanan's model. Chao, Lee, and Jeng (1992), Chao et al. (1996), and Chao and Tsay (1998) used a nonparametric sample coverage approach.

This article discusses strategies for estimating N under M_{th} assumptions when the captured subjects are marked or recorded in such a way that the number of subjects with a particular pattern of being observed or not being observed at each sampling occasion is available. For t sampling occasions, the data can then be displayed in a 2^t contingency table, with a missing observation for the cell corresponding to noncapture at each occasion. For a logit model with a random effect reflecting subject heterogeneity, we study how the estimator of N depends on the variance component and on the extent of sampling. With strong heterogeneity and a small proportion of the population captured in each sample, we see that the $2^t - 1$ observable counts provide little information about N. We also discuss alternative models, of log-linear and latent class form, that account for heterogeneity. Simulations indicate that the log-linear model of homogeneous two-factor association sometimes performs better than the mixed model with respect to both accuracy of the point estimate and coverage and length of the resulting confidence intervals. These simulations also demonstrate the lack of information about N often provided by the latent class model and the overly optimistic nature of models that assume subject homogeneity.

Section 2 presents the mixed logit, log-linear, and latent class models. Section 3 reviews maximum likelihood (ML) estimation of N and an alternative to the standard asymptotic confidence interval for N for these models. Section 4 provides an example, while Section 5 discusses the behavior of the log-likelihood and of point and interval estimators of N in the presence of subject heterogeneity. Section 6 examines the trade-off between confidence intervals being informative and yet maintaining the nominal error rate and makes recommendations based on the results of a simulation study.

2. Mixed Logit Models and Latent Class Models

We first review the models in their traditional application, when all N individuals are observed, before discussing their extensions to the capture-recapture setting in Section 3.

2.1 *A Logit Model with Subject Heterogeneity*

For subject s, $s = 1, \ldots, N$, let $\underline{y}'_s = (y_{s1}, \ldots, y_{st})$ be a vector of t binary measurements (0 or 1), where $y_{sj} = 1$ denotes capture in sample j. Let $p_{sj} = P(y_{sj} = 1)$. We permit subject heterogeneity using the model

$$\operatorname{logit}(p_{sj}) = \alpha_s + \beta_j. \qquad (1)$$

The greater the variability in $\{\alpha_s\}$, the more heterogeneous are the capture probabilities at a given sample. The larger the value of β_j, the greater the probability of capture at occasion j. Original applications of the model (Rasch, 1961) referred to t test items, making the model popular in educational testing, where it is known as the Rasch model. In fitting the model, one assumes independence of responses across occasions for a given subject, termed local independence, and independence between subjects.

Standard ML asymptotics do not apply to this model since, as the number of subjects (N) grows, the number of model parameters also grows. Thus, the ordinary ML estimate of $\underline{\beta} = (\beta_1, \ldots, \beta_t)$ is not consistent (Andersen, 1980). Two approaches are used to overcome the inconsistency. The first, a fixed-effects approach, treats $\{\alpha_s\}$ as nuisance parameters and eliminates them by conditioning on their sufficient statistics, yielding conditional ML (CML) estimates $\hat{\underline{\beta}}^C$. Let $\mathcal{I} = \{(1, \ldots, 1), \ldots, (0, \ldots, 0)\}$ be the set of 2^t possible sequences of responses $(y_{s1}, \ldots, y_{st})$, in lexicographic order. Let $\underline{i} = (i_1, \ldots, i_t)$ be an element of $\mathcal{I}$ and let $n_{\underline{i}} = n_{i_1 \cdots i_t}$ be the number of subjects having that sequence. Tjur (1982) showed that the CML estimates of $\underline{\beta}$ are, equivalently, ML estimates of main effect parameters in a log-linear model of quasi symmetry fitted to the 2^t table of counts $\{n_{\underline{i}}\}$. Specifically, letting $\{\mu_{\underline{i}} = \mathrm{E}(n_{\underline{i}})\}$, the log-linear model is

$$\log(\mu_{\underline{i}}) = \beta_0 + \beta_1 I(i_1 = 1) + \cdots + \beta_t I(i_t = 1) + \lambda(i_1, \ldots, i_t), \qquad (2)$$

where the parameter $\lambda(i_1, \ldots, i_t)$ is invariant to permutations of its arguments and the $I(\cdot)$ function is an indicator.

A second approach treats $\{\alpha_s\}$ as random effects, typically having a normal distribution with mean zero and unknown variance σ^2, for which

$$\operatorname{logit}(p_{sj}) = \sigma Z_s + \beta_j, \qquad (3)$$

with $Z_s \sim \mathrm{N}(0, 1)$. The probability that a subject with ability Z has capture history $\underline{i}$, $\underline{i} \in \mathcal{I}$, is $\pi_{\underline{i}|Z} = \Pi_{j=1}^{t} [\exp\{i_j(\sigma Z + \beta_j)\}/\{1 + \exp(\sigma Z + \beta_j)\}]$. Thus, the probability that a randomly selected subject has that pattern is the marginal probability

$$\pi_{\underline{i}} = \int \pi_{\underline{i}|z}\phi(z)dz = \int \left[\prod_{j=1}^{t} \frac{e^{\{i_j(\sigma z+\beta_j)\}}}{1+e^{(\sigma z+\beta_j)}}\right] \phi(z)dz, \qquad (4)$$

where $\phi(z)$ is the standard normal density. This also satisfies the quasi-symmetry model (2), regardless of the distributional assumption about the random effect (Tjur, 1982). This model implies a positive dependence structure among the t occasions in the form of uniformly nonnegative log odds ratios, both conditional and marginal, in the 2^t table. The marginal multinomial log-likelihood for $(\sigma, \underline{\beta})$, given the cell counts $\underline{\mathbf{n}} = (n_{1 \cdots 1}, \ldots, n_{0 \cdots 0})$, is $l(\sigma, \underline{\beta}; \underline{\mathbf{n}}) \propto \Sigma_{\underline{i} \in \mathcal{I}}\, n_{\underline{i}} \log(\pi_{\underline{i}})$.

Using Gaussian quadrature, one can approximate the marginal probabilities (4) with $\tilde{\pi}_{\underline{i}} = \Sigma_{k=1}^{q} [\Pi_{j=1}^{t} \exp\{i_j(\sigma z_k + \beta_j)\}/ \{1 + \exp(\sigma z_k + \beta_j)\}]\nu_k$ for tabulated $\{z_k\}$ and $\{\nu_k\}$ (Aitkin, 1996). The choice of the number of quadrature points q determines the degree of accuracy, and larger q is needed when σ is larger. Then $\Sigma_{\underline{i} \in \mathcal{I}}\, n_{\underline{i}} \log(\tilde{\pi}_{\underline{i}})$ is the objective function maximized with respect to $(\sigma, \underline{\beta})$.

2.2 *Quasi-Symmetric Log-Linear and Latent Class Models*

This article also discusses two log-linear models for $\mu_{\underline{i}}$ and a latent class model that have connections with the mixed model (3). The log-linear models are simple ones assuming either mutual independence of responses or a homogeneous association pattern. The mutual independence model, which is (2) with constant value for $\lambda(\cdot)$, results from the logistic-normal model with $\sigma = 0$; i.e., it assumes homogeneity of subjects. Darroch et al. (1993) and Agresti (1994) noted that, although the mixed model satisfies the quasi-symmetry model (2) marginally, fitting that marginal model provides no information about N since one of its likelihood equations shows that any value $n_{0 \cdots 0}$ is consistent with the model. They considered special cases of the quasi-symmetry model for which this is not the case. In particular, the log-linear model of homogeneous two-factor association (HO2),

$$\log(\mu_{\underline{i}}) = \beta_0 + \beta_1 I(i_1 = 1) + \cdots + \beta_t I(i_t = 1) + \binom{\sum_{j=1}^{t} i_j}{2} \lambda, \qquad (5)$$

is the special case in which only the second-order interactions differ from zero. This model is the special case of the log-linear model of no three-factor interaction in which all pairwise associations are identical. The mutual independence model is the further special case of (5) in which $\lambda = 0$.

The latent class (LC) model discussed is the special case of model (1) with only two possible values for α_s. It assumes that the population is a mixture of two types, with homogeneity of subjects within each type but with the type of any given subject being unknown. This model is a special case of latent class models introduced by Lindsay, Clogg, and Grego (1991) and Agresti and Lang (1993). The LC model also relates to the normal random effects version (3) of the model, being a generalization of the $q = 2$ Gaussian quadrature approximation of it. Two-point Gaussian quadrature results in a latent class model with two classes that place equal probability ($\nu_k = 0.5, k = 1, 2$) of a subject being in class k. If we relax $\nu_1 = \nu_2 = 0.5$ and instead estimate these weights, the 2^t expected frequencies satisfy

$$\log(\mu_{\underline{i}}) = \sum_{j=1}^{t} \beta_j i_j + \log \left\{ \sum_{k=1}^{2} \exp \left(\sum_{j=1}^{t} \sigma I[k=2] i_j + \lambda_k \right) \right\}, \tag{6}$$

where $\lambda_k = [\log(\nu_k) - \log \Pi_{j=1}^{t} \{1 + \exp(\sigma I[k=2] + \beta_j)\}] - [\log(\nu_1) - \log \Pi_{j=1}^{t}\{1 + \exp(\beta_j)\}]$. This is equivalent to the LC model with two latent classes and common association parameter σ.

This model represents a compromise between the mutual independence model and the logistic-normal model, allowing some heterogeneity yet assuming homogeneity within each latent class. One can fit it using the EM algorithm. The distribution of the complete data has regular exponential form so that only the complete data sufficient statistics must be estimated at each E-step. (See Goodman (1974) for an EM approach to latent class models in general and Agresti and Lang (1993) for this specific case.) In general, let L be the number of latent classes. Lindsay et al. (1991) proved that if $L > t/2$, and if there exists a distribution for α_s in the Rasch model (1) such that the random effects solution can exactly fit the sufficient statistics for α_s, then necessarily the latent class fit will be identical to the conditional maximum likelihood fit. The authors also gave simple numerical and graphical tests of this condition, and in practice this equivalence seems quite common when $L > t/2$. In such cases, the fit is then also identical to that of the quasi-symmetry model and the model is not informative for the capture-recapture problem.

3. ML Estimation of N in Capture-Recapture Studies

In the capture-recapture setting, the cell count $n_{0\cdots0}$ is unknown. In subsequent discussions, we refer to the table with all 2^t cell counts known as the complete table and the one with $n_{0\cdots0}$ unknown as the incomplete table.

3.1 *Point Estimation*

Let $\mathcal{I}^O = \{(1,\ldots,1),\ldots,(0,\ldots,0,1)\}$, $\underline{\mathbf{n}}^O = (n_{1\cdots1},\ldots,n_{0\cdots1})$, and $n = \Sigma_{\underline{i}\in\mathcal{I}^O} n_{\underline{i}}$ denote the observable capture histories, the observable cell counts, and the total number of observed subjects, respectively. For a model indexed by $\underline{\boldsymbol{\theta}}$, we use Sanathanan's (1972a) conditional approach to N-estimation. This approach obtains an N-estimate by maximizing the binomial likelihood of observing n successes in N trials when the probability of success is $1 - \pi_{0\cdots0}(\underline{\hat{\boldsymbol{\theta}}}_C)$, where $\underline{\hat{\boldsymbol{\theta}}}_C$ maximizes the conditional likelihood

$$L_1\left(\underline{\boldsymbol{\theta}}; \underline{\mathbf{n}}^O \mid n\right) = \frac{n!}{n_{1\cdots1}!\cdots n_{0\cdots1}!} \pi'_{1\cdots1}(\underline{\boldsymbol{\theta}})^{n_{1\cdots1}} \cdots \pi'_{0\cdots1}(\underline{\boldsymbol{\theta}})^{n_{0\cdots1}},$$

with

$$\pi'_{\underline{i}} = \pi_{\underline{i}}(\underline{\boldsymbol{\theta}})/\{\Sigma_{\underline{i}\in\mathcal{I}^O}\, \pi_{\underline{i}}(\underline{\boldsymbol{\theta}})\}, \qquad \underline{i} \in \mathcal{I}^O.$$

The resulting estimate $\hat{N}_C = n/\{1 - \pi_{0\cdots0}(\underline{\hat{\boldsymbol{\theta}}}_C)\}$ is the basis of traditional log-linear modelling of capture-recapture experiments (Fienberg, 1972; Cormack, 1989).

One can also use the unconditional ML estimate $\hat{N}$ of the population size. Sanathanan (1972a) commented that necessarily $\hat{N} \leq \hat{N}_C$. It has been our experience that the behavior of the unconditional estimate in the presence of subject heterogeneity is similar to that of the conditional estimate. We focus solely on $\hat{N}_C$ here since it lends itself to the construction of profile likelihood confidence intervals for N.

3.2 *Confidence Interval*

Recent research on confidence intervals for population size has focused on methods for small to moderate samples. The sampling distributions of the N-estimators are then highly skewed, and the asymptotic Wald-type confidence interval for N can have a lower bound falling below the observed number of subjects. In this paper, we use profile likelihood confidence intervals (Cormack, 1992) since Sanathanan's $\hat{N}_C$ has a likelihood-ratio interpretation. Specifically, profile likelihood intervals use the fact that $\hat{n}_{(0\cdots0)C} = \hat{N}_C - n$ is the value of the missing cell count that yields the smallest likelihood-ratio statistic for testing goodness-of-fit of the model to the complete table. Denote the value of this statistic by $G^2(\hat{n}_{(0\cdots0)C})$. The $100(1-\alpha)\%$ profile likelihood interval includes values of $n_{0\cdots0}$, and hence N, that satisfy $G^2(n_{0\cdots0}) - G^2(\hat{n}_{(0\cdots0)C}) \leq \chi^2_{1,\alpha}$, the upper-tail percentage point from a chi-squared distribution with 1 d.f.

Other possible alternatives to Wald-type interval estimation are based on the bootstrap, either a percentile version or Efron's bias corrected and accelerated interval (Buckland and Garthwaite, 1991). Here we focus only on the profile likelihood method since our comparisons, which we plan to report in a separate paper, indicated that this interval holds a substantial advantage over the bootstrap intervals with respect to both coverage and computation for the models we considered.

4. Snowshoe Hare Example

Cormack (1989) reported a capture-recapture study having $t = 6$ consecutive trapping days for a population of snowshoe hares. Table 1, which displays the data, shows that 68 hares were observed. Table 2 summarizes the N-estimates and confidence intervals based on various models. The logistic-normal model using $q = 20$-point quadrature yields $\hat{\sigma}_C = 0.97$ and $\hat{N}_C = 92.0$. Table 1 shows this fit. A profile of $\hat{N}_C$ across $q \geq 2$ reveals that the estimates stabilize for $q > 5$, but it takes a much larger value of q before the plot for the log-likelihood stabilizes; the interval estimates stabilize for $q > 17$.

The log-linear model of homogeneous two-factor association (HO2) gives a similar point estimate of N and fit as the logistic-normal model. All observed cell fitted values were no further than 0.04 from the values shown in Table 1 for the logistic-normal model. However, the HO2 95% profile likelihood interval is narrower than the interval obtained with the logistic-normal model. The log-linear model of mutual independence, which assumes no heterogeneity, yields the narrowest interval. The latent class model also has a much narrower interval than the one for the logistic-normal model.

This example illustrates the considerable variability among the point and interval estimates for N that one usually obtains with different models. The impact of subject heterogeneity on these results is explored in the next section.

Table 1
Results of capture-recapture of snowshoe hares

			Captures 3, 2, and 1							
Capture 6	Capture 5	Capture 4	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
0	0	0	— (24.0)[a]	3 (2.3)	6 (5.4)	0 (0.9)	5 (3.2)	1 (0.5)	0 (1.2)	0 (0.3)
0	0	1	3 (4.8)	2 (0.8)	3 (1.8)	0 (0.5)	0 (1.1)	1 (0.3)	0 (0.6)	0 (0.3)
0	1	0	4 (3.9)	2 (0.6)	3 (1.5)	1 (0.4)	0 (0.9)	1 (0.2)	0 (0.5)	0 (0.2)
0	1	1	1 (1.3)	0 (0.3)	0 (0.8)	0 (0.3)	0 (0.5)	0 (0.2)	0 (0.4)	0 (0.3)
1	0	0	4 (6.8)	1 (1.1)	1 (2.6)	1 (0.6)	2 (1.5)	0 (0.4)	2 (0.9)	0 (0.4)
1	0	1	4 (2.3)	0 (0.6)	3 (1.3)	0 (0.5)	1 (0.8)	0 (0.3)	2 (0.7)	0 (0.4)
1	1	0	2 (1.9)	0 (0.5)	1 (1.1)	0 (0.4)	1 (0.7)	0 (0.3)	1 (0.6)	0 (0.4)
1	1	1	1 (1.0)	1 (0.4)	1 (0.9)	0 (0.5)	0 (0.5)	0 (0.3)	1 (0.7)	2 (0.7)

[a] Fit of logistic-normal model ($q = 20$).
Data from Cormack (1989).

5. Behavior of the Log Likelihood and N Estimator When There Is Subject Heterogeneity

In the capture-recapture problem, large heterogeneity results in strong positive associations among the t capture results and also has a strong impact on the estimation of N. The greater the heterogeneity, as reflected by $\hat{\sigma}_C$, the larger the estimate of $n_{0\cdots0}$ tends to be. For the logistic-normal model with the snowshoe hare data, the first plot in Figure 1 shows $\hat{N}_C$ as a function of an assumed known value for σ. Since $\hat{N}_C$ is a rapidly increasing function of σ, small changes in $\hat{\sigma}_C$ can have a large impact on the ML estimate of N. Plot 2 in Figure 1 displays a profile of $-2 \log L$ in terms of σ, revealing that $\hat{\sigma}_C = 0.97$. The case of large heterogeneity causes difficulties in estimation for the logistic-normal model since a large $\hat{\sigma}_C$ results in a relatively flat likelihood surface, which implies unstable and imprecise N-estimates. Figure 1 shows that this problem is not serious with the snowshoe hare data.

Table 2
N-estimates and profile likelihood confidence intervals for Table 1 produced by the logistic-normal ($q = 20$), latent class, homogeneous two-factor association, and mutual-independence models

Model	$\hat{N}_C$	95% confidence interval
Logistic-normal	92.0	(74.8, 153.6)
Homogeneous two-factor association	90.5	(74.8, 125.1)
Latent class	77.1	(70.8, 87.4)
Mutual independence	75.1	(69.9, 83.3)

As heterogeneity increases, the probability of capturing a subject a relatively small or relatively large number of times increases. If the logistic-normal model provides a reasonable approximation for a particular application, having only one capture for a large proportion of the sampled subjects suggests that considerable heterogeneity exists within the population and/or that the probabilities of capture at each occasion are small. Of course, having only one capture for most subjects could also occur if the model assumptions are badly violated, e.g., when animals exhibit trap avoidance, so that the local independence assumption of the logistic-normal model is inappropriate. Unfortunately, traditional goodness-of-fit tests cannot necessarily differentiate among these cases or between a correct and incorrect model, as discussed in Section 6.

In contrast to the relatively stable point estimates for the snowshoe hare data, consider Table 3 from Chao et al. (1996),

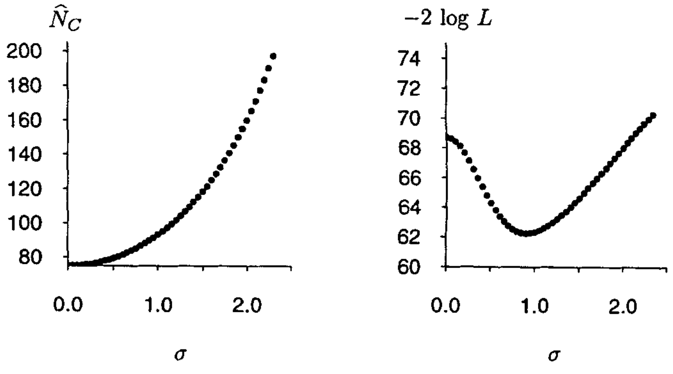

Figure 1. $\hat{N}_C$ and $-2 \log L$ as a function of σ for the logistic-normal model with the snowshoe hare data (Table 1).

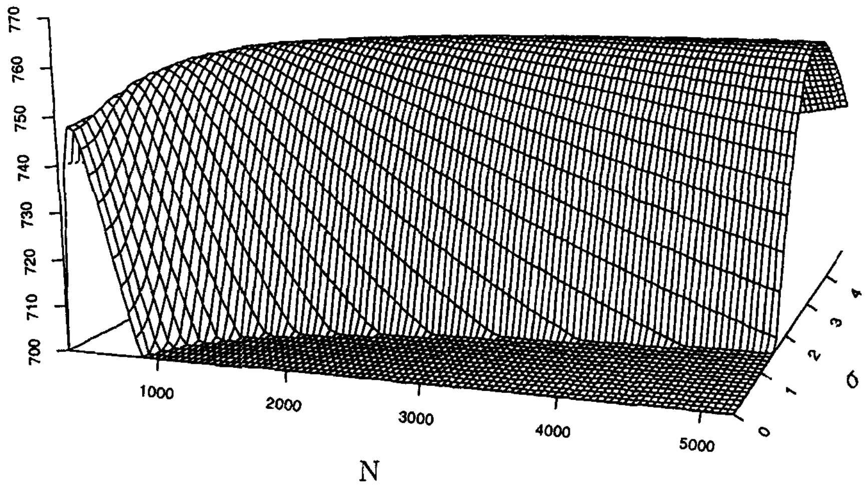

Figure 2. View of the profile log-likelihood surface with respect to N and σ, maximized over $\underline{\beta}$, for the hepatitis data.

which reports the results from an epidemiological study designed to estimate the number of people infected during a 1995 hepatitis A outbreak in northern Taiwan. The 271 observed cases were reported from three sources: records based on a serum test taken by the Institute of Preventive Medicine of Taiwan (P), records reported by the National Quarantine Service (Q), and records based on questionnaires conducted by epidemiologists (E). Here, $\hat{\sigma}_C = 2.9$ is large, and the large negative values for $\hat{\underline{\beta}}_C = (-6.7, -6.8, -6.8)$ reflect the many subjects with one capture (187) compared with the numbers of subjects with two or three captures (56 and 28). In such situations, the data provide relatively little information about N. The log-likelihood is relatively flat with respect to σ, so a wide range of σ values are consistent with the data. The plausible σ values, however, correspond to a wide range of N-estimates since $\hat{N}_C$ increases sharply with respect to σ.

Figure 2 shows the profile log-likelihood surface with respect to N and σ maximized over $\underline{\beta}$. Nothing practical can be said about N except that it is not very small. The flat log-likelihood can cause wild fluctuations in the point estimate due to small changes in numerical precision or in the data themselves. This flat surface produces a 95% profile likelihood interval for N of $(758, \infty)$. Why does the logistic-normal model sometimes provide little information about the population size? The reason is similar to the reason why every $n_{0\cdots0} \geq 0$ is plausible for the quasi-symmetry model. This model results from a completely unspecified mixing distribution, so that each candidate $n_{0\cdots0}$ is plausible. The normal class of mixing distributions is itself rich enough that many values of σ (and of $n_{0\cdots0}$) may be consistent with the data. A wide range of plausible σ values implies that the candidate N values form a wide interval, amounting to little practical information about N.

Instead of allowing each subject to have a different propensity for capture, the latent class (LC) approach requires Z_s to take one of only two values. For the hepatitis example, Figure 3 portrays the deviance profile for the LC model. Because all $N \geq 479$ yield a constant deviance, this model provides no point estimate and a profile likelihood interval of $(407, \infty)$. The flat log-likelihood relates to results of Lindsay et al. (1991) about the model's close relationship to the log-linear model of quasi symmetry, which provides no information about N. Chao et al. (1996) stated that the true population size for the hepatitis data is approximately 545. The complete table satisfies the Lindsay et al. (1991) condition for equivalence of the fit of the LC model with two classes and the conditional ML fit and hence the quasi-symmetry fit.

Flat likelihoods have occurred in other capture-recapture approaches. Burnham and Overton (1978) made a passing reference to the perfomance, in this respect, of the beta-binomial model, and they presented a jackknife estimator. Similarities exist between this problem and the related problem of N-estimation when observing k independent and identically

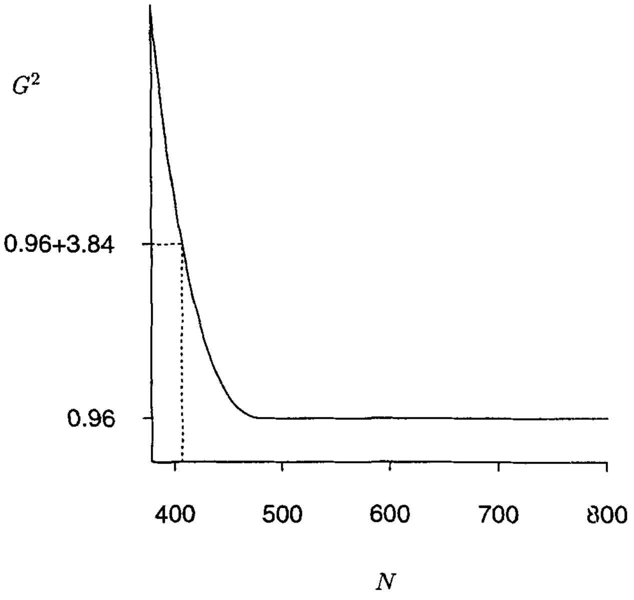

Figure 3. Deviance (G^2) profile for $379 \leq N \leq 800$ for the hepatitis data (Table 3).

Table 3

Capture-history counts and conditional (on n) fitted values for hepatitis study. First row displays 95% profile likelihood confidence intervals for n_{000}.

PQE	Observed count	Quasi-symmetry fit	Logistic-normal fit ($q = 50$)	Latent class fit
0 0 0	—	$(0, \infty)$	$(487, \infty)$	$(136, \infty)$
0 0 1	63	61.0	61.0	61.0
0 1 0	55	58.0	58.0	58.0
0 1 1	18	17.0	17.0	17.0
1 0 0	69	68.0	68.0	68.0
1 0 1	17	20.0	20.0	20.0
1 1 0	21	19.0	19.0	19.0
1 1 1	28	28.0	28.0	28.0

Note: Data from Chao et al. (1996), with P = Institute of Preventive Medicine of Taiwan, Q = National Quarantine Service, and E = epidemiologists.

distributed binomial counts with unknown N and probability parameter (cf., Aitkin and Stasinopoulos, 1989, and references therein). These authors demonstrated that, when the log likelihood is flat, the ML estimator is unstable, with small changes in the data yielding large changes in $\hat{N}$. For the logistic-normal model ($q = 50$) applied to the hepatitis data, $\hat{N}_C$ changes from 3856 to 4551 to 5443 when $n_{1 \cdots 1}$ changes from 27 to 28 to 29.

6. Comparisons and Recommendations

In capture-recapture experiments, we have seen that the point estimate and the associated confidence interval for N depends strongly on the choice of model. This strong dependence reflects the fact that the capture-recapture problem is inherently one of prediction, i.e., in estimating $n_{0 \cdots 0}$, one extrapolates from the range of the observed data, numbers of subjects having $1, 2, \ldots, t$ captures, to the number of subjects with zero captures. The standard goodness-of-fit criteria are of limited help for this extrapolation problem since two models can provide good fits to the observed data yet yield dramatically different estimates for the unobserved count. The hepatitis data illustrates this point. The logistic-normal and latent class models provide identical fits to the incomplete table yet quite different ones to the complete table. The logistic-normal model does not fit the complete table well since $G^2_{\text{complete}} = 9.3$ with 3 d.f. The latent class model, on the other hand, yields $G^2_{\text{complete}} = 1.0$ with 2 d.f. Thus, one cannot definitively test for dependence between samples in the complete table using only the observed counts in the incomplete table.

Given the limited use of goodness-of-fit criteria, the question occurs as to how to select a model. In some applications, subject matter may suggest a particular model. In practice, the probability of capture is often small and most subjects appear in one or none of the samples. There then typically exists an increasing ordering of the magnitude of $\hat{N}_C$ from simple to more complex models. The width of the interval estimates also follow this ordering, reflecting the smaller standard errors obtained with more parsimonious models. The simpler models also have the advantage of greater stability, with the N-estimates not fluctuating so wildly with small changes in the data. We get nothing for free, however. These simpler models either do not account for population heterogeneity or underestimate it, so the point estimates can severely underestimate N and the associated confidence intervals can have actual coverages well below the nominal level. We feel that wide intervals in such cases merely reflect the small amount of information about the population size that results when most subjects are captured only once.

A trade-off clearly occurs in selecting a model. One would like a narrow confidence interval for N but not at the expense of drastic sacrifice in the actual confidence level. Much of the capture-recapture literature has recommended models producing narrow intervals, but we believe that such intervals are usually overly optimistic. An example of this contrast is the difference between the HO2 estimate and a sample coverage estimate $\hat{N}_{SC}$ given by Chao et al. (1996) and Chao and Tsay (1998) for the hepatitis data set. There, $\hat{N}_{HO2}$ has a standard error of about 900, causing the authors to reject this estimator in favor of $\hat{N}_{SC}$, which has a much narrower confidence interval. The simulations of Chao et al., however, indicate that, when the capture history counts are simulated from the logistic-normal model with $N = 200$, the actual coverage for 95% bootstrap confidence intervals generated using $\hat{N}_{SC}$ is 58.5% while the corresponding figure for $\hat{N}_{HO2}$ is 91.5%.

This trade-off is also evident in several simulation studies that we conducted. For lack of space, the details of the simulations and tables listing their results are not shown here, but this material is available from the authors in technical report form. The mutual independence model provides the narrowest intervals but the poorest coverage, while the logistic-normal models often provide close-to-nominal coverage but little practical information on N. For estimates based on the mutual independence model, the profile likelihood coverage is close to the nominal level when that model is the true model and N is large, but this method tends to badly underestimate N in the presence of heterogeneity. Based on the simulation results, we recommend that the logistic-normal model be used as a diagnostic tool when selecting a model. The MLE of σ and the profile likelihood from this model, along with the numbers of subjects captured $0, \ldots, t$ times, provide information about the amount of heterogeneity present and the probabilities of capture. If $0 < \hat{\sigma}_C < 1$ and the occasion parameter estimates are not large negative numbers, we recommend using the profile likelihood interval based on $\hat{N}_C$ from the logistic-normal model. Simulations show that when the subjects are spread somewhat evenly over numbers $0, \ldots, t$ of captures, this model does slightly better than the HO2 model with respect to coverage without producing wider intervals.

If the logistic-normal model yields $\hat{\sigma}_C = 0.0$, one should consider the HO2 log-linear model since there is the possibility of negative dependencies among the t occasions. We simulated a situation in which both population heterogeneity and serial within-subject dependencies exist among the t responses. When trap avoidance causes the log odds ratios for consecutive samples to be negative in the 2^t table, the logistic-normal fit is obtained on the boundary ($\hat{\sigma}_C = 0.0$) and the model will not usually estimate N accurately. This failure of the logistic-normal model is to be expected since generalized linear mixed models containing a random intercept cannot describe a negative dependence structure among the t responses. The HO2 model maintains close-to-nominal coverage for both positive and negative association structures since the homogeneous two-factor association term λ is not constrained to be positive.

When most subjects are captured only once and the continuous mixture model yields a relatively flat likelihood and $\hat{\sigma}_C > 1$, we also recommend using the HO2 model since it yields narrower intervals and much more accurate N-estimates than the mixed model, even when that model truly holds. For all underlying models and parameter settings considered in our simulation studies, this model yielded coverage probabilities that were never far from the nominal level. Moreover, the HO2 model is simple to use, being a log-linear model that accounts for heterogeneity by adding a single association parameter to the log-linear model of mutual independence.

Our simulation studies also suggested that, for large σ, the finite intervals from the latent class model are overly optimistic, the actual coverage probabilities being considerably less than the nominal level. Thus, we are skeptical of the narrow latent class profile likelihood confidence interval for the snowshoe hare data, for which the logistic-normal model has $\hat{\sigma}_C = 0.97$. In general, we do not recommend using the latent class model to estimate N, even if it is the true model; its profile likelihood intervals often have infinite length, and if they have finite length but there is much heterogeneity, then their coverage probability is suspect.

In summary, severe population heterogeneity and/or small probabilities of capture in a capture-recapture experiment make reaching useful conclusions difficult. Mixed capture-recapture models allowing heterogeneous catchability reflect the large amount of uncertainty in estimating N through wide confidence intervals. Flat log-likelihoods can result in unstable estimates that are sensitive to small changes in the data. When this is the case, one must use caution in estimating N with simple capture-recapture models since these models tend to produce overly optimistic confidence statements about N.

Acknowledgement

This research was partially supported by a grant from the National Institutes of Health.

Résumé

A moins que la véritable association soit très forte, de simples intervalles de confiance pour l'odds ratio, établis à partir de la méthode delta et pour de grands échantillons, ont de bonnes performances, même pour des petits échantillons. Ces intervalles comprennent l'intervalle logit de Woolf et l'intervalle associé de Gart, où l'on ajoute .5 avant le calcul de l'estimateur du logarithme de l'odds ratio et de son erreur standard. L'intervalle de Gart équilibre les valeurs observées vers le modèle uniforme, mais on obtient de meilleures probabilités de recouvrement en équilibrant vers le modèle d'indépendance, et en étendant l'intervalle dans la direction appropriée lorsqu'une cellule correspond à une valeur nulle.

References

Agresti, A. (1994). Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics* **50,** 494–500.

Agresti, A. and Lang, J. B. (1993). Quasi-symmetric latent class models, with application to rater agreement. *Biometrics* **49,** 131–139.

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing* **6,** 251–262.

Aitkin, M. and Stasinopoulos, M. (1989). Likelihood analysis of a binomial sample size problem. In *Contributions to Probability and Statistics. Essays in Honor of Ingram Olkin,* L. J. Gleser, M. D. Perlman, S. J. Press, and A. R. Simpson (eds), 399–411. New York: Springer-Verlag.

Andersen, E. B. (1980). *Discrete Statistical Models with Social Science Applications.* Amsterdam: North-Holland.

Buckland, S. T. and Garthwaite, P. H. (1991). Quantifying precision of mark-recapture estimates using the bootstrap and related methods. *Biometrics* **47,** 255–268.

Burnham, K. P. and Overton, W. S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* **65,** 625–633.

Chao, A. and Tsay, P. K. (1998). A sample coverage approach to multiple-system estimation with application to census undercount. *Journal of the American Statistical Association* **93,** 283–293.

Chao, A., Lee, S.-M., and Jeng, S.-L. (1992). Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics* **48,** 201–216.

Chao, A., Tsay, P. K., Shau, W.-Y., and Chao, D.-Y. (1996). Population size estimation for capture-recapture models with applications to epidemiological data. *Proceedings of the Biometrics Section, American Statistical Association* 108–117.

Cormack, R. M. (1989). Log-linear models for capture-recapture. *Biometrics* **45,** 395–413.

Cormack, R. M. (1992). Interval estimation for mark-recapture studies of closed populations. *Biometrics* **48,** 567–576.

Darroch, J. N., Fienberg, S. E., Glonek, G. F. V., and Junker, B. W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association* **88,** 1137–1148.

Fienberg, S. E. (1972). The multiple-recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika* **59,** 591–603.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61,** 215–231.

Lindsay, B., Clogg, C., and Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association* **86,** 96–107.

Norris, J. L. and Pollock, K. H. (1996). Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics* **52,** 639–649.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability,* Volume 4, J. Neyman (ed), 321–333. Berkeley: University of California Press.

Sanathanan, L. (1972a). Estimating the size of a multinomial population. *Annals of Mathematical Statistics* **43,** 142–152.

Sanathanan, L. (1972b). Models and estimation methods in visual scanning experiments. *Technometrics* **14,** 813–829.

Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scandinavian Journal of Statistics* **9,** 23–30.

Received October 1997. *Revised March* 1998.
Accepted March 1998.