Biometrics 61, 868–876
September 2005

The Performance of Mixture Models in Heterogeneous Closed Population Capture–Recapture

Shirley Pledger

School of Mathematics, Statistics and Computer Science, Victoria University of Wellington,
P.O. Box 600, Wellington, New Zealand
email: shirley@mcs.vuw.ac.nz

Summary. Dorazio and Royle (2003, *Biometrics* **59,** 351–364) investigated the behavior of three mixture models for closed population capture–recapture analysis in the presence of individual heterogeneity of capture probability. Their simulations were from the beta-binomial distribution, with analyses from the beta-binomial, the logit-normal, and the finite mixture (latent class) models. In this response, simulations from many different distributions give a broader picture of the relative value of the beta-binomial and the finite mixture models, and provide some preliminary insights into the situations in which these models are useful.

Key words: Beta-binomial; Capture–recapture; Closed populations; Finite mixture; Heterogeneity; Latent class; Mixture model; Species richness.

1. Introduction

Dorazio and Royle (2003), henceforth D&R, explored the behavior of three mixture models for the probability of capture of animals sampled from a closed population. These models are used to allow for heterogeneity of capture among animals as the primary source of variation in capture rates, model M_h in Otis et al. (1978). The number of times animal i is caught is assumed to have a binomial distribution, with capture probability p_i for this animal over k independent samples. The objective is to estimate the total number of animals, N, including those not seen in any sample.

D&R tested models in which p_i had a beta, a logit-normal, or a finite mixture distribution (in which animal i is in class c with probability π_c, $c = 1, \ldots, C$, where within class c the capture probability is a constant θ_c). Details of these models and their multinomial likelihoods are in D&R, Norris and Pollock (1996), and Pledger (2000). However, D&R's recommendation to use the beta distribution was based on restricted simulations, from the beta distribution. Since analysis using a distribution that matches the generating distribution is doomed to success, simulations from a wider range of generating distributions are needed to give a more comprehensive view of the issues of bias and precision of $\hat{N}$ and the coverage of nominal 95% confidence intervals.

Section 2 reports on N estimates from simulations with 18 generating distributions and Section 3 gives interval estimates and further details. Real data sets are considered in Section 4, and Section 5 has discussion of the simulation findings and other comments.

2. All Models Are Wrong

Simulations were done with 18 different generating distributions, covering a wide range of shapes and moments. An appropriate heterogeneity coefficient, used by D&R, is

$$\eta = \frac{\sigma^2}{\mu(1-\mu)},$$

the variance as a proportion of the maximum variance for a distribution on [0,1] with mean μ. The skewness coefficient is

$$\gamma_1 = \frac{E\left[(X-\mu)^3\right]}{\sigma^3}.$$

Each simulation had true $N = 100$, $k = 6$ samples, and 1000 replicated data sets; each generated data set was fitted using models M(0) (null model, no heterogeneity of capture, p_i = constant p), M(h_β) (beta model, individual capture probabilities from a beta distribution model, parameters α and β), and M(h_2) and M(h_3) (the two- and three-point finite mixture, or latent class, models). The generating distributions for individual p_i are shown in Table 1. In Group A, both μ and η are low. The beta distribution (A1) has probability density function (pdf) $f(x) \to 0$ as $x \to 0$ (as the first parameter α is greater than 1). The three two-point mixtures A2–A4 cover cases with skewness coefficient γ_1 less than and greater than the skewness of the corresponding beta distribution (with matching mean and variance). A data-generating distribution with more parameters is tried (the four-point mixture, A5, with seven parameters), and A6, the Uniform[0,0.3] distribution, is included as it is continuous but not a beta distribution. Group B has similar generating distributions but with higher mean and heterogeneity; a quadratic rather than uniform distribution is needed for B6 to obtain a high enough variance. An enormous challenge to capture–recapture analysis is provided in the Group C distributions, which have low mean and

Table 1

Generating distributions for individual capture probability p_i. The mean (μ), variance (σ^2), heterogeneity coefficient (η), and skewness coefficient (γ_1) are given.

Distribution	Details	μ	σ^2	η	γ_1
Group A					
A1. Beta	$B(1.76, 9.99)$	0.15	0.010	0.08	1.02
A2. Two-point	$\pi = (0.942, 0.058)$, $\theta = (0.125, 0.552)$	0.15	0.010	0.08	3.78
A3. Two-point	$\pi = (0.5, 0.5)$, $\theta = (0.05, 0.25)$	0.15	0.010	0.08	0.00
A4. Two-point	$\pi = (0.964, 0.036)$, $\theta = (0.131, 0.669)$	0.15	0.010	0.08	5.00
A5. Four-point	$\pi = (0.4, 0.1, 0.1, 0.4)$ $\theta = (0.05, 0.1, 0.2, 0.25)$	0.15	0.009	0.07	0.00
A6. Uniform	$a = 0$, $b = 0.3$ on [0,b]	0.15	0.010	0.08	0.00
Group B					
B1. Beta	B(1.31, 3.94)	0.25	0.030	0.16	0.80
B2. Two-point	$\pi = (0.866, 0.134)$, $\theta = (0.182, 0.690)$	0.25	0.030	0.16	2.14
B3. Two-point	$\pi = (0.5, 0.5)$, $\theta = (0.077, 0.423)$	0.25	0.030	0.16	0.00
B4. Two-point	$\pi = (0.916, 0.084)$, $\theta = (0.198, 0.822)$	0.25	0.030	0.16	3.00
B5. Four-point	$\pi = (0.4, 0.1, 0.1, 0.4)$ $\theta = (0.06, 0.2, 0.3, 0.44)$	0.25	0.029	0.15	0.00
B6. Quadratic	$f(x) = 85.7(x - 0.4)^2$ on (0.1, 0.6)	0.26	0.030	0.16	1.04
Group C					
C1. Beta	B(0.49, 2.76)	0.15	0.030	0.24	1.54
C2. Two-point	$\pi = (0.935, 0.065)$, $\theta = (0.104, 0.807)$	0.15	0.030	0.24	3.53
C3. Exponential	$\lambda = 6$, truncated to (0,1]	0.16	0.030	0.22	1.68
C4. Log	$f(x) = -\log(x)$ on (0,1]	0.25	0.050	0.27	0.89
C5. Beta mix	50:50 B(0.43, 8.08) and B(9.13, 27.38)	0.15	0.015	0.12	0.27
C6. Beta mix	50:50 B(0.81, 4.57) and B(3.63, 6.74)	0.25	0.030	0.16	0.48

high heterogeneity (C1–C3) and/or $f(x) \to \infty$ as $x \to 0$ (C1, C4–C6). The truncated exponential distribution, C3, has pdf

$$f(x) = \frac{\lambda e^{-\lambda x}}{1 - e^{-\lambda}}$$

on (0,1], $\lambda = 6$.

The N estimates, averaged over the 1000 simulated populations, are given in Table 2, with the M(h_3) results omitted as they largely duplicated the M(h_2) results. As one might expect, with the beta generating distribution the beta model shows less bias in $\hat{N}$ than the two-point, while with the two-point generating distribution the two-point model shows less bias than the beta. Details of the relative importance of the shapes and moments of the generating distributions in predicting bias were obtained from analyses of covariance with response variable the relative bias, $(\hat{N} - N)/N$. The covariates tried were mean μ, variance σ^2, skewness coefficient γ_1, and heterogeneity coefficient η. The first factor tried in the analysis of covariance classified the distributions by distribution type (DT). This factor DT grouped the generating distributions by the amount of probability near zero, with DT = 1 for distributions bounded away from zero (A2–A5, B2–B6, C2), DT = 2 for $f(x) \to c$ (finite) as $x \to 0$ (A1, A6, B1, C3), and DT = 3 for $f(x) \to \infty$ as $x \to 0$ (C1, C4–C6). An alternative factor was tried, using an indicator of whether or not the generating distribution was bounded away from zero. This was labeled DB (for "distribution bounded") with DB = 1 for simulations A2–A5, B2–B6, and C2, otherwise DB = 0. A third alternative factor used an indicator of whether the generating distribution was continuous or discrete. This was labeled DC

Table 2

Average N estimates and their average estimated standard errors over 1000 *replications. The generating distributions, all with true $N = 100$ and $k = 6$ samples, are specified in Table* 1. *Analysis was by three models,* M(0), M(h_β), *and* M(h_2), *and by model averaging* (*Mod. Av.*) *over those three models.*

Generating distribution	Mean $\hat{N}$ and (mean $\hat{se}(\hat{N})$)			
	M(0)	M(h_β)	M(h_2)	Mod. Av.
Group A				
A1. Beta	75 (8)	110 (59)	111 (99)	108 (86)
A2. Two-point	84 (10)	213 (170)	125 (78)	141 (109)
A3. Two-point	72 (7)	87 (30)	101 (79)	96 (63)
A4. Two-point	86 (10)	252 (202)	120 (57)	138 (93)
A5. Four-point	75 (6)	88 (22)	99 (55)	94 (65)
A6. Uniform	78 (8)	90 (27)	106 (94)	98 (67)
Group B				
B1. Beta	76 (3)	103 (33)	92 (29)	96 (34)
B2. Two-point	84 (4)	286 (255)	103 (17)	137 (88)
B3. Two-point	73 (3)	83 (13)	103 (72)	100 (63)
B4. Two-point	87 (5)	418 (371)	102 (10)	127 (61)
B5. Four-point	73 (3)	82 (12)	96 (56)	93 (49)
B6. Quadratic	82 (4)	128 (57)	108 (41)	115 (54)
Group C				
C1. Beta	51 (4)	114 (101)	76 (58)	87 (80)
C2. Two-point	67 (6)	596 (436)	107 (28)	120 (53)
C3. Exponential	62 (5)	161 (147)	91 (57)	109 (95)
C4. Log	67 (2)	109 (55)	82 (22)	92 (39)
C5. Beta mix	63 (5)	72 (22)	81 (89)	77 (44)
C6. Beta mix	75 (3)	91 (20)	94 (38)	95 (38)

for "distribution continuous," with DC = 1 for simulations A1, A6, B1, B6, C1, and C3–C6, and DC = 0 otherwise. Using the Akaike Information Criterion (AIC; Akaike, 1973; Burnham and Anderson, 2002), we find that models M(h_β) and M(h_2) have their bias in $\hat{N}$ dominated by different aspects of the generating distribution. For the beta model, the most important single predictor of bias is the "true" skewness coefficient (AIC 11.0 lower than for the null model), with underestimation of N if the skewness is zero, and an increase to zero bias and then overestimation of N as skewness increases (as seen in Tables 1 and 2). After allowing for skewness, no further predictors are supported by the modeling. For the two-point model, the most important single effect for predicting bias is the DT (AIC lower than for the null model by 11.3). Underestimation of N is well known to be associated with DT = 3, a large proportion of animals with very low capture probability (see, e.g., Otis et al., 1978; Coull and Agresti, 1999), and this is confirmed here (Table 2, C1, C4–C6). After this, inclusion of skewness lowers AIC by 5.6. No further predictors are supported.

Underestimation by the two-point model can occur because it does not adequately model a lot of mass of $f(x)$ near 0; this occurs in simulations B1, C1, and C3–C5 in Table 2. D&R's simulations were similar to cases B1 and C1, and they noted this bias of the two-point model and the better performance of the beta model. However, the extra simulations here show that the beta model overestimates N and the two-point model is preferable when the generating distribution is clear of zero but has high skewness coefficient (cases A2, A4, B2, B4, and C2). In other cases $\hat{N}$ has more serious underestimation from the beta model than from the two-point model, possibly because of low true skewness (A3, A5, B3, B5, C5, and C6). In all the simulations, we note that the heterogeneous models have much higher estimated standard errors of $\hat{N}$ than M(0).

Model selection by AIC (Akaike, 1973; Burnham and Anderson, 2002) was used on each generated data set to select a "best" model from M(0), M(h_β), M(h_2), and M(h_3) (the three-point model). Using the best model for the N estimate gave slightly less bias and narrower intervals throughout. However, this method tends to lead to standard errors that are too low, due to overfitting, so model averaging, which incorporates model uncertainty into the estimation, was also tried (Burnham and Anderson, 2002). This also gave a slight reduction in the bias of $\hat{N}$, compared with the individual models, as shown in Table 2.

Confidence intervals are discussed next.

3. Some Models Are Useful

However, broader considerations show a less gloomy picture of closed population abundance estimation by heterogeneous capture–recapture models.

The simulations in Section 2 have generating distributions and parameters deliberately chosen to make unbiased N estimation difficult for the heterogeneous models (and virtually impossible for the homogeneous null model). Many of the generating distributions had a large proportion of animals with near-zero capture probability, which has long been recognized as treacherous territory for capture–recapture analysis. This is signaled in the model fitting by very flat profile likelihood curves for N with associated wide profile likelihood intervals, and by high standard errors for $\hat{N}$ (as in Table 2).

Table 3

Coverage of nominal 95% *profile likelihood intervals from three models, and confidence intervals from model averaging (Mod. Av.). The poor coverage from the model averaging results from the use of Wald-type confidence intervals, which are not shown for the three basic models. True N is* 100, *there are* 6 *samples, and* 1000 *replications in each simulation.*

Generating distribution	Coverage of 95% PLI			95% CI
	M(0)	M(h_β)	M(h_2)	Mod. Av.
Group A				
A1. Beta	0.313	0.959	0.976	0.719
A2. Two-point	0.622	0.886	0.950	0.943
A3. Two-point	0.233	0.906	0.987	0.513
A4. Two-point	0.674	0.809	0.952	0.971
A5. Four-point	0.306	0.917	0.989	0.505
A6. Uniform	0.423	0.959	0.995	0.583
Group B				
B1. Beta	0.014	0.927	0.865	0.717
B2. Two-point	0.203	0.589	0.961	0.984
B3. Two-point	0.001	0.723	0.946	0.555
B4. Two-point	0.377	0.337	0.956	0.985
B5. Four-point	0.001	0.680	0.948	0.512
B6. Quadratic	0.082	0.928	0.966	0.919
Group C				
C1. Beta	0.000	0.953	0.762	0.588
C2. Two-point	0.132	0.296	0.955	0.990
C3. Exponential	0.013	0.958	0.827	0.766
C4. Log	0.000	0.926	0.668	0.613
C5. Beta mix	0.023	0.764	0.965	0.303
C6. Beta mix	0.007	0.848	0.881	0.582

It is not enough to look only at the point estimate of N and its average bias, as in Section 2. Users of capture–recapture models should be encouraged to look at the standard errors in $\hat{N}$ and the profile likelihood intervals (PLIs), and not place too much emphasis on a single point estimate. Wald-type confidence intervals ($\hat{N} \pm z_{\alpha/2}\hat{\text{se}}(\hat{N})$) can be nonsense, with the lower limit below the number of animals actually observed (see the size of the estimated standard errors in Table 2); profile likelihood intervals are recommended (Cormack, 1992; Coull and Agresti, 1999). Table 3 gives the coverage (proportion of PLIs covering the true value of $N = 100$) for the simulations in Section 2. As we might expect, the beta model gives the best coverage with beta-generated data (as in D&R), and the two-point model is better with a two-point generating distribution. Overall Table 3 shows remarkably good coverage from the heterogeneous models, even in these testing conditions, and especially when compared with the null model M(0).

The Wald-type confidence intervals ($\hat{N} \pm z_{\alpha/2}\hat{\text{se}}(\hat{N})$) gave poor coverage. Since this is the type of confidence interval currently available with model averaging, the poor coverage carried over to the model-averaging estimation (Table 3). A reviewer suggested that an improvement is likely if a log-based interval is used, and that there are other ways to fix this

problem. The development of a profile likelihood type of interval for model-averaged estimates would be very useful.

With many data sets, there will be some heterogeneity, but the extreme conditions of the simulations in Section 2 will not apply. Six more simulations (not shown here) were done, still with $N = 100$ and $k = 6$ occasions, but with less heterogeneity. Both the beta and the two-point models performed better, with the same patterns of bias in $\hat{N}$ as before, but less severe.

If we have higher N, more samples, and/or higher capture probabilities than in the Section 2 simulations, the PLIs are narrower and the heterogeneous models provide useful estimates. Simulations from distributions A1–A6 in Table 1 were repeated with $N = 200$ and $k = 10$ samples. The results (not shown here) when compared with A1–A6 showed small reductions in relative bias and substantial reductions in $\hat{\text{se}}(\hat{N})$ (overall, 36% reduction with $\text{M}(h_\beta)$, 37% with $\text{M}(h_2)$, although only 14% with M(0)). There was a concomitant sharper peaking of the profile likelihood curves and narrowing of the PLIs for N.

4. Analyses of Real Data

We now consider the real data sets discussed by D&R, and one extra data set.

4.1 *Snowshoe Hares*

The snowshoe hare data (Otis et al., 1978) discussed by D&R had $\hat{N}$ at 76.7 for the two-point model, and 90.8 for the beta model. With AIC for the two-point model being only 1.5 lower than for the beta model, a clear choice between these two models is not possible. One scenario could have "truth" with a beta distribution, and the two-point model underestimating N (cf. simulation run B1), while another could have the "true" distribution bounded away from zero but with high positive skewness, making the beta model overestimate N (cf. B2 and B4). More data would be needed to distinguish among these and other scenaria. Interestingly, Cormack (1992) analyzed this data set by removing the two animals seen every time, analyzing the remaining data with the null model, then reincluding the two high-capture hares, giving $\hat{N} = 77$ and a PLI of [70, 87]. The two-point model has $\hat{N} = 76.7$ and a PLI of [71, 87], almost a perfect match. A check of the posterior allocation of hares to the two groups shows the two high-capture hares allocated to one group with capture probability $p = 0.973$, and the remaining 66 hares to the other group, with $p = 0.295$; this is effectively the same as Cormack's ad hoc model, which predated the finite mixture models by several years. This does not of course confirm the latent class model as any more correct than the beta model (D&R) or the logit-normal model employed by Coull and Agresti (1999). With only 68 hares caught and six samples, the data set is sparse (as noted by Cormack, 1992), and more samples would be desirable. The sensitivity of the heterogeneous models to sampling fluctuations in the vector of capture frequencies is easily seen if we pretend that one hare caught every time was caught only five times, changing the capture frequency vector from $\mathbf{n} = (25, 22, 13, 5, 1, 2)$ to $(25, 22, 13, 5, 2, 1)$. For the modified data, the beta model now has $\hat{N} = 86.5$ with PLI [73, 162], and the two-point model has $\hat{N} = 77.5$ with PLI [71, 112]. These two models are now giving a more consistent picture. The sensitivity of the N estimates to minor sampling fluctuation in the capture frequencies is reduced with more data. Since increasing N is scarcely an option, more samples could be used to try to resolve the analysis.

4.2 *Species of Breeding Birds*

The North American Breeding Bird Survey data in D&R has similarly sparse data with sampling fluctuations in the capture frequency vectors. AIC values are too close to distinguish between $\text{M}(h_\beta)$ and $\text{M}(h_C)$ (where C is the number of latent classes indicated by the method of Norris and Pollock, 1996). For the five data sets, $\text{M}(h_\beta)$ has α estimates ranging from 0.296 to 0.409, all less than one, and in all cases $\text{M}(h_\beta)$ is giving higher N estimates than $\text{M}(h_C)$. We cannot distinguish among (i) a scenario with a lot of species virtually undetectable (with the finite mixture underestimating N, and the bias from the beta model depending on "true" skewness, cf. simulation runs C1 and C4), (ii) a scenario with no very low detection probabilities but possibly some other feature such as high enough skewness to make $\text{M}(h_\beta)$ overestimate N (cf. runs A2, A4, B2, B4, and C2), or (iii) some other scenario not simulated here which could have either or both models giving biased point estimates for N.

4.3 *Cottontail Rabbits*

Another data set traditionally used to test heterogeneous models is that of the cottontail rabbits (Edwards and Eberhardt, 1967; Otis et al., 1978). Here N is known to be 135, as the penned rabbits were able to be counted. The null model gives $\hat{N} = 96.3$ (se 6.9) with PLI [88, 107], the beta model gives $\hat{N} = 247.1$ (se 204.0) with PLI [111, 6947], and the two-point latent class model gives $\hat{N} = 135.5$ (se 36.6) with PLI [104, 347]. The beta model has $\hat{\alpha} = 0.389$, the value below one indicating it is fitting a distribution with a large proportion of animals with capture probability near zero; this accords with the high N estimate and PLI upper limit. Since N is known, we may construct an empirical probability distribution for p_i, in which each animal (caught or uncaught) has its p_i taken to be no. captures/k. For example, the 43 rabbits caught once in the 18 samples give a proportion 43/135 having capture probability $X = 1/18$. This distribution of X has $\mu = 0.0584$, $\sigma^2 = 0.0056$, and $\gamma_1 = 1.8317$. For comparison, the fitted beta distribution has $\hat{\mu} = 0.0319, \hat{\sigma^2} = 0.0016$, and $\hat{\gamma}_1 = 10.3890$, indicating reasonable estimates of μ and possibly σ^2 but a complete mismatch of skewness. By contrast, the two-point model (which has an accurate N estimate) has $\hat{\mu} = 0.0582, \hat{\sigma^2} = 0.0027$, and $\hat{\gamma}_1 = 1.9097$, a much better fit for the skewness. This looks like a case where the lack of a third parameter in the beta distribution counts against it. It is interesting that the model with only two support points is so effective in representing an empirical distribution with 19 support points. A similar effect is observed when the simulating distribution has four points but a two-point model is used (Tables 2 and 3, distributions A5 and B5). There are often diminishing returns in adding extra components to the finite mixture after the first few moments have been allowed for, as suggested by Lindsay (1995) and Norris and Pollock (1996). However, I am not suggesting the two-point model is "correct" for these data, or any other data set. For this example, there is an element of luck in the latent class model

giving such an accurate point estimate of N, as the data are sparse and the profile likelihood curve is quite flat. The M(h_2) N estimate would have been far more impressive if the profile likelihood curve had been more peaked and if $\hat{\text{se}}(\hat{N})$ had been lower.

5. Discussion

5.1 *Simulation Findings*

The simulations in Sections 2 and 3 by no means represent a broad sample of possible generating distributions; rather, they have been chosen to investigate the importance of true distribution shape when fitting the beta and two-point finite mixture models. I did not include a big range of μ and σ^2 values, even though these are important—with higher μ and lower σ^2 the heterogeneous models give less biased and more precise N estimates.

The simulations confirmed that (with low μ and moderate to high heterogeneity) if the generating distribution has a lot of probability near zero, the latent class models will underestimate N (as found by D&R) and may have poor coverage. However, D&R's recommendation of using the beta model should be treated with caution, as the bias and coverage depend crucially on the skewness of the "true" distribution. The importance of skewness is not surprising, as we are estimating the number of unseen animals, which are mainly clustered at the low end of the distribution of p_i.

It is essentially an artifact to assume there are two groups of animals, but assuming the generating distribution is of a particular shape (e.g., beta) is equally artificial. However, although these models are wrong, they may give realistic and useful PLIs for N, even in the unfavorable circumstances chosen for these simulations (Table 3), and in more favorable simulations with higher μ (not shown here) the bias in $\hat{N}$ is much reduced.

5.2 *Finite versus Infinite Mixtures*

D&R in their Section 5.1 suggested it was the finite support of the latent class models which made them perform poorly with their beta-based simulations, saying that "adding support points potentially allows finite mixtures to better approximate a highly variable, latent distribution of capture rates," that "as heterogeneity in capture rates increases," finite mixture models with an "insufficient number of support points" may produce biased estimates of N or interval estimates of N that are too narrow and have poor coverage, and that "the beta-binomial and logistic-normal models obviously have an advantage in this situation. Such continuous mixtures are able to specify large variation in capture rates without increasing the number of parameters to be estimated and, in doing so, provide more accurate, if less precise, estimates of N."

There seem to be two concepts of heterogeneity here, one related to variance and the other being the number of support points. If heterogeneity is variance related ($\eta = \sigma^2/[\mu(1-\mu)]$ as in D&R Section 2), the two-point finite mixture can provide high heterogeneity without increasing the number of support points. In fact, for a distribution on [0,1] with mean μ, the highest variance $\mu(1-\mu)$ is attained by the two-point Bernoulli distribution with proportion μ at one and $1-\mu$ at zero. Increasing the number of support points (e.g., using the decision method in Norris and Pollock, 1996) occurs not to increase variance but to match moments of higher order than skewness; this is especially useful for a multimodal underlying distribution.

The number of support points is less relevant than variance in these heterogeneous models. With data generated by the infinite-support beta mixture (C5 and C6), the finite-support two-point mixture model is providing less bias and better coverage than the infinite-support beta model. Also, the factor of continuous versus discrete generating distribution for generating capture probabilities was not selected as important in the analyses of covariance, either for the beta or the two-point model. The two-point mixture does not in general fail when the true distribution has infinite support; it fails in the subset of such distributions where $f(x) \to \infty$ as $x \to 0$.

I prefer the heterogeneity measure to be based on variance, and not on the number of support points. This implies, for example, that a 50:50 finite mixture on support points (0.1, 0.9) has more heterogeneity than a uniform distribution on [0.49, 0.51], despite the latter having infinite support. Coull and Agresti (1999) also use variance to represent heterogeneity.

In view of simulations A5 and B5 (four-point mixtures), C5 and C6 (beta mixtures), and the representation of a 19-point empirical distribution by two points in the cottontail rabbit example, I suggest that the two-point support is often able to provide enough variability to model data from a distribution with more support points, even infinitely many. I am not suggesting that this model is correct, or that we need not try other models.

5.3 *Theoretical Issues*

All capture–recapture models for abundance estimation are sensitive to model structure; essentially it is a forecasting problem, estimating the number of uncaught animals based on some model for the capture patterns of the caught animals. We are not including in the N estimate any animals with zero capture probability; the animals must be "available for capture," although efforts to pin down this concept more precisely have not yet succeeded.

For animals with a very low capture probability, taking more samples (if this can be done while retaining closure) increases the chance of at least one capture. It has been suggested for M(h) that assuming the underlying distribution of p_i is bounded away from zero will remove the problem of different models fitting equally well but giving different $\hat{N}$. However Link (2003) has shown that with this bound on p_i, for a fixed number of samples k, distributions exist which have exactly matched probabilities of $1, 2, \ldots, k$ captures conditional on at least one capture, but different (unconditional) probabilities of zero captures. This means that for a given k we could have two models giving identically good fit but different N estimates, even if N is large. The impact of these findings on capture–recapture analysis has yet to be determined.

The main failures of the beta distribution detected in these simulations are associated with a mismatch of skewness. Perhaps a move to a generalized beta distribution on $[a, b] \subset [0, 1]$ would be useful, with two extra parameters. At least five samples would be needed for this model to be feasible. The

two-point distribution can allow for skewness, and we may increase the number of support points if the data warrant it (Norris and Pollock, 1996), which is effective with a multimodal true distribution. However, the finite mixture models (along with other capture–recapture models) fail if the true distribution of capture probabilities has a large mass near zero ($f(x) \to \infty$ as $x \to 0$).

Profile likelihoods provide useful information with the heterogeneous models; they tend to have appropriate coverage, and will be realistically wide if more data are needed for accurate estimates. PLIs are asymmetric, often with unusably high upper bounds; however, the more stable lower bound may be of greater interest (e.g., for endangered species studies, where there could be serious consequences if N is overestimated).

5.4 *The Way Forward?*

It is important to be aware of the possible failings of capture–recapture models. There are circumstances in which the models will fail to adequately estimate N, for example, a high proportion of animals having zero or very small probability of capture. Nevertheless, there are real data sets where the true value of N was later determined, and in which the heterogeneous models give profile likelihood intervals including N while the homogeneous models do not.

The particular problems discussed in this article will not be solved by abandoning heterogeneous models and reverting to M(0), M(t), M(b), and M($t + b$), as an assumption of no heterogeneity has more impact than an assumption about the shape of the heterogeneity (Table 3).

Fitting a range of models relying on differently shaped distributions is important—a commitment to one model (such as the beta) is unnecessarily limiting, and may result in seriously wrong estimates (as in the cottontail rabbit data). For likelihood-based models, the AIC will indicate whether there is heterogeneity (M(h) versus M(0)), and will compare different heterogeneous models. Model averaging could be used rather than just selecting one best-fitting model. If different models seem to fit equally well, for example with similar AIC, but give very different N estimates or PLIs, the abundance estimates cannot be trusted (cf. Link, 2003). More samples could help to distinguish the models, provided closure of the population is maintained.

If k is large, finite mixtures with more components may be selected (Norris and Pollock, 1996); their nonparametric flavor gives a range of shapes and copes with multimodality of the true distribution. However, underestimation of N will still occur if a large proportion of the population has near-zero capture probability.

We need more investigations of the situations in which various models perform well. Since capture–recapture models do not give useful abundance estimates if a large proportion of the animals are uncaught, it would be useful to identify signals from the data to warn us of such cases. Possible signals could be (i) $\hat{\alpha} < 1$ when the beta model is fitted, (ii) the smaller capture probability from M(h_2) being below some threshold (see Norris and Pollock, 1996), (iii) $\hat{N}_{2pt} << \hat{N}_{\beta}$, or (iv) either model having $\hat{N}/n > Q$ (where n animals were actually seen, and Q is some threshold). To overcome these problems, we need proper study design, sufficient effort to increase the lowest capture probabilities, and attempts to meet assumptions such as closure.

Despite all the problems raised in this article, heterogeneous models are working well for many data sets, and provide a distinct improvement over the homogeneous null model. The price we pay for moving to heterogeneous models is wider PLIs and higher standard errors in $\hat{N}$. I believe this is intrinsic to the heterogeneity situation, a view also taken by Coull and Agresti (1999). If the profile likelihood curve is too flat and the PLI too wide, more data will be needed to provide a more precise estimate of N.

Acknowledgements

The author thanks an associate editor and a reviewer for their many helpful comments, and Martin Ridout for his careful reading and feedback, including correction of mistakes in the tabled beta skewness values.

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, B. N. Petrov and F. Csaki (eds), 267–281. Budapest: Academiai Kiado.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York: Springer-Verlag.

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

Coull, B. A. and Agresti, A. (1999). The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics* **55,** 294–301.

Dorazio, R. M. and Royle, J. A. (2003). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics* **59,** 351–364.

Edwards, W. R. and Eberhardt, L. L. (1967). Estimating cottontail abundance from live trapping data. *Journal of Wildlife Management* **31,** 87–96.

Lindsay, B. G. (1995). *Mixture models: Theory, geometry and applications*. NSF-CBMS Regional Conference Series in Probability and Statistics 5. Hayward, California: Institute of Mathematical Statistics.

Link, W. A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics* **59,** 1123–1130.

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

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs* **62,** 1–135.

Pledger, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics* **56,** 434–442.

Received January 2004. *Revised July* 2004.
Accepted August 2004.

The authors replied as follows:

1. Introduction

Estimating the size N of a closed population in the presence of heterogeneity in detection probability p among individuals has a long history. In an earlier paper on this topic, Burnham (1972) suggested a natural candidate model wherein p is assumed to have a beta distribution. Burnham concluded that this model had poor operating characteristics and instead suggested a jackknife procedure (Burnham and Overton, 1978). Subsequently, other models were developed to specify latent variation in p including logistic-normal mixtures (Coull and Agresti, 1999; Fienberg, Johnson, and Junker, 1999) and finite-mixtures of discrete support points (Agresti, 1994; Norris and Pollock, 1996; Pledger, 2000). In our earlier paper (Dorazio and Royle, 2003), we suggested that Burnham's (1972) assessment of the beta-binomial (BB) mixture was overly pessimistic, and we demonstrated that the BB mixture performs reasonably well across a wide range of situations. This was the primary point of our paper. We assessed the performance of the BB estimator by fitting the BB model to data generated under that model. This approach is widely used throughout statistics as a means of evaluating the performance of model-based estimators of parameters, particularly in small samples. Furthermore, this approach has also been used to assess the performance of other models developed for estimating N in the presence of heterogeneous detection probabilities (e.g., Norris and Pollock, 1996; Coull and Agresti, 1999).

Pledger (2005) takes offense at this approach, noting that "Since analysis using a distribution which matches the generating distribution is doomed to success, simulations from a wider range of generating distributions are needed . . . ." We plead guilty on this count but emphasize that our primary intent was to evaluate the performance of the BB model's estimator of N. Regardless, Pledger's statement is not entirely accurate. We did not *only* consider data simulated from the BB mixture, but also simulated data from logistic-normal models and finite-mixture models (see Section 3.1, Dorazio and Royle, 2003). These simulations were used to illustrate the difficulty in selecting among classes of mixture models for computing accurate estimates of N. In particular, we showed that conventional goodness-of-fit statistics, such as deviance, cannot be relied on for selecting a model that produces valid inferences about N.

Pledger also states "D&R's recommendation to use the beta distribution was based on restricted simulations . . . ." In fact, we did not recommend use of the BB mixture to the exclusion of other classes of mixture models. Instead, we demonstrated that because estimates of N can be sensitive to model-specific assumptions about the latent distribution of p, certain circumstances should lead an investigator to favor *continuous* mixtures over finite mixtures, and on this point our opinion remains unchanged. (We will have more to say about this issue in Section 3.)

2. Pledger's Simulations

Now that we have clarified the fundamental misunderstandings that motivated Pledger's (2005) commentary, we would like to provide some insights and alternative interpretations of her simulation results. In these simulations, 18 data-generating models were used to compare the performance of the BB and 2-point finite mixture (2PT) models in estimating N. A discrete support of 2 or 4 mass points was used in 9 of the 18 data-generating models; the remaining models used a continuously varying but compact support ([0, 1] or a subset of this interval) to specify variation in capture probability among $N = 100$ individuals observed on $T = 6$ sampling occasions. The data-generating models were "deliberately chosen to make unbiased N estimation difficult for the heterogeneous models" in the sense that "many of the generating distributions had a large proportion of animals with near-zero capture probability . . ." (Pledger, 2005).

Using simulation-based estimates of bias and confidence interval coverage for N, Pledger (2005) concludes that the 2PT model is superior to the BB model in those cases where there is substantial skewness in the true latent distribution of p and where the probability density of p does not approach ∞ as $p \to 0$. We disagree with this conclusion and with its implied generality. Among the 18 data-generating models used in the simulation study, we demonstrate here that neither the BB model nor the 2PT model provides a clear inferential advantage in terms of N estimation. In particular, we use Link's (2003) approach and reasoning to evaluate the bias of an approximating model (BB or 2PT) relative to each data-generating model. An advantage of this approach is that it isolates the component of bias in N that arises solely from differences in model-specific assumptions about the form of the latent distribution of p. Additional sources of bias (such as small sample size or method of estimation) that may occur in analysis of data (actual or simulated) are conveniently excluded.

To apply Link's approach, we computed the unconditional probability of being captured x times ($\pi(x)$, $x = 0, \ldots, T$) given a data-generating model M(h) of the latent variation in p. Using these probabilities, we computed the conditional probability of being observed x times given that an animal has been captured at least once ($\pi^C(x) = \pi(x)/(1 - \pi(0))$, $x = 1, \ldots, T$). We then computed the parameter values of the approximating model M(h_m) (where $m = \beta$ denotes BB and $m = 2$ denotes 2PT) that minimizes the Kullback–Leibler distance between its conditional probabilities of capture, say $\pi^C_m(x)$, and those of the data-generating model. These parameter values then were used to calculate the unconditional probabilities of capture $\pi_m(x)$ of the approximating model and the noncentrality of the comparison between models M(h) and M(h_m) as follows: $\lambda_m/n = 2\sum_{x=1}^{T} \pi^C(x) \log(\pi^C(x)/\pi^C_m(x))$. The noncentrality parameter λ_m may be used to calculate the power of an α-level test for distinguishing the approximating model M(h_m) from the data-generating model M(h), given n, the total number of animals captured (Link, 2003). We computed power by substituting (for n) the expected number of animals captured under the data-generating model, i.e., $E(n) = N(1 - \pi(0))$ where $N = 100$. We evaluated the bias of the approximating model by computing the discrepancy between $\pi_m(0)$, the probability of not being captured under model M(h_m), and $\pi(0)$, the "true" probability of not being captured.

Our assessment of the BB and 2PT models is summarized in Tables 1 and 2, respectively. Comparisons, where

Table 1

Comparison of the BB approximating model and various data-generating models $M(\mathrm{h})$ considered in Table 1 of Pledger (2005). $\log(\pi_\beta(0)/\pi(0))$ indicates the relative bias in estimating $\pi(0)$ that is produced by using model $M(\mathrm{h}_\beta)$ to approximate model $M(\mathrm{h})$. Power corresponds to an $\alpha = 0.05$ test for distinguishing model $M(\mathrm{h}_\beta)$ from each data-generating model.

$M(h)$	$\pi(0)$	$\pi_\beta(0)$	$\log(\pi_\beta(0)/\pi(0))$	λ_β/n	Power
A2	0.423	0.871	0.72	0.00577	0.067
A3	0.457	0.334	−0.31	0.00081	0.052
A4	0.415	>0.999	0.88	0.01876	0.111
A5	0.445	0.352	−0.23	0.00050	0.051
A6	0.437	0.358	−0.20	0.00032	0.051
B2	0.260	0.752	1.06	0.01277	0.102
B3	0.328	0.186	−0.56	0.00999	0.086
B4	0.244	0.972	1.38	0.05215	0.316
B5	0.326	0.181	−0.59	0.00632	0.072
B6	0.262	0.377	0.36	0.00146	0.055
C2	0.484	>0.999	0.73	0.14020	0.556
C3	0.480	0.623	0.26	0.00164	0.054
C4	0.370	0.394	0.06	0.00001	0.050
C5	0.492	0.287	−0.54	0.00020	0.050
C6	0.316	0.237	−0.29	0.00027	0.051

approximating and data-generating models are identical, are deliberately omitted because in those cases the noncentrality of the comparison is zero (which implies a constant power of α) and there is no bias induced by model misspecification. Our analysis of the BB and 2PT models suggests that there is virtually no power (above the nominal α level) to distinguish either of these models from the data-generating models, with the exception of two data-generating models (B4 and C2) where inadequacy of the BB model's approximation can be detected with a power of 0.316 and 0.556, respectively. Our power calculations indicate that the BB and 2PT models are often indistinguishable in terms of the observable data, i.e., one can often find parameter values of either model that imply a nearly identical set of conditional capture probabilities $\pi_m^C(x)$ $(x = 1, \ldots, T)$. In this situation, an inferential problem occurs in terms of N estimation if the BB and 2PT models have different *unconditional* capture probabilities for the unobserved individuals because a difference in these probabilities implies a difference in N estimates. This was the main point of Link's (2003) illuminating article. In Tables 1 and 2, it is clear that when the BB and 2PT models are indistinguishable (in the sense described earlier), they often have very different unconditional probabilities of capture $\pi_m(0)$. *In these cases, either model may be used as a data-generating model and induce a "bias" in the other (approximating) model's estimate of* N; therefore, one cannot claim inferential superiority for either model in these cases.

Table 2

Comparison of the 2PT approximating model and various data-generating models $M(\mathrm{h})$ considered in Table 1 of Pledger (2005). $\log(\pi_2(0)/\pi(0))$ indicates the relative bias in estimating $\pi(0)$ that is produced by using model $M(\mathrm{h}_2)$ to approximate model $M(\mathrm{h})$. Power corresponds to an $\alpha = 0.05$ test for distinguishing model $M(\mathrm{h}_2)$ from each data-generating model.

M(h)	$\pi(0)$	$\pi_2(0)$	$\log(\pi_2(0)/\pi(0))$	λ_2/n	Power
A1	0.447	0.367	−0.20	0.00008	0.050
A5	0.445	0.425	−0.05	1.2e-7	0.050
A6	0.437	0.370	−0.17	0.00019	0.051
B1	0.306	0.195	−0.45	0.00134	0.055
B5	0.326	0.259	−0.23	0.00002	0.050
B6	0.262	0.240	−0.09	0.00039	0.052
C1	0.551	0.275	−0.69	0.00314	0.058
C3	0.480	0.286	−0.52	0.00449	0.064
C4	0.370	0.175	−0.75	0.00605	0.073
C5	0.492	0.286	−0.54	0.00003	0.050
C6	0.316	0.181	−0.56	0.00066	0.053

We were puzzled initially that some of our analytical results appear to be inconsistent with Pledger's (2005; see Table 2) simulation results. Specifically, our analysis indicates that the 2PT model's estimate of $\pi(0)$ (and thus N) is negatively biased for each of the non-2PT, data-generating models (Table 2). In contrast, Pledger's simulation-based estimates of the mean $\hat{N}$ for these same data-generating models are either positively biased, negatively biased, or approximately unbiased. To discover the source of the discrepancy between our results, we repeated Pledger's simulation experiments. For the cases under consideration, we found that the sampling distribution of $\hat{N}$ (= the MLE of N under the 2PT model) is so highly skewed that the mean provides a poor measure of central tendency. Furthermore, a nonignorable proportion of simulated samples either failed to converge or produced an MLE near a boundary of the parameter space (e.g., a support point for p near 0 that contains almost all the mass). The latter problem produces estimates $\hat{N}$ that are orders of magnitude higher than N and that can profoundly influence a simulation-based mean if not removed. As an illustration, we simulated 20,000 samples using the Uniform(0,0.3) as a latent distribution for p (i.e., data-generating model A6). After collecting these samples into batches of 1000 replicates (= Pledger's sample size), we computed a mean $\hat{N}$ of 92.4 (Monte Carlo error = 1.1) by excluding 2.3% (on average) of the simulated samples owing to the estimation problems described earlier. We also calculated a median $\hat{N}$ of 82.9 (Monte Carlo error = 0.5), which is substantially lower than the mean owing to the high level of skewness in the sampling distribution of the estimates. The discrepancy between our simulated mean (92.4) and that reported by Pledger (106) exceeds Monte Carlo error and probably reflects differences in the criteria used to diagnose convergence failure and ill-conditioning of the parameter estimates. (Pledger, 2005 does not report these criteria in her simulation study.) Regardless, the important point is that the median $\hat{N}$, a better measure of central tendency, suggests that the 2PT model's estimator of N is negatively biased for data-generating model A6. This particular result, as well as results obtained by simulating samples from each of the other data-generating models (not reported here), is consistent with our analytical findings even though the simulation-based estimates of N no doubt contain additional sources of bias, such as those due to small sample size.

Our assessment of the BB and 2PT models indicates that for the data-generating models under consideration, neither

approximating model provides superior inferences about N. As population size N increases, however, we expect that the power to distinguish between these two approximating models and the data-generating models eventually will increase. But what is the value of these comparisons? For example, it should not be surprising that in sufficiently large populations a 2-point mixture will do a better job of fitting a bimodal generating distribution (e.g., C5 and C6) than a beta distribution—the 2-point mixture places a support point under each mode, whereas the beta distribution splits the difference. Is it really necessary to simulate these pathologies in order to confirm the obvious? We do not deny the relative advantage of the 2PT model in such pathological situations or in other situations where latent classes of individuals detected with different probabilities are thought to provide the primary sources of heterogeneity in p. We simply believe that analysts should carefully consider the sensibility of a model (and its underlying assumptions) within the context of the scientific problem before using the model for inference. We elaborate on this issue in the next section.

3. Choosing Sensible Mixture Models

It is well known that inferences about population size N can be sensitive to model-specific assumptions about the pattern of variation in individual capture probabilities (Coull and Agresti, 1999). An estimator of N essentially amounts to an extrapolation of the number of individuals that have not been captured using only information from those individuals observed in the sample (Fienberg, 1972); therefore, it is not surprising that the extrapolated estimate of the unobserved individuals can be sensitive to model structure.

In data analysis, an inferential problem occurs if different models can appear to fit the observed data reasonably well but yield dramatically different estimates of N. In this situation, we believe that conventional diagnostics used in model selection (such as deviance or Akaike's information criterion) cannot be relied on for selecting a class of mixture models that produces valid inferences about N (Dorazio and Royle, 2003). This view is supported by Link's (2003) thoughtful analysis which proves that N is not identifiable unless one assumes a particular class of distributions to specify the latent variation in p among individuals. Therefore, to compute a valid inference for N, we believe an analyst should carefully consider the distributional form of p (an unverifiable modeling assumption) prior to the analysis of data. An alternative strategy of fitting an arbitrary list of candidate models and then computing an estimate of N that averages over model uncertainty is, in our opinion, scientifically indefensible.

How should an analyst develop one or more plausible models for the heterogeneity in capture probabilities? A useful starting point involves careful consideration of the mechanisms that are likely to have produced different capture rates for different individuals. For example, if the population comprises different age classes or sexes that cannot be observed but are likely to have been captured at different rates, then finite mixtures of discrete support may provide reasonable models of the latent variation in p. On the other hand, if heterogeneity in p is caused by behavioral differences (e.g., activity patterns, habitat preferences, foraging preferences), differences in exposure of individuals to sampling relative to their territories, or other phenomena that are impractical or impossible to observe but likely to vary among individuals, then mixture models of continuous support are needed. In these populations, we reject the idea that a finite mixture of two support points provides a satisfactory approximation of the potentially infinite variation in capture probabilities, as suggested by Pledger (2000, 2005). However, it is conceivable that finite mixtures of continuous support can provide sensible models of latent heterogeneity in p if different groups of individuals are thought to have different capture rates and if heterogeneity in capture rates is thought to exist within each group. Fitting such models would no doubt present some challenges to an analyst.

References

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

Burnham, K. P. (1972). Estimation of population size in multiple capture-recapture studies when capture probabilities vary among animals. Ph.D. Thesis, Oregon State University, Corvallis.

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.

Coull, B. A. and Agresti, A. (1999). The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics* **55,** 294–301.

Dorazio, R. M. and Royle, J. A. (2003). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics* **59,** 351–364.

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

Fienberg, S. E., Johnson, M. S., and Junker, B. W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society, Series A* **162,** 383–405.

Link, W. A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics* **59,** 1123–1130.

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

Pledger, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics* **56,** 434–442.

Pledger, S. (2005). The performance of mixture models in heterogeneous closed population capture–recapture. *Biometrics* **61,** 868–873.

Robert M. Dorazio
U.S. Geological Survey
Florida Integrated Science Center
Gainesville, Florida 32653
email: *bdorazio@usgs.gov*

and

J. Andrew Royle
U.S. Geological Survey
Patuxent Wildlife Research Center
Laurel, Maryland 20708
email: *aroyle@usgs.gov*