Estimating the observable population size from biased samples: a new approach to population estimation with capture heterogeneity

James E. Johndrow ¹, Kristian Lum ², and Daniel Manrique-Vallier ³ §

¹Department of Statistical Science, Duke University
²Human Rights Data Analysis Group
³Department of Statistics, Indiana University

Wednesday 8th June, 2016

Abstract

Capture-recapture methods aim to estimate the size of a closed population on the basis of multiple incomplete enumerations of individuals. In many applications, the individual probability of being recorded is heterogeneous in the population. Previous studies have suggested that it is not possible to reliably estimate the total population size when capture heterogeneity exists. Here we approach population estimation in the presence of capture heterogeneity as a latent length biased nonparametric density estimation problem on the unit interval. We show that in this setting it is generally impossible to estimate the density on the entire unit interval in finite samples, and that estimators of the population size have high and sometimes unbounded risk when the density has significant mass near zero. As an alternative, we propose estimating the population of individuals with capture probability exceeding some threshold. We provide methods for selecting an appropriate threshold, and show that this approach results in estimators with substantially lower risk than estimators of the total population size, with correspondingly smaller uncertainty, even when the parameter of interest is the total population. The alternative paradigm is demonstrated in extensive simulation studies and an application to snowshoe hare multiple recapture data.

KEY WORDS: Capture-recapture; capture heterogeneity; identifiability; observable population size; population estimation; Bayesian nonparametrics; biased sampling.
1 Introduction

Capture-recapture is a class of statistical methods designed to estimate the size of a closed population. The data required for this task are $T$ incomplete but non-disjoint samples of individuals from the population. Each individual is identified across all lists to produce a “capture history”, a length $T$ binary vector giving the pattern of presence and absence on the lists. A joint model for multivariate binary outcomes is then fit to the capture histories of all individuals who appeared in the sample, and used to estimate the number of individuals not captured on any list. Crucially, in order to produce a useful estimate of the total population size $N$, the model must explain not only the observable data, but also the distribution of individuals not observed on any list. Since the model can only be fit to observed individuals, the quality of inference will depend heavily on the relationship between the observed and the non-observed individuals implied by the joint model.

The literature in capture-recapture methods has long recognized the need for specialized methods to deal with populations in which capture probabilities vary from individual to individual. This phenomenon is referred to as capture heterogeneity. In the absence of covariate information that can control for the differences in observability through, for instance, stratified estimation; see Sekar and Deming [1949], most proposed approaches rely on joint sampling models that include some form of individual random effects. The random effects model variation across individuals in the probability of being observed. Early proposals modeled the random effects using a parametric mixing distribution, often chosen because of its mathematical tractability (Sanathanan [1973], Agresti [1994], Darroch et al. [1993]). Other approaches have sought to estimate the mixing distribution in a non-parametric way (Mao [2008]).

The presence of capture heterogeneity is equivalent to bias in the sampling process, since individuals with smaller values of the random effect are necessarily underrepresented on the partial lists. Although not characterized as resulting from sampling bias, this situation was nonetheless recognized in a series of articles, notably by Huggins [2001] and Link [2003], that discussed the identifiability of $N$ in heterogeneous capture-recapture. In particular Link [2003] showed how simple parametric models produced almost identical fitted observed frequencies, while inducing completely different estimates for the number of unobserved individuals. Thus, while using parametric forms for the random effect distribution addresses the immediate issue of identifiability, when the model is misspecified, the resulting population estimates can be wildly inaccurate. This has led some authors to suggest estimating a lower bound on the total population size that is valid for parametric models satisfying certain conditions (for example, see Chao [1987]).

In this article, we show that population estimation with capture heterogeneity is analogous to a latent density estimation problem in the presence of length bias. Length, or size, bias is a type of sampling bias where the probability of observing data is proportional to the magnitude of the observation, as outlined in Patil and Rao [1977, 1978]. There is an extensive literature on density estimation in the presence of length bias, including Vardi [1985], Vardi [1982], Asgharian et al. [2002], de Uña-Álvarez [2004], Jones [1991], and Gill et al. [1988], among others. Curiously, the capture-recapture literature and the sampling bias literature are almost completely orthogonal, possibly because the canonical approach is to consider the individual list capture probabilities rather than the probability of capture on at least one list, as we do here. The only reference to length bias in the capture-recapture literature that we are aware of is Chen and Lloyd [2002], which is a very different setting from ours, requires the availability of individual covariates, and makes the assumption that individual capture probabilities are bounded away from zero. Therefore, a major contribution of this paper is to bring these two literatures closer by making explicit this connection and providing
theoretical results on the risk of capture-recapture estimates in the presence of capture heterogeneity that are a function of nonparametric density estimates under length bias sampling. A secondary contribution is to the literature on Bayesian nonparametric methods for biased sampling, which is relatively thin, see Hatjispyros et al. [2015] and Kunihama et al. [2014]. In this context, our method can be viewed as a Bayesian missing data approach to nonparametric density estimation under length bias.

Our second major contribution is to argue that, while this intrinsic limitation of capture-recapture methods in the presence of capture heterogeneity makes the problem of estimating $N$ essentially ill-posed, there are other meaningful, interpretable quantities related to total population size that can reliably be estimated in this scenario. We thus propose approaching capture-recapture estimation from a different perspective. Instead of estimating the total population size, we advocate estimating the observable population size, which, informally, is the size of the total population that has non-negligible probability of being observed. To make this approach rigorous, we define the concept of $\alpha$-observable populations, the population of individuals with probability at least $\alpha$ of being observed on at least one list. We show that, in addition to being properly defined quantities with an intuitively sound interpretation, estimators of the $\alpha$-observable population size have superior properties to estimators of $N$. Moreover, they often have lower risk as estimators of $N$ than the corresponding unbiased estimator of $N$. We propose a Bayesian approach to inference based on discrete mixtures, and show how this approach is ideally suited to estimating the size of $\alpha$-observable populations. We also outline a strategy for choosing the optimal value of $\alpha$ as a byproduct of model fitting.

2 Heterogeneity in Capture Recapture

Consider a sample of $m$ individuals captured or recorded from a population of unknown size $N$ during $T$ capture attempts. We represent each individual’s capture history using binary indicators, $x_{it}$ for $i = 1, \ldots, m$ and $t = 1, \ldots, T$, which take the value 1 if $i$th individual appeared on the $t$th list and 0 otherwise. These data can be summarized by a $2^T$ contingency table, with cell counts $n(x)$ for $x \in \{0, 1\}^T$ representing the number of individuals in the sample with observed capture history $x$. For example, $n\{(010)\}$ is the number of individuals captured on the second list, but not on the first or third. The count in the cell indexed by all zeroes, $n(\zeta)$ where $\zeta$ is the zero vector, is unobservable by definition, as it represents the total number of individuals in the population that were not recorded on any list. This value is the object of inference, since adding it to $m$ gives an estimate of the total population size $N$.

Capture heterogeneity refers to the case when the probability of capture varies across individuals in the population, resulting in a sample that is biased toward individuals with high capture probability. In our initial development, we abstract from the multiple list structure of the data, and consider only the individual probability of being observed on at least one list. We will demonstrate that knowledge of the distribution of this parameter is sufficient to estimate $N$, and consider a natural estimator.

Let $\pi_{x,i}$ be the probability that individual $i$ has capture history $x$, and $p_i = 1 - \pi_{\zeta,i}$ the probability of appearing in least one list. In the absence of heterogeneity $p_i = p$ for all $i$. For known $p$, the estimator $\hat{N} = mp^{-1}$ is a commonly used estimator of $N$. When $P \sim F$ for a probability distribution $F$ supported on the unit interval, an analogous estimator is

$$m \left\{ \int_{[0,1]} p F(dp) \right\}^{-1} = m\{E_F(P)\}^{-1};$$
when $F$ has a density $f$, we will sometimes write $E_f(P)$ instead of $E_F(P)$. The following remark shows that it is asymptotically equal to a regularized maximum likelihood estimate, where the regularization can be interpreted as a continuity correction for the parameter $N$.

**Remark 2.1.** Suppose $p < 1$. Consider the penalized Binomial log-likelihood with $N$ treated as a continuous parameter

$$
\hat{\ell}(m \mid N, p) = \log \left\{ \frac{\Gamma(N + 1)}{\Gamma(m + 1)\Gamma(N - m + 1)} \right\} + m \log(p) + (N - m) \log(p) - \frac{1}{2} \log \left( \frac{N}{N - m} \right). \tag{1}
$$

Asymptotically, $\hat{\ell}(m \mid N, p)$ is equal to a Binomial($N - 1/2, p$) log-likelihood, so the penalty is analogous to a continuity correction. Moreover, $\hat{N} = \arg\min_N \hat{\ell}(m \mid N, p)$ satisfies

$$
\hat{N} = mp^{-1} + O(m^{-1}). \tag{2}
$$

More generally if $P \sim F$ for a discrete measure with finitely many atoms, then $m \sim \text{Binomial}\{N, E_F(P)\}$, and $\hat{N} = \arg\min_N \hat{\ell}(m \mid N, p)$ satisfies

$$
\hat{N} \mid m, f = m\{E_F(P)\}^{-1} + O(m^{-1}).
$$

This suggests the estimator $m\{E_F(P)\}^{-1}$ for general $F$. So in the presence of capture heterogeneity, nonparametric estimation of $N$ and $\pi(\alpha)$ can be equated with estimation of the expectation of $P$ with respect to $F$, and the associated estimator of $N$ is $m$ divided by $E_F(P)$.

The problem of nonparametric estimation of $F$ differs from typical distribution estimation problems. For simplicity, suppose $F$ has density $f$. In the traditional density estimation setting, data in the neighborhood around the value $p$ appears in the sample in proportion to $f(p)$. In contrast, in the capture-recapture setting, the probability of observing data in a small neighborhood around $p$ is proportional to $pf(p)$. This type of sampling bias is referred to as length bias (Patil and Rao 1977). In this setting, the probability of observing data in the region near zero in finite samples is vanishingly small for bounded densities $f(p)$, even if $f$ has significant mass near zero. Figure 4 shows a graphical representation of this phenomenon, with the density represented on the logit scale for clarity. Specifically, $\eta = \varphi(p)$, where $\varphi : [0, 1] \to \mathbb{R}$ is the logit function. The left panel represents the “true” density $f^*(\eta)$. The center uses shading to represent observability $\varphi^{-1}(\eta)$ of the individuals with random effects in that region; individuals in the dark region are unlikely to be observed. As a result we have the situation depicted in the right panel: regardless of the true density, the three densities shown here fit the observable data equally well, but lead to a dramatically different characterization of the unobserved population.

### 3 Estimation of $\alpha$-observable populations

We now introduce an alternative inferential approach that estimates the $\alpha$-observable population size $N_\alpha$ rather than the total population size $N$. Definition 3.1 formalizes the concept of $\alpha$-observability.

**Definition 3.1.** Individual $i$ is $\alpha$-observable if $1 - \pi_{\alpha,i} = p_i > \alpha$. We define the $\alpha$-observable population size, $N_\alpha$, as $N_\alpha = \sum_{i=1}^{N} 1 \{ p_i > \alpha \}$. We call $[\alpha, 1]$ the $\alpha$-observable region. Any set such that $A \subset [\alpha, 1]$ is an $\alpha$-observable set. Any set that is not $\alpha$-observable is $\alpha$-unobservable.
Figure 1: Left panel: a generic density $f^*(\eta)$. Center panel: same density as in left panel, but with background shading proportional to $\varphi^{-1}(\eta)$ with $\varphi$ the logit function. Right panel: several possible densities that share the same shape in the “observable” region, but different shapes in the “unobservable” region.

In contrast to estimating $N$, estimating $N_\alpha$ requires learning the distribution $F$ only on $[\alpha, 1]$. Then the analogue of $\hat{N}$ is

$$\hat{N}_\alpha = m_\alpha \{1 - F(\alpha)\} \{E_F(P)\}^{-1} = m_\alpha \{E_{F,\alpha}(P)\}^{-1},$$

where $m_\alpha$ is the number of observed individuals having $p \in [\alpha, 1]$ and $E_{F,\alpha}(P) = E_F(P \mid P > \alpha)$. So to estimate $\hat{N}_\alpha$ we only need to integrate over $[\alpha, 1]$, not the entire unit interval.

## 3.1 Risk of estimators of $N$ and $N_\alpha$

We now show that estimators of $N_\alpha$ have lower risk than estimators of $N$. Intuitively, low risk of any nonparametric estimator $\hat{F}$ of $F$ requires that we observe data with high probability in sets whose $F$ measure is bounded away from zero. We refer to these sets as non-negligible.

**Definition 3.2** (Non-negligible sets). For any $\epsilon > 0$, a Borel set $A$ is non-negligible with respect to $F$ and $\epsilon$ if $\int_{p \in A} F(dp) > \epsilon$.

Theorem 3.3 shows that we can guarantee a uniform minimal probability of observing data in non-negligible $\alpha$-observable sets in finite samples, but not in all non-negligible sets. Therefore, non-parametric estimators of $N_\alpha$ are expected to have better finite-sample performance than estimators of $N$. All proofs are deferred to the Appendix.

**Theorem 3.3** (Finite sample bounds in $\alpha$-observable populations). Suppose the true population size is $N$ and fix $\epsilon > 0$. Let $A$ be the set of all $(F, \epsilon)$ non-negligible $\alpha$-observable intervals, and for any $A \subset [0, 1]$, let $M(A) := \sum_{i:p_i \in A} \mathbb{I} \left\{ \sum_{t=1}^T x_{it} > 0 \right\}$, the number of observations with $p \in A$. Then

$$\inf_{A \in A} \text{pr}\{M(A) > 0\} = 1 - \left(1 - \alpha \epsilon\right)^N. \quad (3)$$

That is, for fixed $N$, the probability that some data are observed in any non-negligible set is bounded below by the right hand side of (3), which is a function of $\alpha$. For $\alpha = 0$, the only uniform
bound is the trivial bound \( \inf_{A \in \mathcal{A}} \Pr\{M(A) > 0\} = 0 \). Another implication of Theorem 3.3 is that for fixed \( \alpha \) and \( \epsilon \) and any \( q \in (0,1) \), there exists \( N < \infty \) large enough that the probability of observing data in \( A \in \mathcal{A} \) is uniformly bounded below by \( q \). Of course, no finite value of \( N \) would be sufficient to guarantee a uniform lower bound on the probability of observing data in an arbitrary non-negligible subset of the unit interval. Finally, Theorem 3.3 suggests one approach for choosing values of \( \alpha \) based on the number of unique observed individuals \( m \). Although \( N \) is unknown, it is always the case that \( N \geq m \), so the probability of observing data in non-negligible \( \alpha \)-observable sets can be lower-bounded by \( \epsilon \) with \( N = m \). Therefore, one way to assess different choices of \( \alpha \) is to compute \( \mathbb{E} \) for the observed population size \( m \) and a range of \( \epsilon \) values.

Now we consider the risk of estimators of \( N \) and \( N_\alpha \). Throughout, we will assume that we observe directly length-biased data proportional to \( p F(dp) \). In capture-recapture applications, \( p \) is latent, so the results in this section provide a lower bound on the risk of capture-recapture estimates of \( N \). The distribution of the observed data is

\[
G(dp) = \frac{p F(dp)}{E_f(P)},
\]

(4)
since \( G(dp) \) must integrate to 1. Consider an estimator \( \hat{G} \) of \( G \). We propose to use

\[
\hat{F}(dp) = \frac{p^{-1} \hat{G}(dp)}{E_G(P^{-1})},
\]

(5)
as an estimator of \( F \). This is similar to the approach in [Hatjispyros et al. 2015]. This estimator is appropriate in at least one sense, which is to say that since \( F(dp) \propto p^{-1} G(dp) \), it must be the case that

\[
F(dp) = p^{-1} G(dp) E_f(P) = p^{-1} G(dp) E_G(P^{-1}),
\]
since \( E_G(P^{-1}) = \{E_f(P)\}^{-1} \), and \( \hat{F} \) is the plug-in estimator of \( F \) derived from any estimator of the observable distribution \( \hat{G} \).

We now consider two procedures for estimation of \( \hat{G} \). Critically, when \( F \) is unknown, we cannot estimate \( E_f(P) \) without implicitly choosing an estimator \( \hat{F} \) because of the length bias in the observed data. We consider two possible estimators of \( \hat{G} \): the empirical measure and a histogram estimator. These estimators are classical analogues of the Bayesian Dirichlet process model we propose in [4.2]. Since \( E_G(P^{-1}) = \{E_f(P)\}^{-1} \), in either case the estimator \( \hat{N} \) is given by \( m E_G(P^{-1}) \). When \( \hat{G} \) is the empirical measure, this is just the sample mean of \( p^{-1} \), \( \hat{N} = \sum_{i=1}^{m} \hat{p}_i^{-1} \), and we have the following result for the mean squared error of \( \hat{N} \).

**Theorem 3.4** (Mean squared error for empirical measure estimator). If \( \hat{G}(dp) = \sum_{i=1}^{m} \delta_{\hat{p}_i}(dp) \) then the mean squared error \( \Delta^2(\hat{N}, N) \) is given by

\[
\Delta^2(\hat{N}, N) = N \{E_f(P^{-1}) - 1\}.
\]

(6)

It is immediate that for \( \Delta^2(\hat{N}, N) < \infty \), we need \( E_f(P^{-1}) < \infty \). In many seemingly mundane cases this fails, for example when \( F \) is a Beta\((a,b)\) distribution with \( a \leq 1 \), which includes the uniform distribution on the unit interval. The following Corollary shows that the empirical measure estimator of \( N_\alpha \) always has finite \( L_2 \) risk, and the risk decreases with increasing \( \alpha \).
Corollary 3.5 (Mean squared error for empirical measure estimator of \(N_\alpha\)). For any \(\alpha \in (0,1]\) if \(\hat{N}_\alpha\) is the empirical measure estimator of \(N_\alpha\), then
\[
\Delta^2(\hat{N}_\alpha, N_\alpha) = N_\alpha \{E_{f_\alpha}(P^{-1}) - 1\} \leq N_\alpha (\alpha^{-1} - 1),
\]
and \(N_\alpha^{-1} \Delta^2(\hat{N}_\alpha, N_\alpha)\) is monotone nonincreasing in \(\alpha\).

Finally, in many cases \(\hat{N}_\alpha\) using the empirical measure estimator of \(F\) will have lower \(L_2\) risk for \(N\) than the corresponding empirical measure estimator of \(N\), and \(\hat{N}\) cannot be a minimax estimator of \(N\) in the nonparametric regime.

Corollary 3.6 (Mean squared error for estimation of \(N\) by the empirical measure estimator of \(N_\alpha\)). For any \(\alpha \in (0,1]\) if \(\hat{N}_\alpha\) is the empirical measure estimator of \(N_\alpha\), then
\[
\Delta^2(\hat{N}_\alpha, N) = N_\alpha \{E_{f_\alpha}(P^{-1}) - 1\} + (N - N_\alpha)^2 \leq N_\alpha (\alpha^{-1} - 1) + (N - N_\alpha)^2.
\]
Therefore, \(\Delta^2(\hat{N}_\alpha, N) < \Delta^2(\hat{N}, N)\) whenever \(N_\alpha \{E_{f_\alpha}(P^{-1}) - 1\} + (N - N_\alpha)^2 < N \{E_{f}(P^{-1}) - 1\}\). In particular, \(\hat{N}\) cannot be a minimax estimator of \(N\) when \(F\) is unknown, since there exists \(F\) for which \(\Delta^2(\hat{N}, N) = \infty\), while \(\Delta^2(\hat{N}_\alpha, N) < \infty\) for all \(F\).

We now consider the case where \(F\) has a density \(f\) and \(\hat{g}\) is the histogram estimator. The histogram estimator is a commonly used nonparametric density estimator. We have the following result for the risk of histogram-based estimators of \(N\).

Theorem 3.7 (Mean squared error of histogram estimators of \(N\)). Suppose \(\hat{g}\) is the histogram estimator with bin width \(h\). Assume \(f\) is twice continuously differentiable. Then the mean squared error \(\Delta^2(\hat{N}, N)\) satisfies
\[
0 \leq \Delta^2(\hat{N}, N) - N \left[ E_f(P^{-1}) - \{E_f(P)\}^{-1} \right] + \mathcal{O}(Nh) \leq \frac{N^2 h^2}{4} \left\{ E_f(P^{-1}) + E_f(\|f'(P)\|) \right\}^2 \tag{7}\]
Further, the asymptotic variance (as \(n \to \infty\) and \(h \to 0\)) of \(\hat{N}\) is given by \(N \left[ E_f(P^{-1}) - \{E_f(P)\}^{-1} \right]\), and the asymptotic bias is bounded above by \((N^2 h^2/4) \left\{ E_f(P^{-1}) + E_f(\|f'(P)\|) \right\}^2\).

From equation (7), it is clear that when \(h\) shrinks faster than \(N^{-1/2}\), the asymptotic mean squared error and asymptotic variance are identical. Moreover, for twice differentiable densities \(f\), we can improve upon the asymptotic mean squared error for the empirical measure estimator, since \(\{E_{f}(P^{-1}) - 1\} \geq \{E_{f}(P^{-1}) - \{E_{f}(P)\}^{-1}\}\). Nonetheless, \(\Delta^2(\hat{N}, N) < \infty\) still requires \(E_{f}(P^{-1}) < \infty\). The following Corollary shows that estimating \(N_\alpha\) has similar benefits when we use a histogram estimator of \(g\).

Corollary 3.8 (Mean squared error of histogram estimators of \(N_\alpha\)). Under the conditions of Theorem 3.7 if \(\alpha \in (0,1]\) and \(h = N^{-1+\delta/2}\) for \(\delta > 0\) then
\[
\lim_{N_\alpha \to \infty} N_\alpha^{-1} \Delta^2(\hat{N}_\alpha, N_\alpha) = \left[ E_{f_\alpha}(P^{-1}) - \{E_{f_\alpha}(P)\}^{-1} \right] < \alpha^{-1}.
\]
Furthermore, if \(E_{f_\alpha}(P^{-1}) - \{E_{f_\alpha}(P)\}^{-1}\) is monotone nonincreasing in \(\alpha\), then the asymptotic mean squared error of \(\hat{N}_\alpha\) is also monotone nonincreasing in \(\alpha\).
We conjecture that $E_{f_0}(P^{-1}) - \{E_{f_0}(P)\}^{-1}$ is monotone nonincreasing in general, though something stronger than the obvious convexity argument is required to show this result. Empirically, this holds for all of the distributions we have tested. We also have a result similar to Corollary 3.6 for the histogram estimator.

**Corollary 3.9** (Mean squared error for estimation of $N$ by histogram estimator of $\hat{N}_\alpha$). For any $\alpha \in (0,1]$ if $\hat{N}_\alpha$ is the histogram estimator of $N_\alpha$ and the conditions of Theorem 3.7 hold, then

$$\Delta^2(\hat{N}_\alpha, N) \leq \Delta^2(\hat{N}_\alpha, N_\alpha) + (N - N_\alpha)^2 < \infty$$

It follows that $\hat{N}$ cannot be a minimax estimator of $N$ when $F$ is unknown, since there exists $F$ for which $\Delta^2(\hat{N}, N) = \infty$, while $\Delta^2(\hat{N}_\alpha, N) < \infty$ for all $F$.

Theorems 3.4 and 3.7 show that in many cases, estimators of $N$ have unbounded risk, while Corollaries 3.5 and 3.8 guarantee that for any $\alpha > 0$, there exist estimators of $N_\alpha$ with finite risk for any $F$. Further, these results imply that the choice of $\alpha$ can be viewed as a tradeoff between higher mean squared error and estimating a quantity closer to the total population size. Finally, Corollaries 3.6 and 3.9 show that estimators of $N_\alpha$ often have lower risk for estimation of $N$ than the corresponding estimators of $N$ and are superior in the minimax sense. These are the basic lessons that motivates the methods we propose for estimation of $N_\alpha$ and choice of $\alpha$ in the sequel.

## 4 Model-based estimation of $N_\alpha$

Thus far, the discussion has centered on $p_i = 1 - \pi_{\zeta,i}$ without specifying a model for the other cell probabilities. In capture-recapture, we must use the data on presence or absence on the $T$ lists to estimate the latent distribution $F$ and the population size $N$. This is generally done by specifying a model on the individual-level cell probabilities $\pi_{x,i}$, which induces a model on $p_i = 1 - \pi_{\zeta,i}$, and on $F$ by marginalizing over individuals. We now restrict our attention to a specific class of such models, the $M^{th}$ class. In modeling, it is common to parametrize probabilities via a monotone nondecreasing transformation $\varphi^{-1} : \mathbb{R} \to [0,1]$, such as the logit or probit function. We will follow this convention for the remainder of the paper, and will write $\eta = \varphi(p)$ for the transformed individual observabilities. Densities and distribution functions supported on the real line induced by the transformation of $P \sim F$ via $\varphi$ are represented with superscript $\ast$, for example, $f^\ast(\eta)$ and $F^\ast$.

### 4.1 $M^{th}$ Model

The $M^{th}$ model (Agresti [1994]) is given by the following joint distribution for the list capture random variables $X_1, \ldots, X_T$:

$$\text{pr}(X_{i1} = x_1, \ldots, X_{iT} = x_T \mid \theta, \beta) = \prod_{t=1}^T \varphi^{-1}(\theta_t + \beta_t)^{x_{it}} \{1 - \varphi^{-1}(\theta_t + \beta_t)\}^{1-x_{it}}, \quad \theta_t \overset{iid}{\sim} G^\ast, \quad (8)$$

where the $\theta_t$ are individual-specific observability effects distributed according to a distribution $G^\ast$ supported on $\mathbb{R}$, and the $\beta_t$ are global list effects. By (8), individual $i$ is $\alpha$-observable in this model if

$$p_i = \varphi^{-1}(\eta_i) = 1 - \prod_{t=1}^T \{1 - \varphi^{-1}(\theta_t + \beta_t)\} > \alpha. \quad (9)$$
Clearly, the distribution $G^*$ on $\theta$ induces a distribution $F^*$ for $\eta = \varphi(p)$, the individual observabilities under the transformation $\varphi$. Thus, informally speaking, the parameter $\theta$ is analogous to the individual observability $\eta$. Some additional information on $\alpha$-observability and the induced distribution on $\eta$ are provided in Supplementary Materials.

Historically, the literature has focused on certain parametric choices for $G^*$ that lead to computational tractability. An extensive review is provided in Supplementary materials. To summarize, parametric choices for $G^*$ perform very poorly under model misspecification, since any parametric choice for $G^*$ implies a specific shape for the distribution in the unobservable region. As such, we focus on nonparametric estimation of $G^*$, consistent with our discussion thus far and with the work of Mao [2008]. In particular, we propose estimators of $N_\alpha$ and $N$ in the context of a Bayesian nonparametric $M^{th}$ model, and compare empirical performance to estimation of $N_\alpha$ and $N$ using parametric choices of $G^*$.

4.2 A Bayesian nonparametric prior for $G^*$

Motivated by the desire to estimate $N_\alpha$ using a flexible nonparametric model for $G^*$ in the $M^{th}$ model, we propose the Dirichlet process prior for $G^*$. Mixture models for $G^*$ have been previously proposed by Pledger [2000]. Ford et al. [2015] suggests a Dirichlet process model in the context of capture-recapture, but does not apply the model to population estimation. Conversely, our focus is on population estimation, and particularly on the properties of the model with respect to estimation of $N_\alpha$ and an analogue of the lower-bound estimate of Chao [1987]. We take a Bayesian approach to inference, and provide a Markov chain Monte Carlo algorithm for computation.

The Dirichlet process is a stochastic process specified by a concentration parameter $\alpha$ and a base measure $H$. It is a random probability measure $\mu(\cdot)$ that can be characterized by its stick-breaking representation

\begin{equation}
\mu(A) = \sum_{h=1}^{\infty} \nu_h \delta_{\theta^*_h}(A), \quad \theta^*_h \sim H, \tag{10}
\end{equation}

\begin{equation}
\nu_h = \nu^*_h \prod_{l<h} (1 - \nu^*_l), \quad \nu^*_1, \nu^*_2, \ldots \sim \text{Beta}(1, \alpha_0), \tag{11}
\end{equation}

for $H$-measurable sets $A$, where $\delta_{\theta^*_h}$ is a Dirac measure at $\theta^*_h$. This model can also be represented hierarchically using a countably supported latent class variable $z$

\begin{equation}
\mu(A \mid z) = \delta_{\theta^*_z}(A), \quad \Pr(z = h) = \nu_h, \tag{12}
\end{equation}

a representation that is commonly employed for computation. When a random probability distribution $G^*$ can be characterized by (10)-(11), we write $G^* \sim DP(\alpha_0, H)$. From (10), it is clear that the Dirichlet process with normal base measure is analogous to the empirical measure estimator, in that both are atomic. If the base measure is instead supported on a function space, then the resulting estimator would be analogous to a histogram or kernel density estimator. Thus, the discussion in Section 3.1 is for the most part relevant to the behavior of the Dirichlet process as a prior for $G^*$ in the $M^{th}$ model.

To complete a Bayesian specification for the likelihood in (8) with Dirichlet process prior on $G^*$, we choose a Gaussian base measure and Gamma prior on $\alpha_0$, giving

\begin{equation}
G^* \sim DP\{\alpha_0, N(0, \sigma^2_G)\}, \quad \beta_t \sim N(0, \sigma^2_\beta), \quad \alpha_0 \sim \text{Gamma}(a, b). \tag{13}
\end{equation}
The induced prior on \( F^* \), the distribution of \( \eta \), is also a random measure. In particular, if \( \varphi \) is the logit function, \( \beta_0 = \beta \) for all \( t \), and \( \mu_{G^*} = \sum_{k=1}^{\infty} \nu_k \delta_{\theta_k} \) is the probability measure corresponding to \( G^* \), then \( \mu_{F^*} = \sum_{h=1}^{\infty} \nu_h \delta_{\eta_h(\theta, \beta)} \) where \( \mu_{F^*} \) is the probability measure corresponding to \( F^* \). We use \( \varphi = \Phi \), the probit function, which facilitates computation. We describe an efficient Markov chain Monte Carlo algorithm based on the truncation approximation of Ishwaran and James [2001] in Supplementary Materials.

Finally, we specify the prior \( p(N) \propto N^{-1} \mathbb{1} \{ N \in \mathbb{N}, N \geq m \} \) for the unknown population size. The resulting conditional distribution of \( N - m \) given \( \theta_{[1:K]}^*, \beta_{[1:T]}^* \), and \( \nu_{[1:K]} \) for a \( K \)-length truncation of the stick-breaking process is

\[
p(N - m \mid \theta_{[1:K]}^*, \beta_{[1:T]}^*, \nu_{[1:K]}) \sim \text{NegBin} \left( m, 1 - \sum_{h=1}^{K} \nu_h \prod_{t=1}^{T} \Phi(-\theta_{h}^* - \beta_t) \right)
\]

It follows that

\[
E(N \mid \theta_{[1:K]}^*, \beta_{[1:T]}^*, \nu_{[1:K]}) = m + \frac{m \{1 - E_F(P)\}}{E_F(P)} = \frac{m}{E_F(P)},
\]

so the posterior mean of \( N \) in this model is equivalent to the point estimates of the population size considered in §3 where the estimator for \( F \) is induced by the Dirichlet process prior on \( G^* \) via the transformation in [6].

### 4.3 Estimation of \( N_\alpha \) and choice of \( \alpha \)

At each iteration the Markov chain Monte Carlo algorithm obtains samples of latent class variables \( z_i \) for every individual, as defined by a truncation of [12]. Via these samples, every individual is assigned a value of \( \theta_i \). In addition, the fully Bayesian approach results in sampling of unobserved members of the population conditional on the sampled value of \( N \). Thus, at every iteration of the sampler, every member of the population, including unobserved members, is assigned a value of \( \theta_i \).

Conditional on samples \( N^{(s)} \) of \( N \), \( \theta_i^{(s)} \) of \( \theta_i \) for \( i = 1, \ldots, N^{(s)} \), and \( \beta_t^{(s)} \) for \( s = 1, \ldots, S \), it is straightforward to obtain samples \( N_\alpha^{(s)} \) of \( N_\alpha \) and an estimate of the posterior expectation \( \widehat{N}_\alpha \) by

\[
N_\alpha^{(s)} = \sum_{i=1}^{N^{(s)}} \mathbb{1} \left\{ \varphi^{-1}(\eta_i^{(s)}) > \alpha \right\}, \quad \widehat{N}_\alpha = \frac{1}{S} \sum_{s=1}^{S} N_\alpha^{(s)},
\]

where \( \eta_i^{(s)} \) is defined by [9]. If a value of \( \alpha \) has already been selected then recording samples of \( N_\alpha^{(s)} \) is all that is necessary to obtain approximate samples from the posterior distribution of \( N_\alpha \) under the model in [13]. However, in some cases the heuristics in [9] may not be precise enough to conclusively select a value of \( \alpha \). Here we suggest an approach to making a data-driven selection of \( \alpha \) a posteriori. As a basis for selection of appropriate \( \alpha \) values, we introduce the following condition.

**Definition 4.1** (Informative classes). Consider a data-augmented likelihood given by \( L(x \mid \theta^*, z) = \prod_{i=1}^{m} L(x_i \mid \theta^*_z) \), where \( z_i = 1, \ldots, K \) is a class assignment variable. A class \( h \) with associated class-specific parameter \( \theta_h^* \) is **informative** if, conditional on the class memberships \( z_i, i = 1, \ldots, N \), the Fisher information of \( \theta_h^* \) is nonzero. Define \( \mathcal{H} \subset \{1, \ldots, K\} \) as the collection of informative classes.
In other words, a latent class is considered informative if the likelihood is not flat in the parameters corresponding specifically to that class. The basic intuition is that if the likelihood is insensitive to the value of some parameters, then those parameters and any functionals of them are not identifiable conditional on the current class assignments. For a Dirichlet process with a base measure $H$ that is a member of the exponential family, it is enough to check whether any observed individuals are members of class $h$. Any classes containing at least one observed individual are informative.

We suggest applying Definition 4.1 at each iteration of the Markov chain Monte Carlo algorithm to classify every latent class as informative or uninformative. In particular, since $N_\alpha$ is a function of $\theta^*_h$ for every latent class $h$ such that $1 - \prod_{t=1}^{T} \{1 - \varphi(\theta^*_h + \beta_t)\} > \alpha$, if any of these latent classes are uninformative then the associated sample for $N_\alpha$ may be unreliable. Therefore, we propose choosing $\alpha = \alpha_{\inf,q}$, the minimal $\alpha$ such that $N_\alpha$ is a function of only informative classes with probability $q \in (0, 1)$. This value can be approximated from the quantiles of the sample path $\alpha_{\inf,s}, s = 1, \ldots, S$ where $\alpha_{\inf,s} = \inf\{\alpha : \theta^{(s)}_h > \alpha \Rightarrow h \in \mathcal{H}\}$. We suggest $q = 0.5$ and $q = 0.95$ as reasonable default values. Finally, one can compute an analogue of the lower bound estimate $\hat{N}_{LB}$ by the ergodic average

$$N^{(s)}_{LB} = \sum_{i=1}^{N^{(s)}} \mathbb{I}(z_i \in \mathcal{H}), \quad \hat{N}_{LB} = \frac{1}{S} \sum_{s=1}^{S} N^{(s)}_{LB} \quad (14)$$

an estimate of the total number of individuals in informative classes. Like the lower bound estimate of [Chao 1987], $N_{LB}$ does not pertain to a concrete population. The empirical properties of these approaches are assessed in Section 5.

### 5 Simulation studies

#### 5.1 Simulation setup

We now conduct a series of simulation studies to evaluate the performance of the Dirichlet process prior on $G^*$ in the $M^{th}$ model in population estimation, estimation of $N_\alpha$, and the various approaches to the choice of $\alpha$. Throughout, our approach is compared to the three parametric choices for $G^*$ that are implemented in the Rcapture package for the R environment. All three are exponential tiltings of exponential family distributions, specifically the Poisson, Gamma, and Gaussian distributions. We refer to these as the indirect Poisson, indirect Gamma, and Darroch specifications for $G^*$, respectively. These are common parametric choices for $G^*$ in the population estimation literature, primarily for computational convenience, and are described extensively in Supplementary Materials.

We simulate from the $M^{th}$ model with three choices of $G^*$: (1) the Darroch distribution with $\beta_t = -3.75$ for all $t$ and $\tau = 2$ (see Table 3 in Supplementary Materials); (2) a Normal distribution with mean 2 and variance 14; and, (3) a mixture of two normal distributions with means 0 and -3.5, variances 0.1 and equal weights on the two components. In each of the simulation scenarios, $N = 2000$, $T = 4$, and a $\beta = \beta$ is set such that $m \in [1000, 1500]$ for most simulated datasets. For each generative model, we generate $R = 200$ replicate sets. Analogous results to those that follow are given for seven additional simulation models in the Supplementary Materials.
5.2 Estimation of $f^*(\eta)$, $N$, and $N_\alpha$

We first present results for estimation of $f^*(\eta)$ on the logit scale in each of these models, obtained by transforming from the estimated random effect distribution $\widehat{G}^*$ to the observability distribution $f^*(\eta)$. Figure 2 shows the estimated density $f^*(\eta)$ obtained using the three parametric choices of $G^*$ and the Dirichlet process model for a single replicate simulation. In general, the Dirichlet process is able to recover the correct shape of $f^*(\eta)$ for $\eta > -5$, which includes $\alpha$-observable individuals for $\alpha > 0.01$. The parametric choices performs poorly in this important region in at least one of the simulations.

![Graph showing estimated density](image)

*Figure 2: Estimated density $f^*$ of $\eta$ in the four models considered, plotted with the true $f^*(\eta)$, for three simulation studies. In the figure legend, Dirichlet process is abbreviated DP.*

We also compare estimates of $N_\alpha$ under the Dirichlet process and three parametric estimation models for pre-selected values of $\alpha$ based on the theoretically motivated heuristic in Section 3. If we want to achieve probability 0.5 of observing data in any 0.01 non-negligible set, then by Theorem 3.3, $\alpha \approx 0.05$. If we instead want to achieve probability 0.95, then choose $\alpha$ between 0.2 and 0.3. We calculate $N_\alpha$ for $\alpha \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$, roughly the range of values suggested by the heuristics.

Table 1 shows root mean square error

$$\Delta(\widehat{N}_\alpha, N_\alpha) = \left\{ R^{-1} \sum_{r=1}^{R} \left( \widehat{N}_{\alpha,r} - N_{\alpha,r} \right)^2 \right\}^{1/2}$$

where $N_{\alpha,r}$ is the true value of $N_\alpha$ for replicate $r$ and $\widehat{N}_{\alpha,r}$ is the associated point estimate. $\Delta$ is computed for each of the four models for $G^*$ and the five $\alpha$ values and $R = 200$ replicates, and results are shown in Table 1. The table aggregates results over the three simulation scenarios that vary the true $G^*$. For the Dirichlet process model, we estimate the posterior expectation of $N$ and $N_\alpha$ by their ergodic averages. For the other models, we use the maximum likelihood estimates of $N$ and $N_\alpha$. As predicted by theorem 3.3, for every model considered, $\Delta(\widehat{N}_\alpha, N_\alpha)$ is considerably smaller for every value of $\alpha$ than $\Delta(\widehat{N}, N)$. The best overall value of $\Delta$ across the four estimation models is that for the Dirichlet process model. The Dirichlet process model has a somewhat larger value of $\Delta(\widehat{N}, N)$ than the Darroch choice for $G^*$. This is actually a desirable property, indicating a large degree of uncertainty about an essentially unknowable quantity, namely $N_\alpha$ for $\alpha = 0$. 

12
Table 1: Root mean squared error across all simulation scenarios for fixed $\alpha$. Overall aggregates results over the fixed nonzero $\alpha$ values, the $\alpha = 0$ column shows results for estimation of $N$.

Table 2 shows coverage of 95 percent intervals for total population estimates for each simulation scenario and model for $G^*$. For the three parametric choices of $G^*$, these are 95 percent confidence intervals, and for the Dirichlet process model these are 95 percent posterior credible intervals. Coverage is computed across the 200 replicates of each of the three simulation scenarios. It is clear that when the parametric models are misspecified, they massively and systematically underestimate the uncertainty in $N$, even in cases where there are no modes in the true $G^*$ centered at large negative values. The empirical coverage of the credible intervals for the Dirichlet process model is higher than 95 percent, however, in Bayesian nonparametric models it is not expected that the coverage of credible intervals coincides with the posterior probability of the interval.

Table 2: Coverage of 95% intervals for total population estimates

We next consider the performance of the Dirichlet process model in quantifying uncertainty in $N_\alpha$ for the pre-selected values of $\alpha$, $\alpha_{inf}$, and $\alpha_{LB}$, the identified population estimate defined in (14). Figure 3 shows an estimate of the posterior distribution of $N$ and of $N_\alpha$ for each considered value of $\alpha$ based on Markov chain Monte Carlo output for a single replicate simulation. Manual examination of results for other simulation replicates was similar. As expected, as $\alpha$ decreases, the posterior distribution of $N_\alpha$ becomes more diffuse. However, there is always considerably less uncertainty in $N_\alpha$ than in $N$ for every value of $\alpha$ considered, despite the fact that coverage remains at or above 95 percent (see Table 6 in the Appendix). In many cases, the posterior for $N$ is extremely diffuse and skewed. Thus, the credible intervals reflect the predictions of the theory: estimates of $N$ are in general highly uncertain and unstable. Not only are point estimates of $N_\alpha$ more reliable, uncertainty in the value of $N_\alpha$ is generally much smaller than uncertainty in the value of $N$. Notably, $\alpha_{inf}$ often falls within the range of $\alpha$ values considered and gives similar results. Similar results for six additional simulation scenarios, as well as coverage of 95 percent credible intervals for $N_\alpha$ for the Dirichlet process model, are provided in Supplementary Materials.
Observable population estimation

Figure 3: Violin plots showing an estimate of the posterior distribution for \( N_\alpha \) for several \( \alpha \) values for the Dirichlet process model; also shown is the distribution of \( N_\alpha \) for \( \alpha = \alpha_{\text{inf},0.5} \). These results correspond to a single simulation replicate. The blue dots correspond to posterior means, and the blue lines to a 95 percent credible interval.

6 Application

Here we illustrate the method using real data from the now classic snowshoe hare dataset, originally analyzed by [Otis et al. 1978] and subsequently re-analyzed by several authors. These data give multiple-recapture history of \( n = 68 \) hares across \( T = 6 \) occasions. Previous analyses, including [Cormack 1989, Agresti 1994, Dorazio and Royle 2003, Pledger 2005], generally agree that considerable capture heterogeneity across individuals is present. Unsurprisingly, given the small sample size, these analysts have also noted that different models accounting for heterogeneity seem to fit the data comparably well, yet result in different estimates of the population size, \( N \). For example [Dorazio and Royle 2003] reports estimates ranging from \( \hat{N} = 76.4 \), with 95% confidence interval \([70.2, 86.4]\), for two-component latent-class \( M^{th} \) models; to \( \hat{N} = 91.9 \), with 95% confidence interval \([74.7, 153.0]\), for logistic-normal \( M^{th} \) models.

Figure 4 presents results applying the Dirichlet process \( M^{th} \) model. We show estimates of the posterior distribution of \( N_\alpha \) for fixed values of \( \alpha \), for the estimated \( \alpha_{\text{inf},0.5} = 0.45 \), as well as for \( N_{LB} \). Consistent with the simulation examples, any choice of \( \alpha \) results in substantially smaller posterior uncertainty in \( N_\alpha \) relative to posterior uncertainty in \( N \), or even \( N_{LB} \). All choices of \( \alpha \) result in similar point estimates, around 80, which are in close agreement with the most conservative previous estimates from heterogeneity models, in the vicinity of 77. This was expected, as the reduced sample size cannot inform much about individuals with reduced capture probabilities, even if they existed. We also note that, as expected, the uncertainty in \( N_\alpha \) is larger for smaller values of \( \alpha \).

Figure 4: Estimates of the posterior distribution of \( N_\alpha \) for several fixed \( \alpha \) values and \( \alpha_{\text{inf},0.5} \) using the snowshoe hares data. Also shown is the estimated posterior distribution of \( N_{LB} \). The blue dots correspond to posterior means, the yellow dots to the true value of \( N_\alpha \), and the blue lines to a 95 percent credible interval.
7 Discussion

Although it has been clear for some time that population estimation in the presence of substantial capture heterogeneity can be unreliable, the precise nature and scope of this problem has remained obscure. We have clarified the ultimate cause of this problem by equating heterogeneous population estimation to a length-biased density estimation problem. In this context, it is clear that only with strong parametric assumptions about the catchability distribution is it possible to estimate the total population size in finite samples.

In the absence of a scientific justification for a particular parametric model or strong prior knowledge, we have proposed estimating the “observable” population size and have suggested several data-driven approaches to defining this population. The approach results in substantially lower risk estimators with considerably lower uncertainty. Our approach can be viewed as an alternative to using lower bounds on the total population size, and has the advantage that the resulting estimate is more interpretable and automatically converges toward an estimate of the total population size as the sample grows large. The method is also easy to implement with any model for the catchability distribution. As such, the approach provides a default method when selection of a parametric form for the catchability distribution is unwarranted, as is the case in many applications of population estimation.

There are at least two strong justifications for estimating the observable population size. The first is to view our proposed approach as an extension of the long-standing assumption in population estimation that there are no individuals in the population with zero capture probability. Researchers have long been comfortable with making scientific inferences using traditional population estimates, acknowledging that these estimates exclude all perfectly invisible individuals. Thus, we expect that scientists will find estimates of $N_α$, the population size excluding “nearly invisible” individuals, similarly useful. Secondly, our results show that estimators of the observable population are biased estimators of $N$ that often have lower risk and are superior in the minimax sense for estimation of the total population size, akin to James-Stein estimators and Bayes estimators. As such, regardless of whether one wishes to assume away individuals that are nearly invisible to the sampling design or not, the approach we suggest has significant advantages over existing methods.

Acknowledgments

This work was inspired by research conducted at Human Rights Data Analysis Group. The authors gratefully acknowledge funding support for this work from the Human Rights Data Analysis Group and the National Institutes of Health. The authors thank David Dunson for helpful comments and suggesting the connection to the length bias literature.

References

A. Agresti. Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics*, 50(2):494–500, 1994.

Masoud Asgharian, Cyr Emile M’Lan, and David B Wolfson. Length-biased sampling with right censoring: an unconditional approach. *Journal of the American Statistical Association*, 97(457):201–209, 2002.
Sophie Baillargeon, Louis-Paul Rivest, et al. Rcapture: loglinear models for capture-recapture in r. *Journal of Statistical Software*, 19(5):1–31, 2007.

Anne Chao. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, pages 783–791, 1987.

Song Xi Chen and Chris J Lloyd. Estimation of population size from biased samples using non-parametric binary regression. *Statistica Sinica*, pages 505–518, 2002.

R. M. Cormack. Log-linear models for capture-recapture. *Biometrics*, 45(2):pp. 395–413, 1989. ISSN 0006341X. URL http://www.jstor.org/stable/2531485.

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

Jacobo de Uña-Álvarez. Nonparametric estimation under length-biased sampling and type i censoring: a moment based approach. *Annals of the Institute of Statistical Mathematics*, 56(4):667–681, 2004.

Robert M Dorazio and Andrew J Royle. Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics*, 59(2):351–364, 2003. URL http://onlinelibrary.wiley.com/doi/10.1111/1541-0420.00042/full.

Jessica H Ford, Toby A Patterson, and Mark V Bravington. Modelling latent individual heterogeneity in mark-recapture data with dirichlet process priors. *arXiv preprint arXiv:1511.07103*, 2015.

Richard D Gill, Yehuda Vardi, and Jon A Wellner. Large sample theory of empirical distributions in biased sampling models. *The Annals of Statistics*, pages 1069–1112, 1988.

Spyridon J Hatjispyros, Theodoros Nicoleris, and Stephen G Walker. Bayesian nonparametric density estimation under length bias. *arXiv preprint arXiv:1510.06307*, 2015.

Richard Huggins. A note on the difficulties associated with the analysis of capture–recapture experiments with heterogeneous capture probabilities. *Statistics & probability letters*, 54(2):147–152, 2001.

Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 2001.

Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 2011.

MC Jones. Kernel density estimation for length biased data. *Biometrika*, 78(3):511–519, 1991.

T Kunihama, AH Herring, CT Halpern, and DB Dunson. Nonparametric bayes modeling with sample survey weights. *arXiv preprint arXiv:1409.5914*, 2014.

William A Link. Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics*, 59(4):1123–1130, 2003. URL http://onlinelibrary.wiley.com/doi/10.1111/j.0006-341X.2003.00129.x/full.
D. Manrique-Vallier. Bayesian population size estimation using dirichlet process mixtures. *Biometrics*, 2016. doi: 10.1111/biom.12502.

Chang Xuan Mao. Computing an npmle for a mixing distribution in two closed heterogeneous population size models. *Biometrical Journal*, 50(6):983–992, 2008.

David L Otis, Kenneth P Burnham, Gary C White, and David R Anderson. Statistical inference from capture data on closed animal populations. *Wildlife monographs*, pages 3–135, 1978.

Ganapati P Patil and Calyampudi R Rao. Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, pages 179–189, 1978.

GP Patil and C Radhakrishan Rao. The weighted distributions: A survey of their applications. *Applications of statistics*, 383:405, 1977.

S. Pledger. The performance of mixture models in heterogeneous closed population capture-recapture. *Biometrics*, 61:868–876, 2005.

Shirley Pledger. Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics*, 56(2):434–442, 2000.

Feng Qi and Bai-Ni Guo. An inequality involving the gamma and digamma functions. *arXiv preprint arXiv:1101.4698*, 2011.

Louis-Paul Rivest and Sophie Baillargeon. Applications and extensions of chao’s moment estimator for the size of a closed population. *Biometrics*, 63(4):999–1006, 2007.

L. Sanathanan. A comparison of some models in visual scanning experiments. *Technometrics*, 15: 67–78, 1973.

C. Chandra Sekar and W. Edwards Deming. On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44(245):101–115, 1949. doi: 10.1080/01621459.1949.10483294. URL [http://www.tandfonline.com/doi/abs/10.1080/01621459.1949.10483294](http://www.tandfonline.com/doi/abs/10.1080/01621459.1949.10483294).

Yehuda Vardi. Nonparametric estimation in the presence of length bias. *The Annals of Statistics*, pages 616–620, 1982.

Yehuda Vardi. Empirical distributions in selection bias models. *The Annals of Statistics*, pages 178–203, 1985.

Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

### Appendix

#### A.1 Proof of Remark 2.1

First consider the case where $p$ is the same for all individuals. Then $m \sim \text{Binomial}(N, p)$, and the log-likelihood is given by

\[
l(m \mid N, p) \propto \log(N!) - \log((N - m)!) + (N - m) \log(1 - p).
\]
Introducing the penalty factor of $-\frac{1}{2} \log \{N(N - m)^{-1}\}$,

$$\hat{l}(m \mid N, p) \propto \log \{\Gamma(N + 1)\} - \log \{\Gamma(N - m + 1)\} + (N - m) \log(1 - p) - \frac{1}{2} \log \{N(N - m)^{-1}\}.$$ 

This factor is analogous to a continuity correction for the parameter $N$, since, applying Stirling’s series

$$\hat{l}(m \mid N, p) = (N + 1/2) \log N - N - (N - m + 1/2) \log(N - m) + (N - m)$$

$$+ O(N^{-1}) + O\{(N - m)^{-1}\} - \frac{1}{2} \log\{N(N - m)^{-1}\}$$

$$= N \log N - N - (N - m) \log(N - m) + (N - m) + O(N^{-1}) + O\{(N - m)^{-1}\},$$

so the penalty term gives an asymptotically exact expression for the log-likelihood of Binomial($N - 1/2, p$). Treating $N$ as a continuous parameter, differentiating and setting to zero

$$\frac{\partial}{\partial N}\hat{l}(m \mid N, p) = \psi(N + 1) - \psi(N + 1 - m) + \frac{m}{2N(N - m)} + \log(1 - p) = 0$$

which implies

$$\exp[\psi(N + 1) - \psi(N + 1 - m) + m\{2N(N - m)^{-1}\}] = \frac{1}{1 - p},$$

where $\psi$ is the digamma function. From [Qi and Guo 2011], for $x \in (0, \infty)$,

$$\log(x) + \frac{1}{2x} - \frac{1}{12x^2} < \psi(x + 1) < \log(x) + \frac{1}{2x},$$

so $\psi(x + 1) = \log(x) + (2x)^{-1} - h(x)$ for a function $(12x^2)^{-1} < h(x) < 0$. Substituting

$$\frac{N}{N - m} \exp\{(2N)^{-1} - \{2(N - m)\}^{-1} + m\{2N(N - m)^{-1}\} - \{h(N) - h(N - m)\}\} = \frac{1}{1 - p}$$

$$\frac{N}{N - m} \exp\{-m\{2N(N - m)^{-1}\}^{-1} + m\{2N(N - m)^{-1}\} - \{h(N) - h(N - m)\}\} = \frac{1}{1 - p}$$

$$\frac{N}{N - m} (1 - \{h(N) - h(N - m)\} + O\{(h(N) - h(N - m))^2\}) = \frac{1}{1 - p}$$

So

$$\frac{N}{1 - O(N^{-2})} = \frac{N - m}{1 - p}$$

$$N\left(1 - \frac{1}{1 - p}\right) - O(N^{-1}) = \frac{-m}{1 - p}$$

$$Np + O(N^{-1}) = m$$

$$\tilde{N} = \frac{m}{p} - O\left(m^{-1}\right).$$
This immediately gives a result for \( p \sim F \) where \( F \) consists of \( J < \infty \) atoms. Suppose \( \Pr(P = p_j) = w_j \), and let \( m_1, \ldots, m_J \) be the counts of observed units with \( p = p_j \). Then

\[
(N - m, m_1, \ldots, m_J) \sim \text{Multinomial}\left( N, \left(1 - \sum_j w_j p_j, w_1 p_1, \ldots, w_J p_J\right)\right),
\]

so

\[
m \sim \text{Binomial}\left( N, \sum_j w_j p_j\right),
\]

giving

\[
\hat{N} = \frac{m}{E_F(P)} - O(m^{-1}).
\]

### A.2 Proof of Theorem 3.3

Let \( Y \) be the random variable defined by

\[
Y = 1\{p \in A\} 1\left\{ \sum_{i=1}^T X_i > 0 \right\},
\]

where \( X_i \) is the random variable of which \( x_{it} \) is an observation. Clearly, \( Y \) is binary, and the event \( \{Y = 1\} \) occurs exactly when \( p \in A \) and \( S = \sum_{i=1}^T X_i > 0 \). Since \( \Pr(S > 0) = p \) and \( A \) is \( \alpha \)-observable, \( \Pr(S > 0 \mid p \in A) \geq \alpha \). Since \( A \) is \( f, \epsilon \) non-negligible, it follows that \( \Pr(Y > 0) \geq \epsilon \alpha \).

Therefore, \( M(A) \) is the sum of independent binary random variables \( Y_1, \ldots, Y_N \) each having \( \Pr(Y_i = 1) \geq \epsilon \alpha \). So, \( M(A) \) stochastically dominates a Binom \((N, \epsilon \alpha)\) random variable. So then

\[
\Pr(M(A) > 0) \geq 1 - (1 - \epsilon \alpha)^N
\]

### A.3 Proof of Theorem 3.4

We have \( \hat{N} = \sum_{i=1}^m \hat{p}_i^{-1} \). We will need

\[
E_G(1/P) = \int p^{-1}G(dp) = \int p^{-1}pF(dp)\{E_F(P)\}^{-1} = \{E_F(P)\}^{-1}
\]

\[
E_G(P^{-2}) = \int p^{-2}G(dp) = \int p^{-2}pF(dp)\{E_F(P)\}^{-1} = E_F(P^{-1})\{E_F(P)\}^{-1}
\]

so

\[
\text{var}_G(P^{-1}) = E_F(P^{-1})\{E_F(P)\}^{-1} - \{E_F(P)\}^{-2}.
\]

First compute the variance

\[
\text{var} \left( \hat{N} \right) = \text{var} \left\{ E(\hat{N} \mid m) \right\} + E \left\{ \text{var} \left( \hat{N} \mid m \right) \right\}
\]
\[ \text{where the third step used the variance of } m \text{ derived in the proof of Theorem 3.7. Since} \]

\[ E(\hat{N}) = E\{mE_G(P^{-1})\} = NE_F(P)\{E_F(P)\}^{-1}, \]

\[ \text{the estimator is unbiased, so} \]

\[ \Delta^2(\hat{N}, N) = N\{E_F(P^{-1}) - 1\}. \]

### A.4 Proof of Theorem 3.7

For this result we have the assumptions that \( F \) has density \( f \) that is twice continuously differentiable, which implies that \( G \) also has a density. Consider a histogram estimator of a density \( g \) on the unit interval with bin width \( h \), which is chosen such that \( 1/h \) is an integer. Using (5), the estimator of \( N \) conditional on \( m \) becomes

\[
\hat{N} = m \left\{ \int p \hat{f}(p) dp \right\}^{-1} = m \left\{ \int p \frac{p^{-1} \hat{g}(p)}{\int_0^1 p^{-1} \hat{g}(p) dp} dp \right\}^{-1}
\]

\[
= m \left\{ \frac{\int \hat{g}(p) dp}{\int_0^1 y^{-1} \hat{g}(p) dp} \right\}^{-1} = m \left\{ \frac{1}{\int_0^1 p^{-1} \hat{g}(p) dp} \right\}^{-1} = m \int_0^1 p^{-1} \hat{g}(p) dp.
\]

Let \( B_j \) be the count in the \( j \)th bin. For \( \hat{g} \) the histogram estimator we have

\[
\text{var} \left( \hat{N} \mid m \right) = m^2 \text{var} \left[ \int p^{-1} \hat{g}(p) dp \right]
\]

\[
= m^2 \text{var} \left[ \sum_{j=0}^{1/h-1} \int_{hj}^{h(j+1)} y^{-1} \hat{g}(p) dp \right] = m \text{var} \left[ \sum_{j=0}^{1/h-1} hp_j^{-1} \hat{g}(p_j) \right] \quad (15)
\]

\[
= m^2 \text{var} \left[ \sum_{j=0}^{1/h-1} hp_j^{-1} m^{-1} h^{-1} B_j \right] = m \text{var} \left[ \sum_{j=0}^{1/h-1} p_j^{-1} B_j \right]
\]

\[
= \sum_{j=0}^{1/h-1} p_j^{-2} \text{var} (B_j) + 2 \sum_{j=0}^{1/h-1} \sum_{k=j+1}^{1/h-1} p_j^{-1} p_k^{-1} \text{cov}(B_j, B_k).
\]

where \( p_j \in [hj, hj + h] \) and we applied the mean value theorem in (15). The \( B_j \) are jointly Multinomial\([m, \{G(h) - G(0), G(2h) - G(h), \ldots, G(1) - G(1 - 1/h)\}]\), where \( G(t) = \int_0^t g(p) dp \), so

\[
\text{var} \left( \hat{N} \mid m \right) = \sum_{j=0}^{1/h-1} p_j^{-2} m\{G(hj + h) - G(hj)\}\{1 - G(hj + h) + G(hj)\}
\]
\[ -2 \sum_{j=0}^{1/h-1} \sum_{k=j+1}^{1/h-1} p_j^{-1} p_k^{-1} m \{G(hj + h) - G(hj)} \{G(hk + h) - G(hk)] \]

\[ = \sum_{j=0}^{1/h-1} p_j^{-2} m h g(p_j^*) \{1 - h g(p_j^*)\} - 2 \sum_{j=0}^{1/h-1} \sum_{k=j+1}^{1/h-1} p_j^{-1} p_k^{-1} m h^2 g(p_j^*) g(p_k^*) \]

\[ = m \int_0^1 p^{-2} \{g(p) - h g^2(p)\} dp - 2m \int_0^1 \int_p^1 p_1^{-1} p_2^{-1} g(p_1) g(p_2) dp_1 dp_2 + O(mh), \]

where again we applied the mean value theorem, and \( p_j^* \in [hj, hj + h]. \) Using (4)

\[
\text{var} \left( \hat{N} \mid m \right) = m \int_0^1 p^{-2} \left[ \frac{pf(p)}{\int_0^1 p f(p) dp} - h \left\{ \frac{p^2 f^2(p)}{\int_0^1 p f(p) dp} \right\}^2 \right] dp
\]

\[ - 2m \int_0^1 \int_{p_2=p_1}^1 p_1^{-1} p_2^{-1} \frac{p_1 f(p_1)}{\int_0^1 p f(p) dp} \frac{p_2 f(p_2)}{\int_0^1 p f(p) dp} dp_1 dp_2 + O(mh) \]

\[ = m \left[ \frac{E_f(P^{-1})}{E_f(P)} - h \left\{ \frac{f_f(P)}{E_f(P)} \right\}^2 - \frac{1}{\{E_f(P)\}^2} \right] + O(mh) \tag{16} \]

In (16), we used the fact that

\[ \int_{p_1=0}^1 \int_{p_2=p_1}^1 f(p_1) f(p_2) dp_1 dp_2 \]

is the integral of the product density \( f(p_1, p_2) = f(p_1) f(p_2) \) over the region of the unit square above the line \( p_1 = p_2. \) Since \( f(p_1, p_2) \) is symmetric about the line \( p_1 = p_2, \) we have

\[ \int_{p_1=0}^1 \int_{p_2=p_1}^1 f(p_1) f(p_2) dp_1 dp_2 = \frac{1}{2} \int_{p_1=0}^1 \int_{p_2=0}^1 f(p_1) f(p_2) dp_1 dp_2 = \frac{1}{2}. \]

Since \( E(m \mid N, f) = N \{E_f(P)\}, \) we have

\[ \text{var} \left( \hat{N} \mid N, f \right) = N \left[ E_f(P^{-1}) - \{E_f(P)\}^{-1} \right] + O(Nh). \]

Now we compute the absolute bias

\[
E(\hat{N} \mid m, \hat{g}) - E(N \mid m, g) = mE \left[ \int_0^1 p^{-1} \{\hat{g}(p) - g(p)\} dp \right]
\]

\[ = m \int_0^1 p^{-1} E[\hat{g}(p) - g(p)] dp = m \int_0^1 p^{-1} \text{Bias}[\hat{g}(p) \mid m] dp. \]

The bias of \( \hat{g}(p) \) is \( g'(p)(hj + h/2 - p) + O(h^2), \) see Wasserman 2006, so we have

\[ |E(\hat{N} \mid m, \hat{g}) - E(N \mid m, g)| = m \sum_{j=0}^{1/h-1} \int_0^{hj+h} p^{-1} \left[ g'(p)(hj + h/2 - p) + O(h^2) \right] dp \]
Observable population estimation

\[ = m \left| \sum_{j=0}^{1/h-1} hp_j^{-1} \left[ g'(p_j)(hj + h/2 - p_j) + O(h^2) \right] \right| \]

\[ \leq m \sum_{j=0}^{1/h-1} hp_j^{-1} \left[ |g'(p_j)||hj + h/2 - p_j| + O(h^2) \right] \]

\[ \leq \frac{mh}{2} \left( \int_0^1 p^{-1}|g'(p)|dp + O(h) \right) + O(mh^2) \]

\[ \leq \frac{mh}{2} \int_0^1 pf(p)dp \int p^{-1}|pf'(p)|dp + O(mh^2) \]

where we applied the mean value theorem in the second step. So

\[ \frac{1}{2} \left\{ \text{Bias}(\hat{N} | m) \right\}^2 \leq \left( \frac{mh}{2E_f(P)} \left[ E_f(P^{-1}) + E_f(|f'(P)|) \right] \right)^2 + O(mh^3). \]

Now we just need to calculate \( E(m^2 | N, f) \). Let \( x_i \sim \text{Bernoulli}(p_i) \) be the indicator that the \( i \)th unit is observed, then

\[ \text{var}(m | N, f) = \text{var} \left\{ E(m | p_1, \ldots, p_N) \right\} + E \{ \text{var}(m | p_1, \ldots, p_N) \} \]

\[ = \text{var} \left\{ \sum_{i=1}^N p_i \right\} + E \left\{ \text{var} \left( \sum_{i=1}^N x_i \right) \right\} \]

\[ = \sum_{i=1}^N \text{var}(p_i) + E \left\{ \sum_{i=1}^N p_i(1 - p_i) \right\} \]

\[ = \sum_{i=1}^N E(P_i^2) - \{E(P_i)\}^2 + \sum_{i=1}^N E\{P_i(1 - P_i)\} = N[E_f(P) - \{E_f(P)\}^2], \]

so

\[ E(m^2 | N, f) = N[E_f(P) - \{E_f(P)\}^2] + N^2\{E_f(P)\}^2 \]

\[ = NE_f(P) + (N^2 - N)\{E_f(P)\}^2. \]

So finally

\[ O(Nh) \leq \Delta^2(\hat{N}, N) - N \left[ E_f(P^{-1}) - \{E_f(P)\}^{-1} \right] \leq \frac{N^2h^2}{4} \left\{ E_f(P^{-1}) + E_f(|f'(P)|) \right\}^2 + O(Nh), \]

which gives the result.

B Supplementary Material

B.1 Table of simulation scenarios

Table 3 shows the choices of \( G^* \) for each of the ten simulation studies that contribute to the root mean square error results in Table 1 in the main text; the first three are those discussed in the main
| Scenario | Model                  | \( G^*(d\theta) \propto \)                                      |
|-----------|------------------------|---------------------------------------------------------------|
| 1         | Darroch                | \( \sum_{t=1}^{T} w_t \text{No}(\tau^2 t, \tau^2) \)       |
| 2         | Normal                 | \( \text{No}(0, \tau^2) \)                                   |
| 3         | two normal mixture     | \( \sum_{h=1}^{2} w_h \text{No}(\mu_h, \tau^2_h) \)       |
| 4         | truncated normal       | \( \text{No}(0, \tau^2) \mathbb{I}\{\theta \in [a, b]\} \) |
| 5         | indirect gamma         | \( \sum_{t=1}^{T} w_t \text{Gam}\{\tau, (\lambda + t)^{-1}\} \) |
| 6         | atoms                  | \( \sum_{h=1}^{K} w_h \delta_{\theta_h}(d\theta) \)         |
| 7         | multi-normal mixture   | \( \sum_{h=1}^{K} w_h \text{No}(\mu_h, \tau^2_h) \)       |
| 8         | multi-t mixture        | \( \sum_{h=1}^{K} w_h t_{\nu_h}(\mu_h, \tau^2_h) \)       |
| 9         | Normal, small variance | \( \text{No}(0, \tau^2) \)                                   |
| 10        | Normal, \( T = 7 \)   | \( \text{No}(0, \tau^2) \)                                   |

Table 3: Data generating models for simulations. The weights \( w_t \) for the Darroch and indirect Gamma models are given in Rivest and Baillargeon [2007].

B.2 Parametric choices for \( G^* \) in the \( M^{th} \) model

In this section, we review common parametric choices for \( G^* \) in the \( M^{th} \) model. Our purpose is (1) to provide background that allows comparison of the method we propose in §4.2 to common alternatives; (2) to show that while making a parametric choice for \( G^* \) ensures that \( N \) is identifiable, when the model is misspecified, the results can be highly misleading, and (3) to show that while the most common choices for \( G^* \), which are exponentially tilted exponential family distributions, are mathematically convenient, they are both highly inflexible and imply very strong assumptions about \( G^* \) in \( \alpha \)-unobservable regions.

Certain parametric choices for \( G^* \) lead to convenient analytical expressions for the cell probabilities \( \pi_x \) in \( M^{th} \) models. For this reason, much of the literature has focused on distributions \( G^* \) indirectly defined so that \( G^*_0(\theta) \sim G^*(\theta \mid x = \zeta) \), for \( G^*_0(\theta) \) in a common parametric family such as the normal. Doing so ensures that, after integrating out the random effects, the likelihood belongs to the exponential family. Moreover, the expected cell counts marginal of the individual-level random effects can be expressed in closed form as

\[
E_\theta\{n(x)\} = \exp \left\{ \beta_0 + \sum_i x_i \beta_i + \phi_{G^*_0} \left( \sum_i x_i \right) \right\},
\]

where \( \phi_{G^*_0}(j) \) is the cumulant generating function of \( G^*_0(\theta) \), i.e. \( \phi_{G^*_0}(j) = \log \int_{-\infty}^{\infty} e^{ij\theta} G^*_0(\theta) \). Thus, if the cumulant generating function of \( G^*_0(\theta) \) can be easily calculated, the expected cell counts and
 Observable population estimation

| Scenario | Parameters | $\beta$ |
|----------|------------|---------|
| 1        | $\tau^2 = 2$ | -3.75  |
| 2        | $\tau^2 = 14$ | 2      |
| 3        | $w_h = 0.5$, $\tau_h^2 = 0.1$, $\mu_{1:2} = (0, -3.5)$ | 1      |
| 4        | $\tau^2 = 12$, $a = -2$, $b = \infty$ | -1     |
| 5        | $\tau = 1$, $\lambda = 3.5$ | -2     |
| 6        | $w \sim \text{Dirichlet}(1,\ldots,1)$, $\theta^* = (-5, -3.7, -3.2, -2.75, 0, 1, 3)$ | 0      |
| 7        | $w \sim \text{Dirichlet}(1,\ldots,1)$, $\mu = (-4.2, -2.5, -2.25, 0, 1, 3)$, $\tau^2 = (0.1, 0.05, 0.1, 1, 0.1, 6)$ | 1      |
| 8        | $\nu = 3$, $w, \mu, \tau^2$ same as Scenario 7 | 1      |
| 9        | $\tau^2 = 0.1$ | 0      |
| 10       | $\tau^2 = 10$ | -1     |

Table 4: Parameters for data generating models used in simulations.

Table 5: Common distributions used for $G^*_0(\theta)$.

| Distribution Name     | Definition                |
|-----------------------|---------------------------|
| Darroch               | $g^*_0(\theta) = N(0, \tau)$ |
| indirect Poisson      | $\theta = \log(2)X$, $X \sim \text{Poisson}(\tau)$ |
| indirect Gamma        | $\theta = -X$, $X \sim \text{Gamma}(3.5, \tau)$ |

thus the cell probabilities marginal of $\theta$ are also easily computed. Table 5 shows three distributions commonly used in capture-recapture.

Mixing distributions defined in this way can be problematic. As Baillargeon et al. [2007] show, whenever $G^*_0(\theta)$ is in the exponential family, the resulting $G^*(\theta)$ will be a mixture of exponentially tilted exponential family distributions. Therefore, in general $G^*(\theta)$ will be a multi-modal distribution governed entirely by the parameters of $G^*_0(\theta)$. Moreover, if $G^*_0(\theta) = G^*_0(\theta; \tau)$ has only one parameter, as is typically the case, $\tau$ will define the shape and location of all the modes. For example, if $G^*_0 \sim N(0, \tau)$ – the Darroch distribution in Table 5, then $G^*(\theta)$ will be a mixture of $T + 1$ Gaussian distributions, with means $t\tau$ and variance $\tau$ for $t = 0, 1, \ldots, T$. Thus, the variance $\tau$ in $G^*_0(\theta)$ defines not only the variance of each of the mixture components but also their locations.

In practice, the consequence of choosing these inflexible parametric distributions is that in order to fit the shape of $G^*$ in the observed region, a mode may be introduced in the unobserved region. This can result in highly misleading results, as there is essentially no data to inform about the shape of the distribution in the unobserved region.

To illustrate this point, we simulate data from the model in (8) using different choices for $G^*$, then estimate $G^*$ assuming that it is a member of each of the parametric classes in Table 5. We choose three different true distributions for $G^*$: Darroch, Gaussian ($\theta \sim N(0, \sigma^2)$), and a mixture
of two normals. In the first case, the model is exactly Darroch and the other two parametric choices for \( G^\star \) are misspecified. In the other two data generation cases, all three of the models used for estimation are misspecified.

Figure 5 shows results. In each panel, the true density \( g^\star \) is shown in black, and the estimated density \( \hat{g}^\star \) under each of the parametric models is displayed on the same plot; we refer to these estimates as \( \hat{g}^\star_D \) for the Darroch estimate, \( \hat{g}^\star_G \) for the indirect Gamma estimate, and \( \hat{g}^\star_P \) for the indirect Poisson estimate. These estimates were obtained as transformations of the output of the Rcapture package \( M^\text{th} \) model fit to simulated data from a population of size \( N = 2000 \). As expected, in the first simulation, \( \hat{g}^\star_D \) recovers \( g^\star \) almost exactly, while \( \hat{g}^\star_G \) and \( \hat{g}^\star_P \) perform poorly; \( \hat{g}^\star_G \), in particular, contains a large mode at approximately \( \theta = -7 \) that is not present in the data generating model. A unit with \( \theta = -7 \) has a 0.36% chance of being observed on any of the four lists, making these estimated individuals 0.01-unobservable. Even if all 2000 individuals in the population had \( \theta \leq -7 \), we would expect to observe only seven of them. Similar behavior of \( \hat{g}^\star_G \) is evident when the data generating model is Normal. In this case, \( \hat{g}^\star_D \) also introduces modes where none exist in the true \( g^\star \), though they do not reside as far in the left tail as those in \( \hat{g}^\star_G \). In both the first and second simulations, \( \hat{g}^\star_P \) is improperly truncated on the left. Interestingly, in the third simulation, there is a substantial mode around \( \theta = -3 \) that is completely missed by all three estimators.

![Figure 5: Estimates of the random effects distribution under three estimation models.](image)

Although the three parametric distributions considered are generally ill-suited to characterizing \( g^\star \) in these simulations, the strategy of estimates of \( N_\alpha \) can nonetheless be accurate in many cases. Figure 6 compares the expectation of the \( \alpha \)-observable population size in the true data generating model to the point estimates of \( N_\alpha \) using \( \hat{g}^\star_D \), \( \hat{g}^\star_G \), and \( \hat{g}^\star_P \). In the first two simulations, estimates of \( N_\alpha \) using either \( \hat{g}^\star_D \) or \( \hat{g}^\star_G \) are fairly accurate for values of \( \alpha \) as small as 0.1. The third simulation is the most challenging, with none of the estimators producing reliable estimates over a consistent range of \( \alpha \) values. This reflects that the shape of the true \( g^\star \) even in the observable region is simply not represented by any members of the three parametric classes; use of a more flexible model would likely ameliorate the problem considerably.

### B.3 MCMC algorithm for the nonparametric \( M^\text{th} \) model

We use a truncated stick-breaking prior approximation of the Dirichlet process, see Ishwaran and James [2011]. Let \( K \) be the stick-breaking truncation level. The complete data model can be
represented as

\begin{align*}
x_{it} &= I(\gamma_{ij} > 0) \text{ for } i = 1, ..., N \text{ and } t = 1, ..., T \\
\gamma_{it} &\sim N(\theta_{zi} + \beta_t, 1) \\
\beta_t &\iid N(\phi, \sigma^2) \text{ for } t = 1, ..., T \\
z_i &\iid \text{Discrete}\{\{1, ..., K\}, (\pi_1, ..., \pi_K)\} \text{ for } i = 1, ..., N \\
\theta_h &\iid N(\mu, \tau^2) \text{ for } h = 1, ..., K \\
\nu_h &= \nu_h^* \prod_{l<h} (1 - \nu_l^*) \text{ for } h = 1, ..., K \\
\nu_h^* &\iid \text{Beta}(1, \alpha_0), \text{ for } h < K \text{ and } \nu_K^* = 1 \\
\alpha_0 &\sim \text{Gamma}(a, b), \quad p(N) \propto 1/N
\end{align*}

Let \(n\) be the number of vectors such that \(x_i \neq (0, ..., 0)\). We re-label all the variables that depend on the index \(i = 1, ..., N\) so that \(x = (0, ..., 0)\) for \(i > m\). Let \(\omega = (\omega_1, ..., \omega_K)\) where \(\omega_h = \sum_{i=m+1}^N I(z_i = h)\) is the count of non-observed units in latent class \(h\). Similarly, let \(m_h = \sum_{i=1}^m I(z_i = h)\) be the number of observed units in latent class \(h\). Finally define \(\Gamma_{th}^0 = \sum_{i:i>m, z_i=h} \gamma_{it}\).

1. **Sample from \((z_i)_{i \leq m}\):** For \(i = 1, ..., m\), sample \(z_i \sim \text{Discrete}\{\{1, ..., K\}, (q_1, ..., q_K)\}\), with \(q_i \propto \nu_h \exp \left(-T(\theta_h^*)^2/2 - \theta_h^* \left(\sum_{t=1}^T \beta_t - \sum_{t=1}^T \gamma_{it}\right)\right)\).

2. **Sample from \((\gamma_{it})_{i \leq m}\):** For \(i = 1, ..., m\)

\[
\gamma_{it} \sim \begin{cases} 
\theta_{zi}^* + \beta_t + \left\{ \begin{array}{ll}
\text{TNorm}(0, 1, [-\theta_{zi}^* - \beta_t, +\infty]) & \text{if } x_{it} = 1 \\
\text{TNorm}(0, 1, (-\infty, -\theta_{zi}^* - \beta_t]) & \text{if } x_{it} = 0
\end{array} \right.
\end{cases}
\]

3. **Joint sampling \((\beta, \theta^*)\):** Make \(\nu_h = m_h + \omega_h\) for \(h = 1, ..., K\), and sample \((\beta, \theta^*)' \sim N(\mu, \Sigma)\)
for

\[ \Sigma = \begin{pmatrix} \frac{1}{\sigma^2} + N & I_T \\ \nu_1 & \nu_1 \\ \nu_2 & \nu_2 \\ \vdots & \vdots \\ \nu_K & \nu_K \end{pmatrix} \begin{pmatrix} \nu_1 \\ \nu_2 \\ \vdots \\ \nu_K \end{pmatrix} - 1 \]

and

\[ \mu = \begin{pmatrix} \frac{\phi}{\sigma^2} + \sum_{i=1}^{m_i} \gamma_{it} + \sum_{h=1}^{K} \Gamma_0^h \\ \frac{\phi}{\sigma^2} + \sum_{i=1}^{m_i} \gamma_{it} + \sum_{h=1}^{K} \Gamma_0^h \\ \vdots \\ \frac{1}{\tau^2} + \sum_{i,i \leq m,z_i=1} \gamma_{it} + \sum_{t=1}^{T} \Gamma_t^h \\ \frac{1}{\tau^2} + \sum_{i,i \leq m,z_i=2} \gamma_{it} + \sum_{t=1}^{T} \Gamma_t^h \\ \vdots \\ \frac{1}{\tau^2} + \sum_{i,i \leq m,z_i=K} \gamma_{it} + \sum_{t=1}^{T} \Gamma_t^h \end{pmatrix} \]

4. Sample from \( \nu \): For \( h = 1, \ldots, K - 1 \) sample

\[ \nu_h^* \sim \text{Beta} \left( 1 + \nu_h, \alpha_0 + \sum_{l=h+1}^{K} \nu_l \right) \]

where \( \nu_h = m_h + \omega_h \). Let \( \nu_K^* = 1 \) and make \( \nu_h = \nu_h^* \prod_{l<h} (1 - \nu_l^*) \) for all \( h = 1, \ldots, K \).

5. Sample from \( \alpha_0 \sim \text{Gamma}(a - 1 + K, b - \log \nu_K) \)

6. Sample from \( (N, (z_i)_{i \geq m}, (\gamma_{it})_{i \geq m}) \): For this we use a conditional sampling strategy, similar to [Manrique-Vallier 2016]. We factor this joint distribution into:

\[ p(N, (z_i)_{i \geq m}, (\gamma_{it})_{i \geq m} | \theta^*, \beta, \nu) = p(N | \theta^*, \beta, \nu^*) p((z_i)_{i \geq m} | \theta^*, \beta, N, \nu) p((\gamma_{it})_{i \geq m} | \theta^*, \beta, N, (z_i)_{i \geq m}, \nu) \]

and sample from the three partial conditional distributions sequentially. Since we only need \( N, \omega, \Gamma_{jk}^0 \), we sample these variables instead.

(a) Sample from \( N | \theta, \beta, \pi \): Let \( N_0 = N - n \). The partial conditional distribution of \( N_0 \) for prior \( p(N) \propto 1/N \)

\[ N_0 \sim \text{NegBinomial} \left( m, 1 - \sum_{h=1}^{K} \rho_h \right) \]

where \( \rho_h = \nu_h \prod_{l=1}^{T} \Phi(-\theta_h^* - \beta_l) \) and \( \Phi(\cdot) \) is the distribution function of a standard normal distribution. Set \( N = N_0 + m \).

(b) Sample from \( \omega \sim \text{Multinomial}(N_0, (q_1, \ldots, q_K)) \) where \( q_k \propto \rho_k \).

(c) Sample \( \Gamma_{jk}^0 \): For \( i = 1, \ldots, \omega_h \) get \( \gamma_{ith}^0 \sim \text{TNorm}(\theta_h^* + \beta_l(1, [0, \infty)) \) and make \( \Gamma_{jk}^0 = \sum_{i=1}^{\omega_k} \gamma_{ith}^0 \)
B.4 Additional information on the $M^{th}$ model

If we choose $\varphi = \ell$, the likelihood simplifies to $\prod_{t=1}^{T} e^{(\theta_i + \beta_t) x_{it}}$, and under the additional restriction $\beta_t = \beta$ for $t = 1, \ldots, T$, we can solve for $\eta$ in terms of $\theta$ analytically, giving

$$\eta = \log \left[ (1 + e^{\theta_i + \beta})^T - 1 \right] = \eta(\theta, \beta)$$

This function is monotone increasing in $\theta$, which is consistent with the basic intuition that higher “catchability” should be equivalent to smaller no-capture probabilities. In this case we can also give the condition

$$\theta_i > \log \left[ \left( \frac{1}{1 - \alpha} \right)^{1/T} - 1 \right] - \beta$$

for individual $i$ to be $\alpha$-observable. If we further assume that $G^*$ has density $g^*$, we can give the induced distribution exactly

$$f^*(\eta) = \frac{1}{T} \left( 1 + e^{\eta} \right)^{(1-T)/T} e^\eta \left( \log \left[ (1 + e^{\eta})^{1/T} - 1 \right] - \beta \right). \quad (18)$$

When $|\eta|$ is large, the Jacobian factor in (18) behaves like a constant, and the term $\log[(1 + e^{\eta})^{1/T} - 1] \approx \eta$. So the tails of $f^*$ behave much like the tails of $g^*$.

B.5 Additional Simulation Studies

In this section, we provide results for six additional simulation studies – scenarios 4-9 in Table 3 in the Appendix. The simulations are constructed as described in §5.1 and the choice of $G^*$ for data generation is varied across each simulation. The six additional cases considered here are mostly mixture models, though we also consider an atomic case and one in which the data generating model is the indirect Gamma. Figure 7 shows the distribution of observed population size $m$ across the 200 replicates of all nine simulations. In most cases, $m \in [1000, 2000]$. Recall that in every simulation, the total population size $N = 2000$.

Shown in Fig. 8 are estimates $\hat{f}^*$ of the density $f^*(\eta)$ on the logit scale induced by the estimated distribution of the random effects $\hat{G}^*$. The Dirichlet process model is generally successful at identifying the locations of significant modes, while the parametric models often miss one or more of the modes.

Figure 9 shows summaries of the estimated distribution of $N_\alpha$ for a single simulation replicate for a range of values of $\alpha$, as well as $\alpha_{inf,0.5}$. The results correspond to simulation scenarios 4 through 9 from Table 3 in the Appendix. The results are similar to those shown in the main text for simulation scenarios 1 through 3. As the value of $\alpha$ decreases, posterior uncertainty about $\alpha$ increases. However, any nonzero value of $\alpha$ allows for much smaller credible regions than those for estimates of $N$. This is expected and appropriate, since the value of $N$ cannot be accurately estimated in finite samples. Additionally, point estimates of $N_\alpha$ are relatively accurate and fall within the credible regions.
Figure 7: Distribution of values of $m$ in the simulation studies

Figure 8: Estimated density $f^*$ of observability $\eta$ on the logit scale in the four models considered, plotted with the true $f^*(\eta)$, for six of the ten simulation studies
Observable population estimation

Figure 9: Estimates of the posterior distribution for $N_\alpha$ for several $\alpha$ values for the Dirichlet process model; also shown are results for $\alpha = \alpha_{\text{inf}}$. The blue dots correspond to posterior means, the yellow dots to the true value of $N_\alpha$, and the blue lines to a 95 percent credible interval.

|                | 0.25 | 0.2 | 0.15 | 0.1  | 0.05 | 0   |
|----------------|------|-----|------|------|------|-----|
| darroch       | 1.00 | 1.00| 1.00 | 1.00 | 1.00 | 1.00|
| normal        | 1.00 | 1.00| 1.00 | 1.00 | 1.00 | 1.00|
| 2 normal mixture | 1.00 | 0.98| 1.00 | 1.00 | 1.00 | 0.99|
| truncated normal | 1.00 | 1.00| 1.00 | 1.00 | 1.00 | 0.99|
| gamma         | 1.00 | 1.00| 1.00 | 1.00 | 1.00 | 1.00|
| atoms         | 0.54 | 0.91| 0.98 | 0.99 | 0.99 | 1.00|
| multi-normal mixture | 1.00 | 1.00| 1.00 | 1.00 | 1.00 | 0.97|
| multi-t-mixtue | 1.00 | 1.00| 1.00 | 1.00 | 1.00 | 0.99|
| normal, small variance | 0.92 | 0.93| 0.94 | 0.95 | 0.97 | 0.96|
| eco example   | 1.00 | 0.99| 0.99 | 0.99 | 1.00 | 1.00|

Table 6: Coverage of 95% intervals for $N_\alpha$ for several values of $\alpha$. 