AN INTRODUCTION TO BAYESIAN INFERENCE AND MCMC METHODS FOR CAPTURE-RECAPTURE

Trinity River Restoration Program
Workshop on Outmigration: Population Estimation

October 6–8, 2009
An Introduction to Bayesian Inference

1. The Binomial Model
   - Maximum Likelihood Estimation
   - Bayesian Inference and the Posterior Density
   - Summarizing the Posterior Density

2. MCMC Methods and the Binomial Model
   - An Introduction to MCMC
   - An Introduction to WinBUGS

3. Two-Stage Capture-Recapture Models
   - The Simple-Petersen Model
   - The Stratified-Petersen Model
   - The Hierarchical-Petersen Model

4. Further Issues in Bayesian Statistics and MCMC
   - Monitoring MCMC Convergence
   - Model Selection and the DIC
   - Goodness-of-Fit and Bayesian p-values

5. Bayesian Penalized Splines
1 The Binomial Model
- Maximum Likelihood Estimation
- Bayesian Inference and the Posterior Density
- Summarizing the Posterior Density
1 The Binomial Model

- Maximum Likelihood Estimation
- Bayesian Inference and the Posterior Density
- Summarizing the Posterior Density
Maximum Likelihood Estimation
The Binomial Distribution

Setup

- a population contains a fixed and known number of marked individuals ($n$)

Assumptions

- every individual has the same probability of being captured ($p$)
- individuals are captured independently

Probability Mass Function

The probability that $m$ of $n$ individuals are captured is:

$$P(m|p) = \binom{n}{m} p^m (1 - p)^{n-m}$$
If $n = 30$ and $p = .8$: 

![Probability Mass Function](image-url)
**Definition**

The likelihood function is equal to the probability mass function of the observed data allowing the parameter values to change while the data is fixed.

The likelihood function for the binomial experiment is:

\[
L(p|m) = \binom{n}{m} p^m (1 - p)^{n-m}
\]
If $n = 30$ and $m = 24$: 

![Likelihood Function Graph](image)
Definition
The maximum likelihood estimator is the value of the parameter which maximizes the likelihood function for the observed data.

The maximum likelihood estimator of $p$ for the binomial experiment is:

$$\hat{p} = \frac{m}{n}$$
If \( n = 30 \) and \( m = 24 \) then \( \hat{p} = \frac{24}{30} = .8 \):
Imagine that the same experiment could be repeated many times without changing the value of the parameter.

**Definition 1**
The standard error of the estimator is the standard deviation of the estimates computed from each of the resulting data sets.

**Definition 2**
A 95% confidence interval is a pair of values which, computed in the same manner for each data set, would bound the true value for at least 95% of the repetitions.

The standard error for the capture probability is:

\[ SE_p = \sqrt{\hat{p}(1 - \hat{p})/m}. \]

A 95% confidence interval has bounds:

\[ \hat{p} - 1.96SE_p \text{ and } \hat{p} + 1.96SE_p. \]
If $n = 30$ and $m = 24$ then:

- the standard error of $\hat{p}$ is: $SE_p = .07$
- a 95% confidence interval for $\hat{p}$ is: $(.66,.94)$
An Introduction to Bayesian Inference

1 The Binomial Model
   - Maximum Likelihood Estimation
   - Bayesian Inference and the Posterior Density
   - Summarizing the Posterior Density
Pilot Study

- Data: $n = 20, m = 10$
- Likelihood: $\binom{20}{10} p^{10} (1 - p)^{10}$

Full Experiment

- Data: $n = 30, m = 24$
- Likelihood: $\binom{30}{24} p^{24} (1 - p)^6$

Combined Analysis

- Likelihood: $\binom{20}{10} \binom{30}{24} p^{34} (1 - p)^{16}$
- Estimate: $\hat{p} = \frac{34}{50} = .68$
Prior Beliefs

- Hypothetical Data: \( n = 20, \ m = 10 \)
- Prior Density: \( \binom{20}{10} p^{10}(1 - p)^{10} \)

Full Experiment

- Data: \( n = 30, \ m = 24 \)
- Likelihood: \( \binom{30}{24} p^{24}(1 - p)^{6} \)

Posterior Beliefs

- Posterior Density: \( \binom{20}{10} \binom{30}{24} p^{34}(1 - p)^{16} \)
- Estimate: \( \hat{p} = \frac{34}{50} = .68 \)
Bayesian Inference and the Posterior Density
Combining Data with Prior Beliefs I
Prior Beliefs

- Hypothetical Data: \( n = 2, \ m = 1 \)
- Prior Density: \( \binom{2}{1} p(1 - p) \)

Full Experiment

- Data: \( n = 30, \ m = 24 \)
- Likelihood: \( \binom{30}{24} p^{24}(1 - p)^6 \)

Posterior Beliefs

- Posterior Density: \( \binom{30}{24} \binom{2}{1} p^{25}(1 - p)^7 \)
- Estimate: \( \hat{p} = \frac{25}{32} = .78 \)
Bayesian Inference and the Posterior Density
Combining Data with Prior Beliefs II

---

Introduction to Bayesian Inference: The Binomial Model, Bayesian Inference and the Posterior Density
An Introduction to Bayesian Inference

1. The Binomial Model
   - Maximum Likelihood Estimation
   - Bayesian Inference and the Posterior Density
   - Summarizing the Posterior Density
**Fact**
A Bayesian posterior density is a true probability density which can be used to make direct probability statements about a parameter.
Classical Point Estimates

- Maximum Likelihood Estimate

Bayesian Point Estimates

- Posterior Mode
Classical Point Estimates

- Maximum Likelihood Estimate

Bayesian Point Estimates

- Posterior Mode
- Posterior Mean
- Posterior Median
Summarizing the Posterior Density
Bayesian Uncertainty Estimates

Classical Measures of Uncertainty

- Standard Error
- 95% Confidence Interval

Bayesian Measures of Uncertainty

- Posterior Standard Deviation:
  The standard deviation of the posterior density.
- 95% Credible Interval:
  Any interval which contains 95% of the posterior density.
1. Bayesian inference for the binomial experiment
   File: Intro_to_splines\Exercises\binomial_1.R
   This file contains code for plotting the prior density, likelihood function, and posterior density for the binomial model. Vary the values of \( n \), \( m \), and \( \alpha \) to see how the shapes of these functions and the corresponding posterior summaries are affected.
2 MCMC Methods and the Binomial Model
   • An Introduction to MCMC
   • An Introduction to WinBUGS
2 MCMC Methods and the Binomial Model

- An Introduction to MCMC
- An Introduction to WinBUGS
Concept

If the posterior density is too complicated, then we can estimate posterior quantities by generating a sample from the posterior density and computing sample statistics.
Definition
A Markov chain is a sequence of events such that the probabilities for one event depend only on the outcome of the previous event in the sequence.

Key Property
If we choose construct the Markov chain properly then the probability density of the events can be made to match any probability density – including the posterior density.
**Definition**
A Markov chain is a sequence of events such that the probabilities for one event depend only on the outcome of the previous event in the sequence.

**Key Property**
If we choose construct the Markov chain properly then the probability density of the events can be made to match any probability density – including the posterior density.

**Implication**
We can use a carefully constructed chain to generate a sample any complicated posterior density.
2 MCMC Methods and the Binomial Model
   • An Introduction to MCMC
   • An Introduction to WinBUGS
1) Model definition

```r
## Likelihood function
m ~ dbin(p,n)
```

```r
## Prior distribution
p ~ dbeta(1,1)
```

2) Data list

```r
list(n=30,m=24)
```

3) Initial values

```r
list(p=.8)
```
1. WinBUGS for the Binomial Experiment

Use the provided code to implement the binomial model in WinBUGS. Change the parameters of the prior distribution for $p$, $a$ and $b$, so that they are both equal to 1 and recompute the posterior summaries.
3 Two-Stage Capture-Recapture Models

- The Simple-Petersen Model
- The Stratified-Petersen Model
- The Hierarchical-Petersen Model
Two-Stage Capture-Recapture Models

- The Simple-Petersen Model
- The Stratified-Petersen Model
- The Hierarchical-Petersen Model
**The Simple-Petersen Model**

**Model Structure**

### Notation
- \( n/m = \# \) of marked individuals alive/captured
- \( U/u = \# \) of unmarked individuals alive/captured

### Model
- Marked sample: \( m \sim \text{Binomial}(n, p) \)
- Unmarked sample: \( u \sim \text{Binomial}(U, p) \)

### Prior Densities
- \( p: p \sim \text{Beta}(a, b) \)
- \( U: \log(U) \propto 1 \) (Jeffrey’s prior)
Intro_to_splines\Exercises\cr_winbugs.txt
Three

Two-Stage Capture-Recapture Models
- The Simple-Petersen Model
- The Stratified-Petersen Model
- The Hierarchical-Petersen Model
The Stratified-Petersen Model

Model Structure

Notation

- $n_i/m_i = \#$ of marked individuals alive/captured on day $i$
- $U_i/u_i = \#$ of unmarked individuals alive/captured on day $i$

Model

- Marked sample: $m_i \sim \text{Binomial}(n_i, p_i), \quad i = 1, \ldots, s$
- Unmarked sample: $u_i \sim \text{Binomial}(U_i, p_i), \quad i = 1, \ldots, s$

Prior Densities

- $p_i$: $p_i \sim \text{Beta}(a, b), \quad i = 1, \ldots, s$
- $U_i$: $\log(U_i) \propto 1, \quad i = 1, \ldots, s$
The Stratified-Petersen Model

WinBUGS Implementation

Intro_to_splines\Exercises\cr_stratified_winbugs.txt
1. The Stratified-Petersen Model

Intro_to_splines\Exercises\cr_stratified_winbugs.txt

Use the provided code to implement the stratified-Petersen model for the simulated data set and produce a boxplot for the values of $p$ (if you didn’t specify $p$ in the sample monitor than you will need to do so and re-run the chain). Notice that the 95% credible intervals are much wider for some values of $p_i$ than for others. Why is this?
3 Two-Stage Capture-Recapture Models
- The Simple-Petersen Model
- The Stratified-Petersen Model
- The Hierarchical-Petersen Model
The Hierarchical-Petersen Model

Model Structure

Notation

- \( n_i/m_i = \# \text{ of marked individuals alive/captured on day } i \)
- \( U_i/u_i = \# \text{ of unmarked individuals alive/captured on day } i \)

Model

- Marked sample: \( m_i \sim \text{Binomial}(n_i, p_i), \ i = 1, \ldots, s \)
- Unmarked sample: \( u_i \sim \text{Binomial}(U_i, p_i), \ i = 1, \ldots, s \)
- Capture probabilities: \( \log(p_i/(1 - p_i)) = \eta_i^p \)

Prior Densities

- \( \eta_i^p: \eta_i^p \sim N(\mu, \tau^2), \ i = 1, \ldots, s \)
- \( \mu, \tau: \mu \sim N(0, 1000^2), \tau \sim \Gamma^{-1}(0.01, 0.01) \)
- \( U_i: \log(U_i) \propto 1, \ i = 1, \ldots, s \)
The Hierarchical-Petersen Model
WinBUGS Implementation

Intro_to_splines\Exercises\cr_hierarchical_winbugs.txt
Exercises

1. Bayesian inference for the hierarchical Petersen model
   Intro_to_splines\Exercises\cr_hierarchical_2_winbugs.txt
   The hierarchical model can be used even in the more extreme case in which no marked fish are released in one stratum or the number of recoveries is missing, so that there is no direct information about the capture probability. This file contains the code for fitting the hierarchical model to the simulated data, except that some of the values of $n_i$ have been replaced by the value NA, WinBUGS notation for missing data. Run the model and produce boxplots for $U$ and $p$. Note that you will have to use the gen inits button in the Specification Tool window to generate initial values for the missing data after loading the initial values for $p$ and $U$. 
Further Issues in Bayesian Statistics and MCMC

- Monitoring MCMC Convergence
- Model Selection and the DIC
- Goodness-of-Fit and Bayesian p-values
Further Issues in Bayesian Statistics and MCMC

- Monitoring MCMC Convergence
- Model Selection and the DIC
- Goodness-of-Fit and Bayesian p-values
**Definition**

The *traceplot* for a Markov chain displays the generated values versus the iteration number.

**Traceplot for \( U_1 \) from the hierarchical-Petersen model:**

![Traceplot](image)
Monitoring Convergence
Traceplots and Mixing

Poor Mixing

Good Mixing
**Definition**

The **MC error** is the amount uncertainty in the posterior summaries due to approximation by a finite sample.

**Posterior summary for** $U_1$ **after 10,000 iterations:**

![Node statistics](image)
**Definition**

The **MC error** is the amount uncertainty in the posterior summaries due to approximation by a finite sample.

Posterior summary for $U_1$ after 100,000 iterations:

<table>
<thead>
<tr>
<th>Node</th>
<th>mean</th>
<th>sd</th>
<th>MC_error</th>
<th>val2.5pc</th>
<th>median</th>
<th>val97.5pc</th>
<th>start</th>
<th>sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>U[1]</td>
<td>9731.0</td>
<td>297.5</td>
<td>11.74</td>
<td>9175.0</td>
<td>9731.0</td>
<td>10310.0</td>
<td>1</td>
<td>100000</td>
</tr>
</tbody>
</table>
**Definition**

A chain is **thinned** if only a subset of the generated values are stored and used to compute summary statistics.

Summary statistics for $U[1] – 100,000$ iterations:

![Node statistics](image1)

Summary statistics for $U[1] – 100,000$ iterations thinned by 10:

![Node statistics](image2)
**Definition**

The burn-in period is the number of iterations necessary for the chain to converge to the posterior distribution.

**Multiple Chains**

The burn-in period can be assessed by running several chains with different starting values:
**Definition**

The *Brooks-Gelman-Rubin* convergence diagnostic compares the posterior summaries for the separate samples from each chain and the posterior summaries from the pooled sample from all chains. These should be equal at convergence.

Brooks-Gelman-Rubin diagnostic plot for $\mu$ after 100,000 iterations:
1. Bayesian inference for the hierarchical Petersen model: convergence diagnostics

Intro_to_splines\Exercises\cr_hierarchical_bgr_winbugs.txt

This file contains code to run three parallel chains for the hierarchical-Petersen model. Implement the model and then produce traceplots and compute the Brooks-Gelman-Rubin diagnostics. To initialize the model you will need to enter 3 in the num of chains dialogue and then load the three sets of initial values one at a time.
Further Issues in Bayesian Statistics and MCMC

- Monitoring MCMC Convergence
- Model Selection and the DIC
- Goodness-of-Fit and Bayesian p-values
Model Selection
The Principle of Parsimony

Concept
The most parsimonious model is the one that best explains the data with the fewest number of parameters.
**Definition 1**

The $p_D$ value for a model is an estimate of the effective number of parameters in the model – the number of unique and estimable parameters.

**Definition 2**

The **Deviance Information Criterion** (DIC) is a penalized form of the likelihood that accounts for the number of parameters in a model, as measured by $p_D$. Smaller values are better.
Further Issues in Bayesian Statistics and MCMC

- Monitoring MCMC Convergence
- Model Selection and the DIC
- Goodness-of-Fit and Bayesian p-values
**Concept**

If the model fits well then new data simulated from the model and the parameter values generated from the posterior should be similar to the observed data.
**Definition 1**
A discrepancy measure is a function of both the data and the parameters that assess the fit of some part of the model.

**Example**

\[ D(u, U, p) = \sum_{i=1}^{n} (\sqrt{u_i} - \sqrt{U_i p_i})^2 \]

**Definition 2**
The Bayesian p-value is the proportion of times the discrepancy of the observed data is less than the discrepancy of the simulated data. Bayesian p-values near 0 indicate lack of fit.
Bayesian Penalized Splines
**Concept**

We can control the smoothness of a B-spline by assigning a prior density to the differences in the coefficients.

Specifically, we would like our prior to favour smoothness but allow for sharp changes if the data warrants.
Bayesian Penalized Splines

Model Structure

\[ y_i = \sum_{k=1}^{K+D+1} b_k B_k(x_i) + \epsilon_i \]

Error \( \epsilon_i \sim N(0, \sigma^2) \)

Hierarchical Prior Density for Spline Coefficients

Level 1 \( (b_k - b_{k-1}) \sim N(b_{k-1} - b_{k-2}, (1/\lambda)^2) \)

Level 2 \( \lambda \sim \Gamma(.05, .05) \)

The parameter \( \lambda \) plays the same role as the smoothing parameter:

- if \( \lambda \) is big then \( b_k \approx b_{k-1} \) and the spline is smooth,
- if \( \lambda \) is small then \( b_k \) and \( b_{k-1} \) can be very different.