Introduction to Bayesian Statistics

Carrie Groth
Postdoctoral Fellow
Department of Preventive Medicine (Biostatistics)
Feinberg School of Medicine
Northwestern University

December 18, 2018
Basic Properties of Bayesian Statistics

- Bayes Rule
- Likelihood, Priors, and Posterior

OpenBUGS Programming Example

- Model
- Data
- Initial Values
- General Steps in BUGS
- Background BUGS Processes
- Results of a Bayesian model
- Model Diagnostics

Big Picture:

- Classical vs. Bayesian Methods
- When might Bayesian methods be useful to me?

Bayesian Missing Data

Resources on Bayesian Statistics
Bayesian Statistics

- Area of statistics that combines information in the form of **priors** and a **likelihood** to obtain a posterior distribution and develop inference on the parameters of interest
  - **priors**: Information on unknown parameters; information provided by a clinician, other studies, or other information on particular unknown parameters
  - **likelihood**: considered the data; same as a likelihood under maximum likelihood
- Relies on conditional probability and Bayes Rule
Bayes Rule

Conditional Probability \((P(A|B))\): The probability of event \(A\) occurring given that \(B\) has already occurred.

- In terms of distributions: the distribution of \(A\) given these values of the parameters or data

Bayes Rule:

\[
P(A|B) = \frac{P(A) \times P(B|A)}{P(B)}
\]

\[
P(A|B) = \frac{P(A) \times P(B|A)}{\sum \{P(A) \times P(B|A)\}}
\]

Posterior Distribution = \(\frac{\text{prior} \times \text{likelihood}}{\sum \{\text{prior} \times \text{likelihood}\}}\)

Consider \(\sum \{\text{prior} \times \text{likelihood}\}\) to be a normalizing constant. Then, Primary Bayesian Statistics Principle:

\[
\text{Posterior} \propto \text{prior} \times \text{likelihood} \quad (1)
\]
Linear Regression Example: Predicting Birthweight

- Dataset in MASS R package
- Response of interest (Y): infant birthweight (in grams)
- Predictors:
  - mother’s weight (W)
  - mother’s race (R)
  - mother’s smoking status (S)
  - mother’s history of hypertension (H)

\[ Y_i = \beta_0 + \beta_1 W_i + \beta_2 R_i + \beta_3 S_i + \beta_4 H_i + \epsilon_i \quad \epsilon_i \sim iid \sim N(0, \sigma^2) \quad (2) \]

- We are interested in the regression coefficients (\( \beta \)) and the conditional variance \( \sigma^2 \)
- Assuming all data is complete (no missing data)
Likelihood

- Review from maximum likelihood
- Describes the model and the data
- Example: Birthweight

\[ Y_i \overset{iid}{\sim} N(\beta_0 + \beta_1 W_i + \beta_2 R_i + \beta_3 S_i + \beta_4 H_i, \sigma^2) \quad (3) \]
Priors: Distributions on parameters used for inference

Priors range in “informativeness” (differing opinions)

**Jeffrey’s Priors**: prior settings that are equivalent to doing a classical analysis using maximum likelihood

Priors may be based on information from clinicians or data from other studies

In general, the more informative the prior, the more the final inference will be based on the prior (instead of the data)

**Example: Birthweight**

- Weakly Informative Model: $\beta \sim N(0, 1000000)$ and $\sigma^2 \sim IG(0.01, 0.01)$
- Informative Model: $\beta_0 \sim N(0, 10000)$ $\beta_1 \sim Unif(0, 5)$ $\beta_2 \sim Unif(-250, 0)$ $\beta_3 \sim Unif(-500, 0)$ $\beta_4 \sim Unif(-500, 0)$ and $\sigma \sim Unif(0.1, 1000)$
Final inference is obtained from the posterior distributions

We will obtain a posterior distribution of each parameter of interest (posterior samples)

Inference will be dependent on the priors and the data
  - If the priors are non-informative or weakly informative, the inference will be mostly from the data
  - If the priors are informative, priors will have a larger impact on the inference (data will not matter as much)
Common Bayesian Programs

- OpenBUGS/WinBUGS
- JAGS
- STAN
- Built into SAS
- Built into R (rjags, R2OpenBUGS)
Parts of a Bayesian Model in BUGS

- Model statement
- Data
- Initial Values
model {

**Likelihood:**
for (i in 1:N) {
  bwt[i] ~ dnorm(mu[i],tausqy)
           beta[4]*smoke[i] + beta[5]*ht[i]
}

**Priors:**
for (j in 1:5) {
  beta[j] ~ dnorm(0,0.000001)
}
  tausqy ~ dgamma(0.01,0.01)
  sigma1 <- 1/tausqy
}

**Notes:**
dnorm(mean, precision): Uses notation where tausqy is the precision (1/variance)
dgamma(shape, rate): is a gamma distribution
list(x = c(8.0, 15.0, 22.0, 29.0, 36.0), # vectors
xbar = 22, N = 30, T = 5,

Y = structure( # matrix
.Data = c(151, 199, 246, 283, 320,
145, 199, 249, 293, 354,
147, 214, 263, 312, 328,
155, 200, 237, 272, 297,
135, 188, 230, 280, 323,
159, 210, 252, 298, 331,
141, 189, 231, 275, 305,
159, 201, 248, 297, 338,
177, 236, 285, 350, 376,
... .Dim = c(30,5)) )
## Data: Column Format

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>151</td>
<td>199</td>
<td>246</td>
<td>283</td>
<td>320</td>
</tr>
<tr>
<td>145</td>
<td>199</td>
<td>249</td>
<td>293</td>
<td>354</td>
</tr>
<tr>
<td>147</td>
<td>214</td>
<td>263</td>
<td>312</td>
<td>328</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>153</td>
<td>200</td>
<td>244</td>
<td>286</td>
<td>324</td>
</tr>
<tr>
<td>END</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Initial Values**: plausible values of each parameter. These values are where the chains for each parameter will start and what the chain will use to estimate the first iteration.

For each chain enter a list:

```r
# Chain 1 Initial Values
list(beta=c(0,0,0,0,0),tausqy=1)
# Chain 2 Initial Values
list(beta=c(1,1,1,1,1),tausqy=10)
# Chain 3 Initial Values
list(beta=c(-1,-1,-1,-1,-1),tausqy=5)
```
General Steps in BUGS

1. Model > Specification
   Highlight the word model and click check model

2. Highlight the word list in front of data; click load data

3. Identify number of chains

4. Click Compile

5. Highlight the word list in front each set of initial values and click load inits

6. Inference > Samples
   In node box, type the name of each parameter and select set after each parameter. Then place a * in the node box

7. Model > Update
   Type the desired number of iterations in updates box
   Click update
In order to obtain posterior samples, programs like BUGS use Markov Chain Monte Carlo (MCMC) based methods. The basis for these methods is based on probability theory. Some examples of these methods include:

- Gibbs Sampling
- Metropolis Hastings
- Rejection Sampling
- Slice Sampling
- Reversible Jump MCMC

These methods are required because it is not always possible to obtain a posterior for an parameter in closed form. If a closed form posterior is available, then may use direct sampling.
Gibbs Sampling

Uses full conditional distributions of each parameter of interest.

Example: Birthweight
We can solve for the full conditionals of $\beta$ and $\sigma^2$:
Then within each iteration ($j$):

1. Sample $\beta_j \mid \sigma^2_{j-1}, W, R, S, H, Y$
2. Sample $\sigma^2_j \mid \beta_j, W, R, S, H, Y$

Repeating steps 1 and 2 leads to posterior distributions of $\beta$ and $\sigma^2$.

Important point: one parameter depends on the other. So once we have sampled the $\beta$, we use that sample to estimate $\sigma^2$ within that iteration.
Once the chain is done updating, one can look at various statistics/results.

- In Sample Monitor Tool Window, click **history**. This will give you what we call **traceplots**
Mixing: Refers to if you see the chains shown in different colors overlap with each other.
Convergence: Appearance of Traceplots

*Convergence*: Refers to if the chains have settled down and are wiggling around around a single value; often first looked at through observing traceplots.

![Chain: Lack of Convergence](image1.png)

![Chains: Converged](image2.png)
To look at the distribution of the parameter estimates, we can click on \textit{density}.

Every parameter will have a distribution to it because we sampled a number of iterations. Therefore, all summaries are over the iterations.
To obtain statistics, we can click on stats. The percentiles desired are highlighted on the right of the Sample Monitor Tool.

For each parameter you will be displayed with:

- Mean
- standard deviation
- MC Error (Monte Carlo Error)
- 2.5 Quantile
- Median
- 97.5 Quantile
- Start
- Sample (number of iterations)
95th Percentile Credible Interval: There is a probability of 0.95 that the parameter is within this range
- More direct statement than a confidence interval
- Obtained using percentiles of the posterior samples (example: 2.5 percentile and 97.5 percentile of the posterior distribution)
- Posterior samples of different parameters can be combined using mathematical formulas to obtain another posterior distribution of a parameter relying off of the parameters in the model

95th Percentile Confidence Interval: We are 95% confident that the true parameter falls within this range
- The value either falls inside or it doesn’t
- This is not a probability statement
Model Diagnostics

- **Convergence Assessment**
  - Assessed with multiple chains with different starting values
  - Observe trace plots
  - Gelman Rubin Diagnostics (want to observe below 0.2)
  - Levels of MC Standard Error (lower is better)

- **Compare Models**
  - Deviance Information Criterion (DIC) (Inference > DIC before you update the model)
  - Posterior predictive model assessment tools such as widely applicable information criterion (WAIC) and leave one out cross validation (LOO); statistics focus on reproducing the data using the model (predicting the data)

- **Autocorrelation**
  - Too much correlation in the chain between one iteration to the next is bad. To avoid this, people often *thin* the chain, meaning they only use every X number of samples in the final chain.
  - Assessed using autocorrelation plots (Sample Monitor Tool > auto cor)
Birthweight Example:

We are going to compare results from a few different models:

1. Classical analysis

2. Weakly Informative Model:
   \[ \beta \sim N(0, 1000000) \]
   \[ \sigma^2 \sim IG(0.01, 0.01) \]

3. Informative Model:
   \[ \beta_0 \sim N(0, 10000) \]
   \[ \beta_1 \sim Unif(0, 5) \]
   \[ \beta_2 \sim Unif(-250, 0) \]
   \[ \beta_3 \sim Unif(-500, 0) \]
   \[ \beta_4 \sim Unif(-500, 0) \]
   and \[ \sigma \sim Unif(0.1, 1000) \]
Maximum Likelihood Results:
Example: Birthweight

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3335.5</td>
<td>(3158.8, 3512.1)</td>
</tr>
<tr>
<td>Mother’s Weight (centered)</td>
<td>4.6</td>
<td>(1.4, 7.9)</td>
</tr>
<tr>
<td>Mother’s Race (1=non-white)</td>
<td>-409.8</td>
<td>(-613.7, -206.0)</td>
</tr>
<tr>
<td>Mother’s Smoking Status (1=Yes)</td>
<td>-398.7</td>
<td>(-606.7, -190.7)</td>
</tr>
<tr>
<td>Mother’s Hypertension Status (1=Yes)</td>
<td>-521.7</td>
<td>(-928.7, -114.7)</td>
</tr>
</tbody>
</table>

**Table 1.** Regression coefficients under a normal maximum likelihood analysis
Weakly Informative Bayesian Analysis: Example: Birthweight

Priors: $\beta \sim N(0, 1000000)$ and $\sigma^2 \sim IG(0.01, 0.01)$

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
<th>Credible Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3302.5</td>
<td>(3121.9, 3477.2)</td>
</tr>
<tr>
<td>Mother’s Weight (centered)</td>
<td>4.6</td>
<td>(1.3, 7.9)</td>
</tr>
<tr>
<td>Mother’s Race (1=non-white)</td>
<td>-384.1</td>
<td>(-587.6, -178.3)</td>
</tr>
<tr>
<td>Mother’s Smoking Status (1=Yes)</td>
<td>-373.8</td>
<td>(-578.1, -163.4)</td>
</tr>
<tr>
<td>Mother’s Hypertension Status (1=Yes)</td>
<td>-496.0</td>
<td>(-892.6, -94.6)</td>
</tr>
</tbody>
</table>

Table 2. Regression coefficients under a weakly informative Bayesian analysis
Informative Bayesian Analysis: Example: Birthweight

Priors: $\beta_0 \sim N(0, 10000)$ $\beta_1 \sim Unif(0, 5)$ $\beta_2 \sim Unif(-250, 0)$ $\beta_3 \sim Unif(-500, 0)$ $\beta_4 \sim Unif(-500, 0)$ and $\sigma \sim Unif(0.1, 1000)$

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
<th>Credible Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1950.6</td>
<td>(1832.3, 2067.7)</td>
</tr>
<tr>
<td>Mother’s Weight (centered)</td>
<td>3.3</td>
<td>(0.4, 4.9)</td>
</tr>
<tr>
<td>Mother’s Race (1=non-white)</td>
<td>-8.5</td>
<td>(-44.1, -0.3)</td>
</tr>
<tr>
<td>Mother’s Smoking Status (1=Yes)</td>
<td>-10.6</td>
<td>(-54.8, -0.4)</td>
</tr>
<tr>
<td>Mother’s Hypertension Status (1=Yes)</td>
<td>-84.4</td>
<td>(-367.8, -3.0)</td>
</tr>
</tbody>
</table>

Table 3. Regression coefficients under an informative Bayesian analysis. Inference would be quite different here compared to the MLE case or weakly informative case. Estimates are quite different.
Conclusions from this Analysis

- The weakly informative Bayesian analysis developed results similar to a classical analysis.
- The informative prior scenario developed results that were quite different than what was found in the classical analysis.
The Big Picture
Bayesian Inference vs. Classical Inference

- **Support of Bayesian Statistics:**
  - Allows for incorporation of information from clinicians, older studies, and other information
  - Can be more flexible than maximum likelihood methods
  - More direct probability statements

- **Critiques of Bayesian Statistics**
  - Information may not be drawn only from the data. Priors may be too informative
  - Bayesian methods can be computationally intensive (involving Markov Chain Monte Carlo Methods)
  - P-values are not available in Bayesian Statistics
When might Bayesian methods be useful to me?

- When you want to incorporate expert information or background information into an analysis.
- You want to have a more exact probability statement.
- When it is easier to think of your problem in an iterative process (latent class modeling).
- When you want to preserve uncertainty in multiple points in your model that otherwise would have been disjoint.
- When you want an alternative to a complex process or a resampling procedure (maybe to avoid LASSO, Jacknife, EM algorithm, or other sampling procedures).
- When you have missing data.
Missing Data in Bayesian Statistics
Missing Data in Bayesian Statistics

- Treat the missing data like a parameter that we are estimating
- Draw repeated samples of the missing value obtaining a distribution of the missing data value
Example Birthweight
Let’s now imagine that we have missing data in some of our covariates and/or responses. One potential method that is used to impute the missing data is Gibbs sampling:

Within each iteration ($j$):

1. Sample missing data given $\sigma_{j-1}^2$, $\beta_{j-1}$, and the observed data. This gives you a complete set of measurements within this iteration to sample the remaining parameters.

2. Sample $\beta_j \mid \sigma_{j-1}^2, W, R, S, H, Y$ (including sampled missing data).

3. Sample $\sigma_j^2 \mid \beta_j, W, R, S, H, Y$ (including sampled missing data).

This will give us a posterior distribution of each missing value. Thus we have quantified the uncertainty on the missing data value.
Real World Applications

*Deepwater Horizon* Oil Spill Airborne Chemical Exposure Assessment: Gulf Longterm Follow-up Study (GuLF STUDY)

- Developed statistical methods to quantify airborne chemical exposures for workers who were involved in the response and clean-up of the *Deepwater Horizon* oil spill
- Missing data present: We had to estimate exposures accounting for uncertainty
Nutrition: Understanding the True Intake of Nutrients

- Treat the true intake of nutrients as missing data and try to estimate it accounting for the fact that self-reports are plagued with measurement error

Clinical Trials:

- Accounting for missing data in a Bayesian perspective has recently become popular
- Adaptive Bayesian clinical trial designs
Resources on Bayesian Statistics

Bayesian Statistics Books:
Bayesian Data Analysis
Bayesian Methods for Data Analysis
Applied Bayesian Statistics
Doing Bayesian Data Analysis

Programming References:
RJAGS Tutorial:
JAGS User Manual:
OpenBUGS User Manual and Other Information:
   http://www.openbugs.net/w/Manuals/
Bayesian Organizations

- Bayesian Statistical Section of the ASA
- International Society for Bayesian Analysis (ISBA)
- Special Interest Conferences:
  - Bayesian Non-parametrics
  - Spatial Analysis tools
  - Clinical Trials
References

- OpenBUGS manual and examples
- MASS R package birthweight dataset