Biol 206/306 – Advanced Biostatistics
Lab 12 – Bayesian Inference
Fall 2016

By Philip J. Bergmann

0. Laboratory Objectives

1. Learn what Bayes Theorem and Bayesian Inference are
2. Reinforce the properties of Bayesian Inference using a simple example
3. Learn what the Markov Chain Monte Carlo approach entails
4. Learn to obtain a posterior probability distribution from MCMC
5. Learn to test an hypothesis using MCMC output

1. Bayes Theorem and Bayesian Inference

Bayesian Inference (BI) is a statistical approach that is founded on Bayes Theorem. Bayes Theorem is a probabilistic formula that allows one to calculate a **posterior probability distribution**, given a dataset, a model or hypothesis, and a **prior probability distribution**. Bayes Theorem can be expressed as:

$$P(H_A|Data) = \frac{P(Data|H_A)P(H_A)}{P(Data)}$$

In Bayes Theorem, the approximation of *P(Data|H_A)* is actually a familiar quantity, in that it is the likelihood of your data given a model – this is something that you have obtained from analyses that you have done in previous labs. The term *P(Data)* is considered a scaling term that is important to making the mathematics work out, and will not be considered further today.

The term *P(H_A)* is the prior probability distribution, often simply referred to as the **prior**. This is a distribution that characterizes our understanding of the phenomenon being modeled before we look at our data. The prior can be obtained from previous studies published in the literature, a pilot study, a guess, or it can be **uninformative**. Each parameter in a model will have a prior. For example, if a parameter of interest can range from zero to one, and we have a pretty good idea that the parameter's value is >0.5, we might use a prior probability distribution that gives high probabilities for values 0.5-1, but low probabilities below 0.5 (for example, a normal distribution with mean 0.8 and variance 0.2, expressed as *N(0.8,0.2)*). If we have no idea what parameter value is most likely, we might use a prior that is uniform between zero and one.

Finally, the term *P(H_A|Data)* is the posterior probability distribution, also simply called the **posterior**. It takes both our data (the likelihood of our data, given the model) and the prior into consideration.

BI is attractive because it offers two main advantages over frequentist approaches. First, the posterior represents the probability of an hypothesis/model, given the data, and this is what one ultimately would like to estimate, not the probability of the data given the hypothesis. Second, BI allows for the incorporation of prior knowledge into a study. Since science works by building on previous work, a formal, quantitative framework to do this is attractive. However, when using BI, one must also be careful to interpret the results correctly. Specifically, the posterior represents the probability of our belief in a certain hypothesis, and not the relative frequency of events. In addition, parameters (including hypotheses, likelihoods, priors, and posteriors) are viewed as random variables that are sampled from a distribution, not as fixed values that we are trying to estimate.

2. A Simple Example Using Bayesian Inference

We will use a simple example to get more comfortable with Bayesian Inference. There is no dataset for this example, but it will show how BI can be used. The general format of this exercise is modified from Shahbaba (2011. Biostatistics with R. Springer: New York). In this example, you are studying how a trematode parasite influences reproduction in female stickleback fish. Since you discovered this new parasite of stickleback, you know nothing of how it influences reproduction of the fish. You collect some very basic data in the lab, counting how many females successfully lay eggs when parasitized and when not. Your results are as follow:

	Parasitized	Not Parasitized
# of Females that Lay Eggs	16	20
Total # of Females	20	20

You are interested in using the data in the table above to find the posterior probability distribution of a female reproducing if she is parasitized. ***From the information given, what would you choose the prior probability distribution to be? Simply describing its shape is okay.***

Population proportions (proportion of individuals doing something) is frequently modeled as a **beta distribution**, which is just another type of distribution, like normal or uniform. This is convenient because it also makes the calculation of a posterior from some data and a prior computationally easy (note that most models and applications of BI actually require extremely intensive math). The beta distribution ranges from 0 to 1, and has two parameters that define its shape, called α and β. The mean of a beta distribution can be calculated as:

$$mean = \frac{\alpha}{\alpha + \beta}$$

You can also plot the shape of this distribution using the beta density function, dbeta, in R:

```
> plot(seq(0,1,0.01),dbeta(seq(0,1,0.01),α,β))
```

In the **dbeta** function, the first argument is a vector that lists the values for which to calculate the beta density, and the second and third arguments are simply the two shape parameters.

Try this for the following values of the parameters. Describe each distribution.

α	β	**Description & Mean**
2	8	
1	1	
4	4	
8	2	

Given your answer to the question on the previous page, which of these options would you use for the prior probability distribution in the current analysis?

As mentioned above, the posterior can easily be calculated from the prior if you are happy modeling each with a beta distribution. You can calculate the posterior population proportion (proportion of females reproducing) as follows:

$$Posterior = Beta(\alpha + y, \beta + n - y)$$

In this equation, *Beta(α,β)* is the prior probability distribution, *y* is the number of individuals undergoing the event of interest (*i.e.,* laying eggs), and *n* is the total number of individuals studied. ***For the dataset provided on the previous page, calculate the posterior beta distribution by hand. What are its two shape parameters? Express your answer as "Beta(#,#)".***

For the sequence of beta distribution values given above (in your plots), save the beta distribution for your prior as object "prior1", and for your posterior as "post1". Assemble the sequence of beta distribution values, your prior densities and your posterior densities into a data frame. Plot the prior and posterior on the same plot against beta value, save the graph using the menus, and paste the plot here:

Assignment: (4 points to end of section)

What is the mean of your posterior? How does it compare to just calculating the proportion of females that laid eggs in the study? Why are the two values different?

As you may have noticed, using 20 individuals may not result in the most accurate estimates of how the trematode affects female reproduction. You repeat your study with more female sticklebacks. You also do the study in two lakes. Lake A is the same as was used in your pilot study, Lake B is new. This time you use 80 females that you infect and 80 as controls from each lake. Again, all the controls from both lakes lay eggs. Of the 80 infected females, 67 lay eggs from Lake A, and only 36 lay eggs from Lake B. ***Repeat the analysis for each lake, but use the posterior from your pilot study as the prior for both of these new analyses. Calculate the posteriors, expressing each as Beta(α, β):***

Also save the density functions for each of your posteriors along your sequence of beta values, as objects "postA" and "postB". Plot your new prior and both posteriors against the beta values, and paste the plot below.

3. The Markov Chain Monte Carlo Approach to BI

As mentioned above, for virtually all applications of BI, the mathematics are very daunting. In fact, it was not until the 1990s when scientists discovered how to get around the mathematical hurdles and estimate a posterior probability distribution for sophisticated problems. Shortly after, BI began to be applied in a variety of applications. In all of these applications, a Markov Chain Monte Carlo (MCMC) approach is used. An MCMC run starts at a random starting point in parameter space and follows a directed random walk through parameter space until likelihood is maximized. Once an MCMC arrives at near the maximal likelihood, it wanders in its vicinity, visiting various relatively likely values of parameters. It "visits" each parameter value in direct proportion to that value's posterior probability, and so a large sample of these visits produces a posterior probability distribution for each parameter and the likelihood. This is explained and illustrated in lecture.

MCMC analyses often run for millions or billions of iterations, and so can take hours, days, or even months to complete, and this is certainly a limitation of the approach. (In the estimation of phylogenetic trees, where any analysis can take a long time, BI actually provides a computational savings because it simultaneously estimates clade support values.) MCMC analyses can be done in R, but due to time limitations, we will not be doing the actual analyses. Instead, we will examine the raw output of some MCMC analyses that are already done and test hypotheses of trait evolution. Each of the MCMC analyses that you will use took approximately eight hours to run on a laptop.

The dataset that was analyzed consists of data for eighteen species of phrynosomatine lizards and a phylogenetic tree (Bergmann et al. 2009. Directional evolution of stockiness in lizards coevolves with ecology and locomotion in lizards. Evolution 63: 215-227). You will be testing for an evolutionary correlation between maximal relative horn length and relative velocity. Maximal relative horn length is a measure of the degree to which a species is protected by horns to ward off predators; and relative velocity is the maximal speed a lizard can run in body lengths per second. Defending oneself with horns and running away are two strategies for not getting eaten by a predator. The goal is to test whether slower lizards have longer horns (because they need to be better defended). We will test for trait co-evolution by using one MCMC analysis that uses a BM model of trait evolution for each trait, and one MCMC analysis that uses a BM model, plus has a parameter for the covariance between two traits. By comparing these models, we can test whether two traits have evolved in a correlated manner. ***What alternative approach could you use to test these hypotheses?***

Download the Excel spreadsheet for this lab and open it. The spreadsheet contains two worksheets, one for each MCMC analysis of relative horn length and relative velocity. One analysis assumes no correlation among traits, and one provides an extra parameter for covariance. Looking at the output, you will see that the first column contains the iteration number of each sampled iteration. The tree column contains which phylogenetic tree was used for each iteration. Since we had only one phylogeny, this is always the same. If you had multiple phylogenies that spanned the range of supported phylogenies, you could incorporate phylogenetic uncertainty into the analysis. People sometimes will use a posterior of phylogenetic trees to do this. The next column is the log-likelihood of the model with the parameter values for that iteration. The HMean column is the harmonic mean of log-likelihoods for the sampled parameter values up to the current iteration. The alpha parameter for each trait is the ancestral reconstruction for the base of the tree. The variance for each trait is the σ^2 parameter – the rate of evolution. Finally, the Covar column is the covariance between the two traits. If it is zero in all iterations, then the model does not include a covariance parameter and the traits are assumed to follow independent BM models of evolution. ***How many iterations was each MCMC analysis run for?***

The large number of iterations makes it necessary to sample from the MCMC run only once in a while because each iteration would otherwise take up one line in the spreadsheet, resulting in an excessively long spreadsheet. ***How frequently were the MCMC runs you have at your disposal sampled?***

As an MCMC run moves from its random starting position to the likelihood maximum, it is not sampling from the posterior probability distribution. This happens only once it has arrived near the likelihood maximum. The part of the MCMC where it is traveling towards the optimal area must be discarded from consideration, and is called **burn-in**. Each parameter value, and the likelihood will undergo a burn-in period and the longest burn-in should be used to determine what to discard. Only when the MCMC is stationary (in the vicinity of the likelihood maximum), can you study its posterior distributions. You can estimate the burn-in period by simply plotting log-likelihood or a parameter value against iteration number. The harmonic mean of the log-likelihood is provided because it gives the most rigorous estimate of the burn-in period and it will be used later to compare models. ***In Excel, plot the log-likelihood, the harmonic mean of log-likelihood, and each of the parameter values against iteration number. What is the approximate length of the burn-in period in each case to the nearest 10 million iterations? Fill in your findings in the table. Note that "Horn" is trait 1 and "rVel" is trait 2.***

Measure	Horn-rVel no correl	Horn-rVel correl
log(L)		
H. Mean (L)		
Alpha 1		
Alpha 2		
Variance 1		
Variance 2		
Covariance	N/A	

For one of these two MCMC analyses, provide a plot of log-likelihood and its harmonic mean against iteration number. Identify the burn-in period that you would use for that analysis. Using Excel to make your graph is fine, no need to use R. Insert the graph here.

Now that burn-in is identified for each analysis, you can calculate mean and standard deviation for the log-likelihood and parameters. For mean log-likelihood, you can simply take the harmonic mean from the last iteration of the analysis. Calculate its standard deviation in the usual way from the likelihood column.

Assignment: (6 points – to end of worksheet)
Delete the rows in each dataset that constitute the burn-in period, and then calculate mean (=AVERAGE()), and standard deviation (=STDEV()) for each quantity in the table below. The "fill down" capability in Excel will also be useful.

	Horn-rVel no correl		**Horn-rVel correl**	
Burn-in				
Measure	Mean	SD	Mean	SD
Likelihood				
Alpha 1				
Alpha 2				
Variance 1				
Variance 2				
Covariance	N/A	N/A		

The next step in the analysis is to compare models. Similar to the labs using a model-based approach, you have fitted two models to the pair of traits with the intent of finding the better model. ***What are the two models that you fit to the pair of traits?***

You could calculate AIC values to compare models, but a Bayesian Inference approach suggests that you should use a Bayesian approach to comparing models as well. A Bayesian Information Criterion (BIC) exists and can be used for this purpose. Another approach is to use **Bayes Factors** (BF). BFs compare two models, rely on a quantity called a **marginal likelihood**, and can be interpreted in a similar way to an AIC Δ value. When BF < 2, there is weak support for one model over the other – the two models are about equal. When BF > 10, there is very strong support for one model over another. BFs between 2 and 10 indicate moderate support for the better model. The marginal likelihood is estimated with the harmonic mean of the log-likelihood. Bayes Factors can be calculated as follows:

$$BF = 2(\log(\bar{L}_{Better\ Model}) - \log(\bar{L}_{Worse\ Model}))$$

Since the output already provides harmonic means of log-likelihoods, you can calculate *BF* simply by taking the difference in harmonic means of log-likelihood for the two models and multiply by two. ***Do this for your analysis. Provide the BF, and give your conclusions about which model is preferred and how strongly it is supported. What are your biological conclusions about the evolution of the traits?***

Using the best model, which trait is evolving faster (in the tables above, the traits are listed in order they appear in the output)? Explain your answer.

If the correlated evolution model is better supported, calculate the evolutionary correlation between the traits so that it is more easily interpretable. A correlation can be calculated from a covariance and standard deviations for two traits, which are estimated as the parameters of the model:

$$\boldsymbol{R = \frac{cov(X,Y)}{s_X s_y}}$$