Integrative Biology 200A Spring 2012

University of California, Berkeley Michael Landis – March 6th

**Phylogenetic Trees VI: Bayesian Phylogenetics**

**Why use Bayesian statistics?**

Your plane crashes in the Amazon. You are the only survivor. Nothing is salvageable from the wreckage except for a wildlife guide. Days pass, and you are on the brink of starvation. By fortune, you happen upon a fruit-bearing tree, and you gorge yourself on its pulpy red fruit. You realize this fruit could feed you for several weeks, perhaps until you find help. Things are looking up, until you recall one (suddenly) relevant table from your wildlife guide:

| | |
|---|---|
| 1% | of fruit are poisonous |
| 90% | of poisonous fruit are red |
| 40% | of non-poisonous fruit are red |

Should you despair? Bayes theorem says probably not:

$$
\begin{aligned}
&P(\text{poisonous} \mid \text{red}) \\
&\quad = P(\text{red} \mid \text{poisonous}) * P(\text{poisonous}) / P(\text{red}) \\
&\quad = 0.90 * 0.01 / (0.90 * 0.01 + 0.40 * 0.99) \\
&\quad = 0.0222
\end{aligned}
$$

Without Bayes theorem, this sort of question ranges from being unclear to unanswerable for the given probabilities. This calculation is the result of some very simple rules of probability. However, the formulation of Bayesian probabilistic models makes some scientists, statisticians, and philosophers uneasy. We'll start with some of the theory, then discuss the implications in phylogenetics, then review how Bayesian statistics are computed in practice.

**1. Some basic definitions and rules in statistics (cheat sheet)**

To understand how Bayes theorem works, we will review some basic terms and rules in probability by applying them to the poisonous red fruit example above.

- **Random variable**

  A, B

  Let's describe a random fruit by two random variables, A and B. We allow A to take on the values red or not red, and allow B to take on the values of poisonous or edible. The possible values assigned to random variables are called outcomes.

- **Probability distribution**

  P( . )

  A probability distribution assigns a probability to each outcome. An outcome must have a probability between 0 and 1. The sum of the probabilities of all outcomes must equal 1.

- **Conditional probability**

$$
\begin{aligned}
&P(A \mid B) && = P(A, B) / P(B) \\
&P(B = \text{poisonous} \mid A = \text{red}) && = P(A = \text{red}, B = \text{poisonous}) / P(A = \text{red}) \\
& && = 0.009 / 0.405 \\
& && = 0.0222
\end{aligned}
$$

- **Joint probability**

  $P(A, B) = P(A \mid B)\, P(B)$

  $P(A = \text{red}, B = \text{poisonous}) = P(A = \text{red} \mid B = \text{poisonous}) * P(B = \text{poisonous})$
  $= 0.90 * 0.01$
  $= 0.009$

- **Marginal probability**

  $P(A) = \sum_i P(A, B_i)$

  $P(A = \text{red}) = P(A = \text{red} \mid B = \text{poisonous}) * P(B = \text{poisonous})$
  $+ P(A = \text{red} \mid B = \text{edible}) * P(B = \text{edible})$
  $= 0.90 * 0.01 + 0.99 * 0.40$
  $= 0.405$

**2. Bayes Theorem**

Using the rules above, we can derive Bayes Theorem

$P(B \mid A) = P(A, B) / P(A)$
$= P(A \mid B)\, P(B) / P(A)$
$= P(A \mid B)\, P(B) / \sum_i P(A \mid B_i)\, P(B_i)$

or,

**Posterior = Likelihood * Prior / Data**

**How to think about Bayesian inference**

Your beliefs *a priori* → the likelihood of new evidence → your beliefs *a posteriori*

**Prior –** The probability of each hypothesis occurring before looking at the data.
**Likelihood** – The probability of the data given a hypothesis and model.
**Posterior –** The probability of the hypothesis occurring after looking at the data.

See Figure 1. The prior distribution is a Normal distribution around the mean of 0. The data is modeled as a Normal distribution whose likelihood is maximized when the mean is 1. The posterior distribution is proportional to the product of the prior and the likelihood. Notice how the posterior distribution is intermediate between prior beliefs and current evidence. Without any evidence, the posterior distribution is identical to the prior distribution. As the amount of evidence increases, the posterior distribution is comes to resemble the likelihood function.

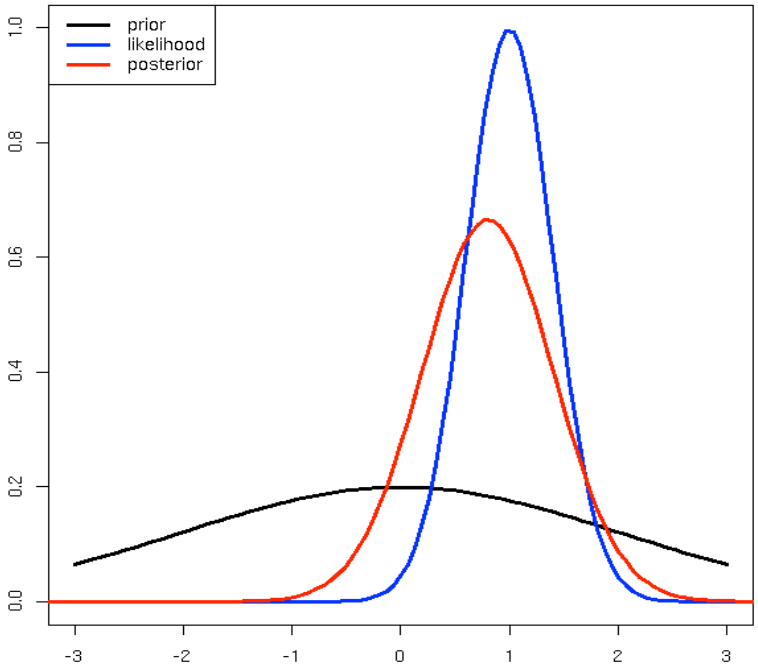



**Figure 1.** Prior, likelihood, and posterior.

## 2. An example of Bayesian inference

To understand these components, let's look at a simple example. Suppose I provide a coin and ask you to flip it ten times. We agree that I will buy you one donut per tails (T) and you will buy me one donut per heads (H). We observe the following sequence of coin flips: HHTHTHHTHTH.

### The likelihood function

The likelihood function is a way to explain the probability of the data given some parameters and some model of reality. Maximum likelihood will explain the data with parameters that optimize the likelihood function. For example, we could be interested in the probability of seven heads (*r*) and four tails on a fair coin for 11 trials (*n*), then model the observations using a Binomial distribution:

$P(r = 7 \mid p = 0.5, n = 11) = 0.1611328$

However, the maximum likelihood is observed to be where p=7/11 (from the data):

$P(r = 7 \mid p = 7/11, n = 11) = 0.2438488$

Say you've heard rumors I'm an insatiable donut addict and you doubt the integrity of the coin, you could then calculate the z-score

$z = (7/11 - 0.5) / (0.5^2 / 11)^{0.5} = 0.9045339 < 2.3263$

and conclude that there is not strong enough evidence to reject the null hypothesis (fair coin) at $p<0.01$, since for $z<2.3263$.

### The posterior distribution

The likelihood is the probability of the data given the parameters, but the posterior is the probability of the parameters given the data. From the previous example, we're seek the probability of the parameterization of the Binomial distribution given the coin flips we observe – that is, how fair or unfair is the coin given the coin-flip sequence is HHTHTHHTHTH? Consider these parameters to be competing hypotheses. The posterior distribution assigns the probability of any defined hypothesis being true given the observed data. This depends on what you assign as your prior distribution (recall the figure above).

### The prior distribution

The prior distribution assigns the probability of each event occurring in the absence of any data. This can be a little tricky to imagine, since it's precisely the data that you want to inform your model (the likelihood function). One extreme example may be to assign equal probabilities to all possible parameters, such as with Beta(1,1) above. This supposes you are ignorant to the probability of a coin coming up heads or tails, despite common sense and past experiences. Or oppositely, suppose you are absolutely certainty the coin was fair, and place that P(p=0.5) = 1.0, then the only posterior probability with a non-zero value would be for p=0.5. This is an extreme example, and returns us to a frequentist perspective where a single "true" parameter exists!

Hopefully you see how the prior distribution gives you immense control over how you input your beliefs into your model.

### Objective Bayesians

Objective Bayesians are interested in P(H | D), which is computable only through the use of P(H). However, Objective Bayesian adhere to the *principle of indifference,* and attempt to avoid inserting their beliefs or biases into their analyses by specifying an uninformative prior. In some situations, an uninformative prior distribution is easy to identify and is effectively neutral.

Of course, to assign "neutral" probabilities to all possible hypotheses, you must know the

set of all possible hypotheses. For discrete probabilities (i.e. coin-flips, die-rolls), this is possible. For continuous probabilities (i.e. durations of time, distances, etc.), how to do so may be unclear.

Another option is to parameterize your prior distributions with yet another distribution, called the hyperprior. The hyperprior, too, can be assigned its own hyperprior. But eventually, the modeler must eventually submit and declare his or her prior beliefs, be it on the first prior or on the umpteenth hyperprior.

Often, default settings for Bayesian software specify uninformative priors, which makes them suitable for general analyses. However, a misspecified uninformative prior can lead to unexpected results. For example, the default branch length prior in MrBayes is the exponential distribution, and has been held accountable for generating excessively long branches when compared to ML methods (Yang and Rannala, 2005).

**Subjective Bayesians**

Subjective Bayesians are interested in P(H | D), but wish to leverage the power of P(H) by using an informative prior. For instance, suppose one is modeling the rate of extinction for a specific clade of interest. One might consult the paleobiology literature and construct an empirical estimate of the rate of extinction across all lineages, then construct an *informative prior* that reflects a reasonable guess of what the extinction rate may be for a randomly sampled clade. An informative prior can drastically alter the posterior distribution, for better or worse. There's no formula how to mold expert prior beliefs into a probability distribution, which is part of the thrill of Subjective Bayesianism – you can justify your prior beliefs by saying, "well, this is how I see things." For example, Huelsenbeck et al. (2002) surveyed a panel of amateurs and experts, then treated their responses as a prior probability distribution for clade monophyly.

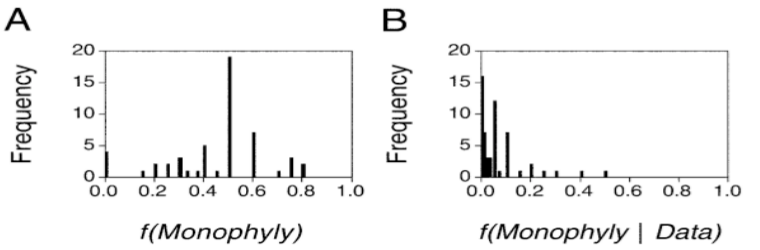


FIGURE 1. Frequency histograms of the responses concerning *Ipomoea* monophyly. (A) Prior beliefs. (B) Updated beliefs.

**Figure 3.** Prior distribution of surveyed beliefs (Huelsenbeck *et al*, 2002)

**Wait, what about the probability of the data occurring?**

To interpret the posterior as a probability distribution, we have to account for the probability of the data occurring, a.k.a. find the normalization constant. Nature does not typically supply probabilities for data occurring, so this can be difficult or impossible to define. Later, we'll peek at some of the statistical machinery that allows us to handle this problem. For the mean time, we'll suppose it's of no concern. Often, you'll see "the posterior is *proportional* to the likelihood times the prior", where the probability of the data remains an unknown "normalization constant".

**3. Philosophical Implications**

Karl Popper's requirement of falsifiability for hypothesis testing defines a meaningful hypothesis as one that can be rejected. Popper also concluded one can never hold something to be irrefutably true through induction – that is, theories can only be supported by immense amounts of evidence, yet new evidence can always cause the theory to be rejected. This introduces asymmetry between accepting and rejecting hypotheses. Whether or not you agree with Popper, his philosophy highlights an important distinction between frequentists and Bayesians.

**$P(D \mid H_i)$**

The frequentist asks, "What is the likelihood of the data given a hypothesis?" That is, the frequentist says there exists a true parameter, and a true model, and the data is generated through repeated trials. The frequentist perspective forbids assigning probabilities to hypotheses.

The frequentist can falsify a null hypothesis by using the p-value. Imagine you have a probability distribution, and you are interested in whether or not some random variable is reasonably described by that distribution. The p-value is a way to test whether or not the variable is in the middle of the distribution, or in tails of the distribution (i.e. an outlier). If the p-value is small enough, then the frequentist says, "My p-value is $< 0.05$, so this random variable is *practically* outside of my expectations under some null hypothesis, and so I accept an alternative hypothesis."

But which alternative hypothesis should we favor? Just because your data is not explained by a null hypothesis does not mean the alternative hypothesis you have in mind is the correct hypothesis. What if they are not all defined (or even known)? Since Popper says you can never fully accept a hypothesis, only reject one, is this necessarily bad?

**$P(H_i \mid data) = P(data \mid H_i)\, P(H_i)$**

The Bayesian asks, "What is the probability of a hypothesis given the data?" Bayesian analysis affords a very natural interpretation of the probability of one hypothesis versus another. This is a main attraction of Bayesian analysis, for which frequentist statistics have no equivalent. The null hypothesis becomes just one of all hypotheses that may be explored.

The freedom to describe the probabilities of arbitrary hypotheses has consequences. What if you fail to describe all possible hypotheses? Do your posterior probabilities represent what you think they do?

Let's take the Popperian point of view again. How do you falsify the probability of a belief when beliefs are non-falsifiable (i.e. how do you falsify the use of a specific prior distribution)? You can perform model consistency checks on simulated data, but what about real-world data?

## 4. Bayesian Phylogenetics

Now that we understand what the Bayesian perspective can offer to science, let's see what it can do for phylogenetics.

### Branch support measures

The likelihood function specifies the probability of observing the traits at the leaf nodes when we assume some evolutionary parameters and underlying phylogeny to be true. The maximum likelihood method attempts to find tree and parameters that maximize the likelihood of the observed data. Frequentist phylogeneticists use the bootstrap and jackknife methods to establish confidence intervals on the ML topology. Last week, we discussed the difficulties in interpreting these support measures.

The posterior defines the probability of a topology and its evolutionary parameters being true given the observed data and some model of evolution. This allows you ask questions such as, "What is the probability of having this particular clade and with traits evolving at this particular rate, given that I see this pattern of diversity and disparity at the tips?" just as you would ask, "What is the probability the coin is fair, given that we observed seven heads for 11 coin flips?"

In theory, the Bayesian approach solves this problem quite nicely. However, in practice, posterior probabilities of clades can sometimes produce misleading results. Yang and Rannala

(2005) demonstrated high prior probabilities for long internal branch lengths artificially increase the posterior probability of clades.

**Model testing (model comparisons)**

Suppose you are interested in whether the nucleotides in a protein of interest are better described by a Jukes-Cantor or a General Time-Reversible model of evolution. This calls for model testing, which can appear deceptively simple.

A frequentist may use the Likelihood Ratio Test (LRT) to determine which model produces the highest likelihood of the data occurring, but ignores model complexity and may favor the overfitting of complex models over simple models. Other tests, such as Akiake Information Criteria (AIC), attempt to correct for this by penalizing for additional parameters in an ad hoc fashion. An example LRT:

$$\Lambda(x) = \frac{L(\theta_0|x)}{L(\theta_1|x)} = \frac{f(x|\theta_0)}{f(x|\theta_1)}$$

The Bayesian equivalent is the Bayes Factor (BF), which is the ratio of the marginal likelihoods. Unlike LRT, the BF integrates over all possible parameters, so the dangers of model overfitting are eliminated. This is a good thing. An example BF:

$$K = \frac{\Pr(D|M_1)}{\Pr(D|M_2)} = \frac{\int \Pr(\theta_1|M_1)\Pr(D|\theta_1, M_1)\,d\theta_1}{\int \Pr(\theta_2|M_2)\Pr(D|\theta_2, M_2)\,d\theta_2}.$$

However, these integrals must be approximated. Notice that the integral relies on the prior distribution, P(theta). This introduces some difficulties into producing an accurate approximation. The harmonic mean estimator (the default in MrBayes) is quick but often overestimates model likelihood, and can produce incorrect model testing results. Alternative methods such as thermodynamic integration and the stepping-stone method resolve this issue, but are computationally intensive.

**5. Methods**

**Markov chain Monte Carlo (MCMC) & Metropolis-Hastings (MH) algorithm**

For phylogenetic inference, we're interested in exploring the posterior distribution of the trees and parameters given our data and model. One would like to directly calculate the likelihood for every single value for every single parameter. However, there are (2n-3)!! possible topologies, which by itself makes this an impossible task.

MCMC is a statistical method to approximate an arbitrary probability distribution. The algorithm explores the distribution step-by-step, preferring moves into regions of higher probability to those of lower probability. At intervals, MCMC records the current parameter values that correspond to a likelihood. After many steps (often millions), the MCMC will have spent time in each part of the probability distribution proportional to its posterior probability. By combining these samples (like a histogram), you can constitute the posterior probability.

MCMC decides how to move by proposing a new parameter value to explore. If MCMC prefers to explore regions of high probability, you may be worried MCMC would get stuck on local likelihood maxima. To avoid this problem, the Metropolis-Hastings algorithm is used. MH allows MCMC to move downhill into lower probability regions instead of only climbing uphill onto peaks.

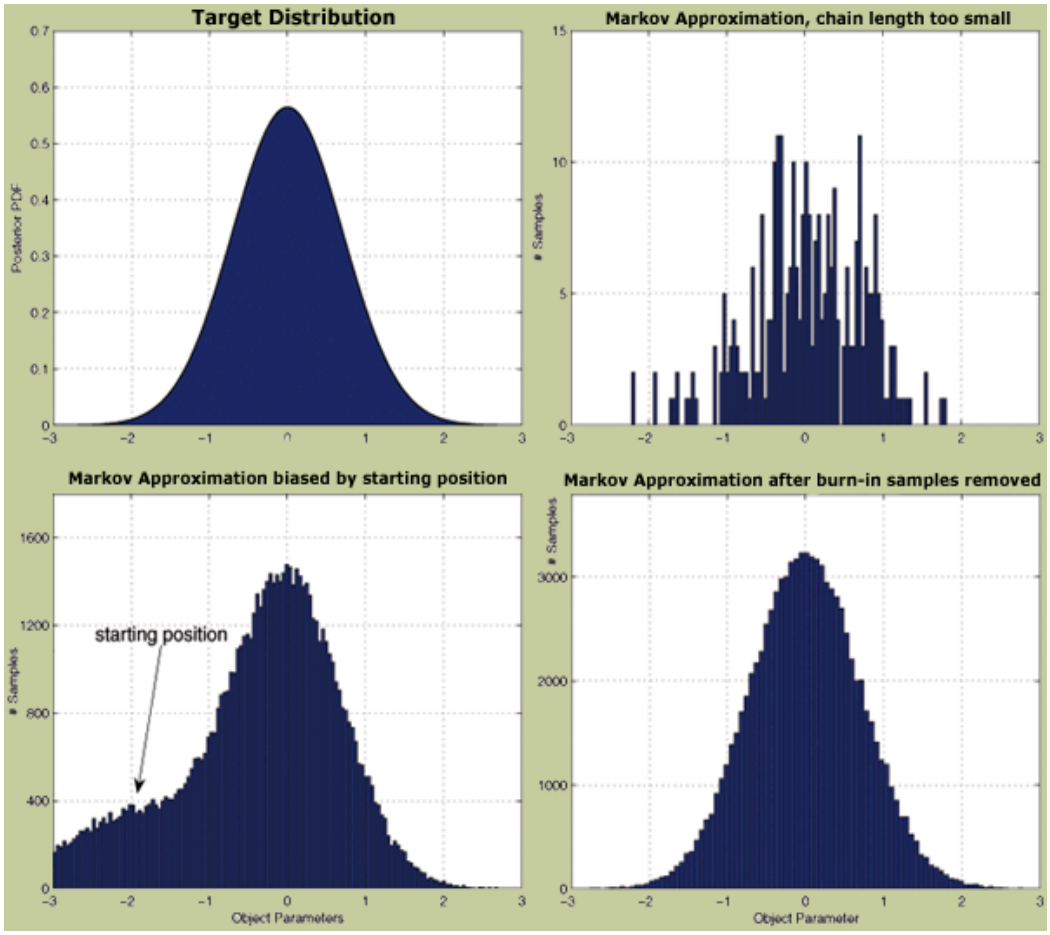
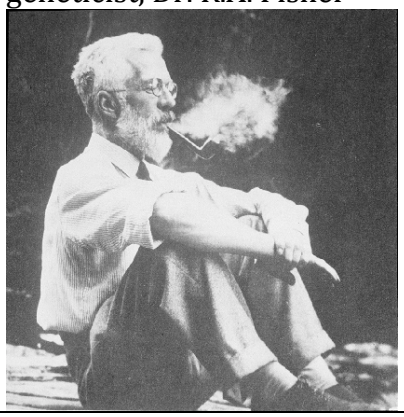



**Figure 4. Examples of MCMC output.**
Upper left: The "true" distribution. Upper right: Few MCMC samples. Lower left: Many MCMC samples, including burn-in. Lower right: Many MCMC samples, excluding burn-in, approximating the "true" distribution.

## 6. Summary

There's no obligation to be only a frequentist or only a Bayesian. Before choosing a method (or, inadvertently, a philosophy!), consider the question at hand:

| | **Frequentist** | **Bayesian** |
|---|---|---|
| **Question asked** | "Probability of data given hypothesis?" | "Probability of hypothesis given data?" |
| **Perspective** | Data is random.<br>Parameters are real. | Parameters are random.<br>Data are real. |
| **Functions used** | Likelihood | Posterior = Likelihood x Prior / Data |
| **As probability** | P(D \| H) | P(H \| D) = P(D \| H) x P(H) / P(D) |
| **Objective/subjective?** | Objective, if you believe there are "true" parameters and distributions producing your data. | Flexible, depending on how you decide to define your prior distribution. |
| **Methods** | Simple | Often complex (e.g. MCMC) |
| **Hypothesis test** | Can I reject my null hypothesis? | What are the relative probabilities of my hypotheses? |
| **Confidence Measure** | Significance test (p-value, bootstrap) | Posterior distribution |
| **How to Interpret Confidence Measures** | Unclear, debated | Very clear |
| **Model Comparison** | LRT, AIC | BF |
| **Engimatic Icon** | Statistician and population geneticist, Dr. R.A. Fisher  | Amateur mathematician and minister of the Nonconformists, Reverend Thomas Bayes  |

## References


Huelsenbeck J. P., Larget, B., Miller, R. E, Ronquist, F. (2002) *Potential Applications and Pitfalls of Bayesian Inference of Phylogeny*. Syst. Biol. 51(5):673-688.

Jaynes, E. T. (2003) *Probability Theory: The Logic of Science.* Cambridge Univ. Press. Edited by G. Larry Bretthorst.

MacKay, D. J. C. (2003) *Information Theory, Inference, and Learning Algorithms.* Cambridge Univ. Press.

Yang, Z. and Rannala, B. (2005) *Branch-Length Prior Influences Bayesian Posterior Probability of Phylogeny*. Syst. Biol. 54(3):455-470.


## Links

http://plato.stanford.edu/entries/epistemology-bayesian/
http://videolectures.net/mlss09uk_jordan_bfway/
http://yudkowsky.net/rational/bayes
http://www-gap.dcs.st-and.ac.uk/~history/Biographies/Bayes.html