1 Introduction

The goal of this study is to provide a model to estimate abundance and detectability for many bird species observed in the Minnesota Forest Breeding Bird Project (MNFB). In particular, we incorporate the interval-censored structure of the bird counts to inform our model. We also examine whether a more fine-grained partitioning of intervals provides benefits over a coarser partioning.

2 Single Location Model

2.1 Binomial Model

For a binomial model with unknown N, p, the distribution for a single count is $p(n|N,p) = \binom{N}{n} p^y (1-p)^{N-n}$. In the case of multiple iid observations, the sufficient statistic for N and p will be $(n_{(k)}, \sum n_i)$, where $n_{(k)}$ is the largest number of birds observed, but in the single-observation case n is the sufficient statistic for both N and p.

[Is it worthwhile describing at all the history of point estimators, or should we just dive in with Bayesian estimation? If so, I should probably read Olkin, Petkau, and Zidek (1981). According to Raftery, Olkin et al. show that both the MME and the MLE for N are unstable in that a small change in data can lead to large changes in the estimate of N.]

If we place a $Beta(\alpha, \beta)$ prior on p, we find:

$$p(N|n) \propto \binom{N}{n} p(N) \frac{\Gamma(\beta+(N-n))}{\Gamma(\alpha+\beta+N)},$$

where $p(N)$ is our unstated prior on N. [Note: Link (2013) provides a posterior for N_i in the case where N_i varies from observation to observation].

The prior for N poses some difficulties in generating a proper posterior. Kahn (1987) warns particularly against a uniform prior on N. He demonstrates that the tail for the posterior on N for large N is proportional to $N^{-(\alpha+s)}$, where the prior for N is of the form N^{-s}. Thus the posterior is only proper if $\alpha + s > 1$. Additionally, if we want a finite expectation for N, then we require $\alpha + s > 2$. Berger (2008) favors $p(N) = \frac{1}{\sqrt{N}}$ when p is known. And when p is unknown, he favors $p(N) = \frac{1}{N}$ with either the Jeffrey's prior, Beta$(\frac{1}{2}, \frac{1}{2})$, or the uniform prior on p, depending on whether p is likely to take on extreme values (Jeffrey's handles the extremes better).

Additionally, Berger addresses the Beta-Binomial model, which is very like the binomial with a beta prior – the difference being that, in the beta-binomial, separate observations n_i derive from separate realizations of $p \sim Be(\alpha, \beta)$ with (α, β) known. In this situation, Berger recommends $p(N) = \frac{1}{\sqrt{N(N+\alpha+\beta)}}$ as being handily superior to $p(N) = \frac{1}{N}$.

2.2 Binomial-Exponential Model

For the dataset on hand, we have bird counts with interval-censored times of observation for each bird. These times of observation potentially provide information beyond the binomial model for estimating abundance and detectability. For the moment, we ignore the interval censoring and assume that observation times are precisely known. The observation period is a window from time = 0 to τ, after which no new birds are recorded. If we now assume an exponential waiting time with rate parameter ρ, the resulting distribution of observation times is:

$$p(x|\rho, N) = \left(\prod_i \rho e^{-\rho x_i}\right) e^{-\rho\tau(N-n)}; \quad x = (x_1, \ldots, x_n);$$

where, as before, N is the total population present and n is the number of individuals observed in one session. The MLE's are $\hat{N} = x_n$ and $\hat{\rho} = \frac{n}{\sum x_i + \tau(\hat{N}-n)}$.

An important feature of the exponential model is that ρ dictates the probability of detection. That is, the probability that a bird is observed during the τ-minute observation period is $p(p) = 1 - e^{-\tau\rho}$. We can use this to create a binomial-exponential model :

$$p(x, n|N, \rho) = \binom{N}{n}\rho^n \exp\left\{-\rho\left[\sum_{i=1}^{n} x_i + \tau(N-n)\right]\right\}.$$

From here, it is possible to assign priors to N and ρ. For ρ, the conjugate prior is $Ga(a, b)$. Meanwhile, for N, we continue to consider among the priors discussed above for binomial models, which we will refer to generically as $p(N)$. The result leads to a posterior for N:

$$p(N|x, n) \propto p(N)\binom{N}{n}\left[\tau(N-n) + \sum x_i + b\right]^{-(a+n)}$$

2.2.1 Equivalent priors on p and ρ

The standard priors on a binomial p are Beta(0,0), Beta($\frac{1}{2}, \frac{1}{2}$), and Beta(1,1). For consistency, we would like in our binomial-exponential model to adopt equivalent priors for the transformation $\rho = -\frac{1}{\tau}\log(1-p)$. Univariate transformation shows that a generic $p \sim \text{Beta}(\alpha, \beta)$ is equivalent to:

$$p(\rho) = ke^{-\tau\beta\rho}(1 - e^{-\tau\rho})^{\alpha-1} \text{ where } k = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\tau.$$

For the three beta priors listed, this leads to:

$$p(p) \sim \text{Beta}(1, 1) \longrightarrow p(\rho) = \tau e^{-\tau\rho}$$

$$p(p) \sim \text{Beta}(\tfrac{1}{2}, \tfrac{1}{2}) \longrightarrow p(\rho) = \tfrac{\tau}{\pi}e^{-\frac{\tau\rho}{2}}(1 - e^{-\tau\rho})^{-1/2}$$

$$p(p) \sim \text{Beta}(0, 0) \longrightarrow \lim_{\alpha=\beta\to 0} p(\rho) = 0, \text{ except at } \rho = 0, \text{ where the limit tends to } \infty$$

The first case is a Gamma(1,τ) prior on ρ, which is conjugate in the context of the binomial-exponential model. The second case has no clean analytical simplification. We will approximate it using numerical methods (or, we may try to find a gamma approximation for it at a later date). And the third case is clearly not at all practical.

3 Literature Notes – not a review as of yet, just notes

3.1 Berger (2008)

Reference priors for discrete parameter spaces.
Berger, J, Bernardo, J. and Sun, D. (2008). Technical Report.

Main interest is in a method for using continuous parameters as a mean of devising effective discrete priors. However, along the way, he addresses a problem very similar to ours.

Section 2: Binomial-Exponential Convolution

$N \sim Exponential(1/\lambda)$. Stops after R observations, with units stopping at times $t_1, \ldots, t_R$. So, N and λ are unknown, while R is known. Then:

$$p(t_1, \ldots, t_R|N, \lambda) = \tfrac{N!}{(N-R)!}\lambda^R \exp\left\{-\lambda\left[t_1 + t_2 + \ldots + t_R + (N-R)t_R\right]\right\}.$$

This is equivalent to what we have derived, where Berger's λ, R, t equates to our ρ, n, x:

$$p(x, n|N, \rho) = \binom{N}{n}\rho^n \exp\left\{-\rho\left[\sum_{i=1}^{n} x_i + \tau(N-n)\right]\right\}.$$

There are two differences between the models. The key difference seems to be that Berger uses t_R as the final observation time, while we use τ. In other words, he stops after R observations, while we stop at time τ. He cites Goudie & Goldie's (1981) derivation of the joint distribution of the sufficient statistics $(V = \frac{\sum t_i}{t_R}, W = t_R)$ [see Goudie & Goldie below for more details]. Apparently, the marginal density of V depends only upon R and N, not λ, and this is nice for reasons that are beyond me. The second difference is his ordering of x's so that $t_1 < t_2 < \ldots < t_R$, which leads to the inclusion of an extra $n!$? As seen later, this is the approach adopted by Goudie & Goldie (1981).
Some of the upshot is that there is no unbiased estimate of N. Additionally, the MLE and moment estimates of N are also 50% unlikely to exist. [Note: I think this claim is a bit exaggerated. The claim in Goudie & Goldie is that the MLE will not exist if the mean observation time is roughly greater than half the last observation time]. Also, $p(V|N) \to 1$ as $N \to \infty$, so that the posterior does not exist with a constant prior or even a $1/N$ prior.

Section 5: Binomial-Beta Distribution: N, p unknown

Berger's methodology generates two candidates for the prior on N: $\pi_1(N) = \frac{1}{\sqrt{n(n+\alpha+\beta)}}$ and $\pi_2(N) = 1/N$. Both yield proper posteriors. Berger argues that π_1 has an intuitive appeal in that, when α and β are large (representing strong knowledge about p), then π_1 behaves similarly to $1/\sqrt{n}$, their recommended objective prior in the case when p is known. He also examines frequentist coverage of credible sets arising from each prior and finds that π_2 can significantly overstate the coverage probability of credibility sets, while π_1 is both more accurate and more conservative. Furthermore, as expected, π_1 is better behaved as values of α and β grow larger.

Section 6: Binomial (no Beta)

First, Berger consider the case where p is known. This section is not so much directly appropriate for our study. As noted above, through the application of similar methods, Berger concludes that $1/\sqrt{n}$ is a good prior on n.
Then, in Section 6.2, he moves on to the Binomial with unknown n & p. Here he favors, for the sake of simplicity, the joint prior: $\pi^*(p, n) \propto \frac{1}{n} \times Be(p|, \frac{1}{2}, \frac{1}{2})$. However, he also considers a prior suggested by Raftery (1988a in Berger's bibliography): $\pi^R(p, n) \propto \frac{1}{n}$. Ultimately, Berger concludes that π^* handles extreme values of p more accurately, while π^R better addresses non-extremes. This conclusion conforms to what we might expect, given that π^R is uniform on p, while π^* adopts a Jeffrey's prior on p.

3.2 Goudie & Goldie (1981)

Initial size estimation for the linear pure death process. Biometrika 68(2): 543-550.

Abstract. Estimating N based on observed times of the first n deaths (i.e., a censored sample); n predetermined. They show that the only reliable non-Bayesian method is interval estimates. Even these are dubious. At last, they consider a truncated sampling model.

Their model is the same as Berger's (but adopting our notation):

$$p(x, n|N, \rho) = n!\tbinom{N}{n}\rho^n \exp\left\{-\rho\left[\sum_{i=1}^{n} x_i + \tau(N-n)\right]\right\}; x_1 < x_2 < \ldots < x_n$$

This is very similar to our model with the exceptions that their observations are ordered, and they cease observation at observation number n (which happens to occur at time τ), while we cease observation at time $= \tau$ (before which time there have been n observations). The sufficient statistic is $(\sum x_i/\tau, \tau)$. Because they treat n as fixed, they can assign a density to

$$f(\tau|\lambda, n) = n\tbinom{N}{n}\rho(1 - e^{-\rho t})^{n-1} \exp\{-\rho\tau(N - n + 1)\}(\tau > 0)$$

From here, they go on to solve the joint distribution of sufficient statistics and the marginal distribution of $\sum x_i/\tau$ (which they name α). It gets just a little hairy. However, their first figure starkly illustrates the

trouble for estimating N. They plot $f(\alpha|N)$ vs. α for $n = 10$ and show that the distribution asymptotically approaches a normal distribution with mean that looks to be $\frac{n+1}{2}$. Consequently, unless α is small, it becomes difficult to tell distributions apart. Furthermore, if α should happen to fall above $\frac{n+1}{2}$, then the MLE of N (which is drawn from $f(\alpha|N)$) will fail to exist. This non-existence turns out to hold for method of moments estimators as well.

Additionally, in Section 3.3 of their paper, Goudie & Goldie prove that there is *no* unbiased estimator of N.

Section 4 of their paper takes up confidence intervals from uniform most powerful tests. Their solutions are largely in terms of generic functions and regions such as $K(N)$ and $A^*(N)$. However, they generate a table of confidence intervals depending upon α and the confidence coefficient. Again, if α is very high or if the confidence coefficient is too strict, the confidence interval extends to ∞. However, at least there is a finite lower bound. They conclude that "these interval estimates therefore suggest that only fairly weak inference statements about the value of n are justified by the data, and that the apparent precision of any point estimate will often be largely spurious."

Section 5 of the paper deals with what they call Truncated Sampling, but which I called a Binomial-Exponential model above. It is exactly the same as our version of the model, except they continue to order their observations $x_1 < x_2 < \ldots < x_n$, and thus they continue to include $n!$ in their model. It is thus the exact same as their initial model with the exception that they fix τ instead of fixing n. However, now the sufficient statistics are $(n, \sum x_i/\tau)$. For the joint distribution of sufficient statistics, they refer us to Hoem (1969).

Again, the MLE fails to exist if $\alpha > \frac{n+1}{2}$. A method of moments estimatorfor N cannot be written independent of ρ. But it is easily shown that $\alpha > \frac{n}{2}$ is still a conundrum. Also, there remains no unbiased estimator of N. But worst of all <drumroll> **even interval estimation is unsatisfactory**. The problem is getting rid of the ρ. Furthermore, the likelihood ratio statistic is intractable. Sounds like this is a real bugger of a model.

On the bright side, the authors' recurring references to non-Bayesian frameworks suggest that Bayesian statistics may forge a more suitable solution.

3.3 Royle 2004

N-Mixture Models for Estimating Population Size from Spatially Replicated Counts

Abstract. He describes a class of mixture models which allow population estimation from temporally repeated sparse data counts and which use detection probability. The key is to veiw N as random variables distributed according to a mixing distribution.

His dataset consists of 50 sites sampled 10 times in one month by one observer. Thus, he is able to model each site as having a common N_i, and there is a common p across all sites. Nor is there any removal sampling involved.

His Literature Review. Repeated observations of a site yield a MME and MLE; however the estimators can be unstable for low p. One solution to this problem was proposed by Carroll and Lombard (1985); it involved integrating out p under a $Beta(a, b)$ prior, where both a and b are fixed a priori. But this estimator encounters troubles.

Royle's Model. Site-specific abundance is viewed as a random effect, where marginal likelihood of counts is obtained by integrating the binomial likelihood over possible values of abundance. This is similar to Carroll and Lombard, except that he's integrating over N instead of p (he argues that integrating out p does little to simplify the problem). He calls this an N-mixture model. He selects $N_i \sim \text{Poisson}(\lambda_i)$, where λ is the abundance per sampled area. The MLEs for p and for hyperparameters of λ are then maximized numerically; Royle uses these MLEs as plug-in estimates for N.

For his prior on N, Royle chooses the negative binomial, which results from a gamma prior on λ_i:

$f(N|\alpha, r) = \frac{\Gamma(N+\alpha)}{\Gamma(\alpha)N!} r^\alpha (1-r)^N$

This can simplified with $\mu = \alpha(1-r)/r$.

Covariates can be included as predictors of abundance or of detectability. For instance, we could model $\log(\lambda_i) = \sum_{j=1}^{r} x_{ij}\beta_j$. A logit model on p makes sense, too.

Critique of Carroll and Lombard. Before summarizing Royle, I just wish to note that the Carroll and Lombard approach seems to mirror our own. Develop a joint distribution for N, ρ and then integrate out ρ. Royle notes that the CL model requires specification of the beta parameters a, b in the beta prior of p, with unfortunate consequences in that estimators of N will be sensitive to this choice. Indeed, the CL estimator is biased for any p without a fortuitous choice of a, b. Additionally, for sparse counts, the CL estimator tends to return $N = n$ (which, when you think about it, is really bad, since sparse counts may indicate low p while $N = n$ indicates $p = 1$). Also zero-counts pose difficulties in the CL model. By pooling data across sites, one can overcome these issues, but then one has lost ones ability to say anything specific about sites (and, more to the point, about covariates that vary across sites). Finally, the CL approach of integrating out p means that one cannot model covariate effects upon p.

In the end, Royle's estimator was biased, but it's median/mode were okay. However, the CL estimator had dramatic bias. [Note: estimations of p were okay for Royle]. Coverage of the true N was decent for Royle's estimator but dubiously dubious for CL.

3.4 Raftery 1988

The useful part of Rafter from our perspective is his Poisson model. However, he does not have interval censored data, so the value of his model stops there.

Raftery develops an Bayesian model where $N \sim \text{Poisson}(\mu)$ with detection probability p. Then, for each observation period, we have $n_i \sim \text{Poisson}(\mu p), i = 1, \ldots, k$ where we can substitute $\lambda = \mu p$. This framing of the problem has the natural advantage that he can place priors on n_i, an observed value, rather than on N, which is never directly observed. Another benefit of his approach is that μ is continuous, whereas N is discrete. The posterior distribution for $N \geq n_{(k)}$ can be written:

$$p(N|n) \propto (N!)^{-1} \left\{ \prod_{i=1}^{k} \binom{N}{n_i} \right\} \int_0^1 \int_0^\infty p^{-N+S}(1-p)^{kN-S} \lambda^N e^{-\lambda/p} p(\lambda, p) d\lambda dp$$

where $S = n_1 + \ldots + n_k$. Raftery chooses a uniform prior on p and an inverse prior on λ so that $p(\lambda, p) \propto \lambda^{-1}$. This is equivalent to $p(N, p) \propto N^{-1}$. Carrying through the integrations, he gets

$$p(N|n) \propto \left\{ \frac{(kN-S)!}{(kN+1)!N} \right\} \left\{ \prod_{i=1}^{k} \binom{N}{n_i} \right\}.$$

For a single observation, this simplifies to $p(N|n) = \frac{n_1}{N(N+1)}; N \geq n_1$; with a median of $2n_1$. Raftery also provides a posterior distribution assuming a Gamma prior on λ, which I will not copy here.

Raftery uses a MSE loss function,the posterior mode, and the posterior median to derive point estimates for N. Then, he proceeds to generate interval estimates using the posterior distribution of N. His analysis follows a pattern established by Olkin et al. (1981) and Carroll and Lombard (1985). Ultimately, his point estimator proves more stable than either of theirs.

References to scout out: Kramer & Starr (1990) – from Berger
Basu & Ebrahimi (2001) – from Berger
! Blumenthal & Marcus (1975) – from Goudie & Goldie
Hoem (1969) – from Goudie & Goldie
! Olkin et al. (1981) – from Royle