Likelihood: Frequentist vs Bayesian Reasoning

Stochastic Models and Likelihood

A *model* is a mathematical formula which gives you the probability of obtaining a certain result. For example imagine a coin; the model is that the coin has two sides and each side has an equal probability of showing up on any toss. Therefore the probability of tossing heads is 0.50.

Models often have parameters, these are numerical variables that can take different values. Let's imagine that our coin does not have a 50% chance of turning up heads, but instead that the coin has probability α of turning up heads; α is now the only parameter in our coin flipping model.

To calculate the probability of multiple independent outcomes of our model we multiply the probability of each outcome together. For example the probability of getting heads on the first flip and tails on the second flip would be $\alpha(1-\alpha)$. In fact we can write a general formula for any combination of heads and tails:

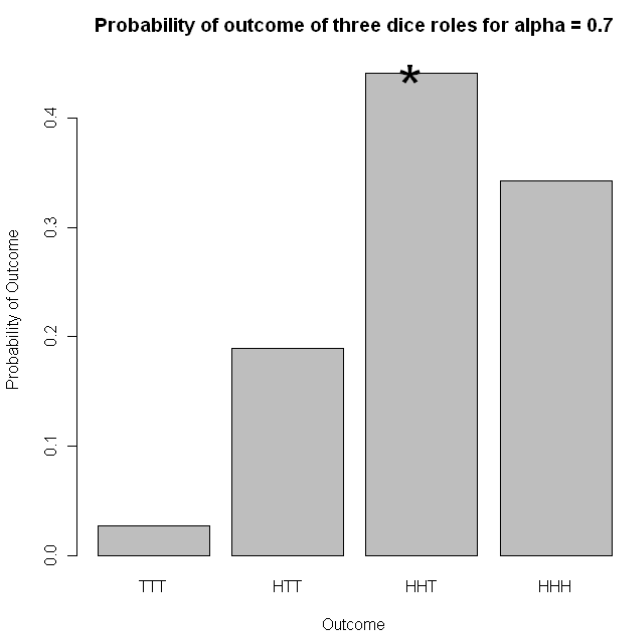
$$P(H,T \mid \alpha) \propto \alpha^{H}(1-\alpha)^{T}$$

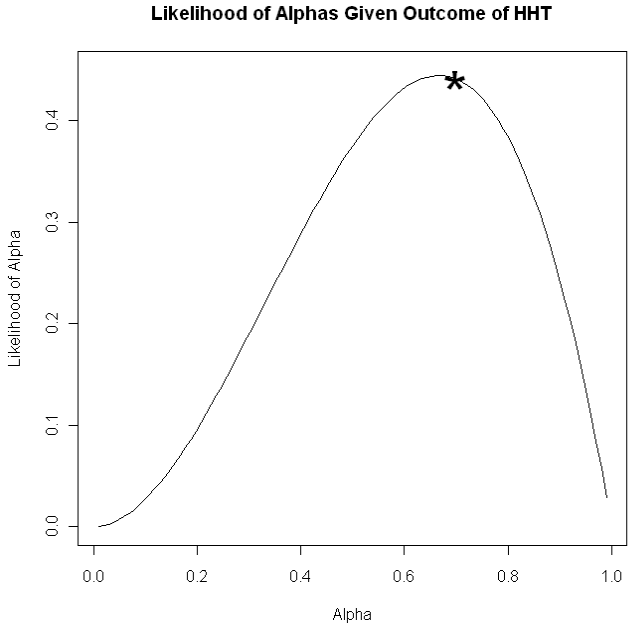
Often in the literature the distinction between a model and it's parameters is not entirely clear. Sometimes "model" refers to just the formula without the parameter values plugged in and some times "model" refers to the formula with a specific value for all the parameters. As a practical matter, the formula is usually held constant, and what we actually consider are different possible values for the parameters. Thus when we talk about the likelihoods of different models we are actually talking abut the likelihoods of different sets of parameter values. The cases where we actually consider different formulas are called *Model Tests* and we will deal with that at the end of the lecture. Generally speaking I will use the word "model" to mean both the model and the parameters; until we start talking about hypothesis testing, then "model" will refer to just the formula.

The *likelihood* of a model is the probability of the data given the model.

Likelihood of Model and Parameters = P(Data | Model (Parameters))

Up until now we have talked about the probabilities of outcomes given a model. However, what we are really interested in is picking the correct model. Calculating the probability of the model is not a straight forward business, but calculating the likelihood is relatively easy. It is important to distinguish between probabilities and likelihoods. The probabilities of all the different possible outcomes of a model must add up to 1. On the other hand the likelihoods of all the different possible models to explain a set of data do not have to add up to one. In fact the sum of the likelihoods will often have dimensions, a bad property for a probability.





Frequentist vs Bayesian Perspectives on Inference

The probability of a model given the data is called the *posterior probability*, and there is a close relationship between the posterior probability of a model and its likelihood that flows from some basic probability math:

$P(A\&B)=P(A|B)P(B) \quad P(A\&B)=P(B|A)P(A)$

$P(Model|Data)P(Data)=P(Data|Model)P(Model)$

$$\frac{P(M_1 \mid Data)}{P(M_2 \mid Data)} = \frac{P(Data \mid M_1)P(M_1)}{P(Data \mid M_2)P(M_2)}$$

Maximum Likelihood relies on this relationship to conclude that if one model has a higher likelihood, then it should also have a higher posterior probability. Therefor the model with the highest likelihood should also have the highest posterior probability. Many common statistics, such as the mean as the estimate of the peak of a normal distribution are really Maximum Likelihood conclusions.

Bayesian statistics on the other hand maintain that you can in fact calculate the *posterior probability* of each model using the formula below:

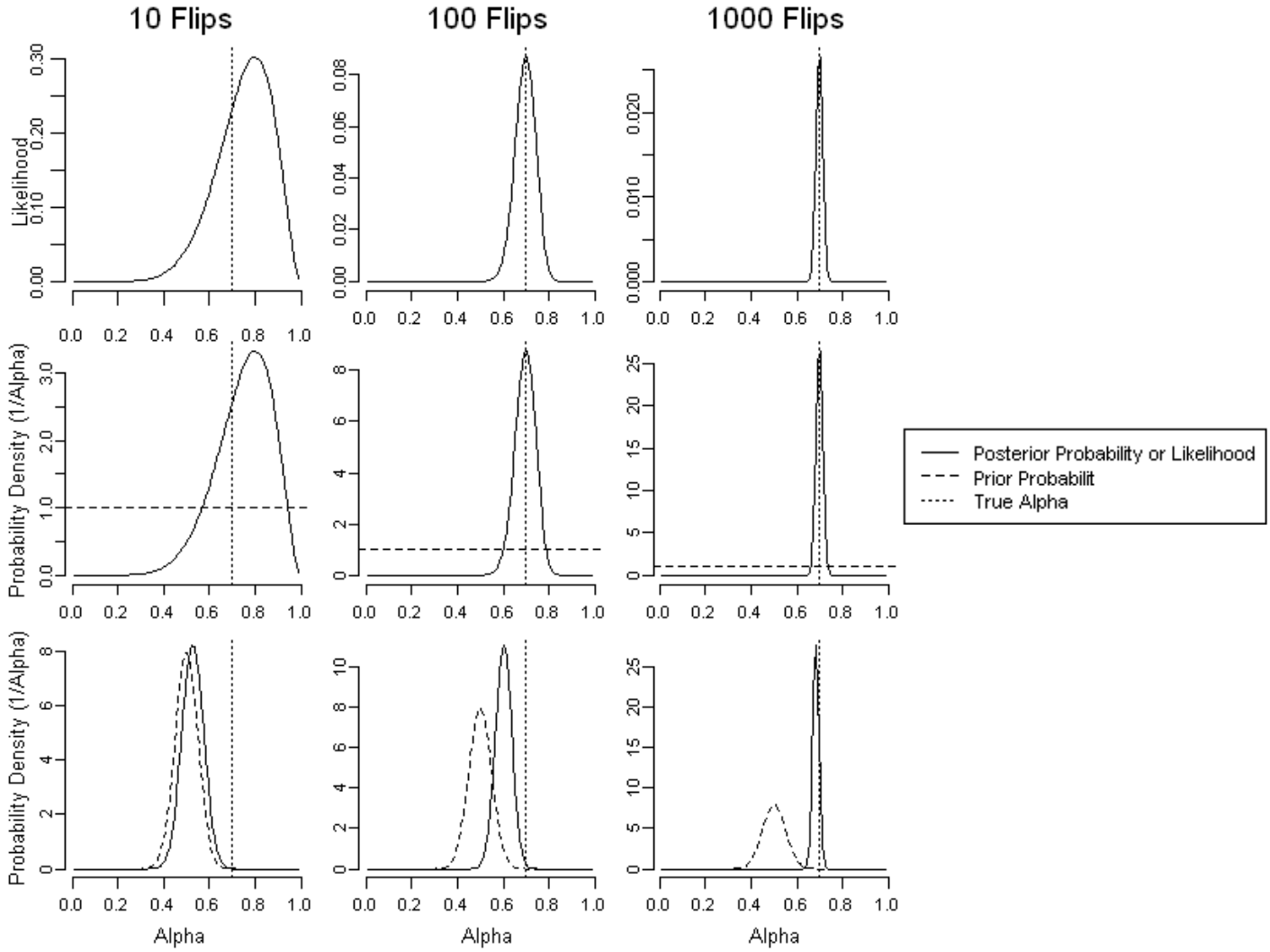
$$P(M_0 \mid Data) = \frac{P(Data \mid M_0)P(M_0)}{P(Data)} = \frac{P(Data \mid M_0)P(M_0)}{\sum_i P(Data \mid M_i)P(M_i)}$$

There are two things to note about this formula. The denominator is calculated as a sum over all the models. This can be very cumbersome to calculate, especially if there are many possible models. I will explain how this is dealt with shortly.

The more controversial part of this formula is *P(M)*, the *prior probability* of the

model. The idea is that this represents your idea about the probability of a model, before you consider the data. Determining what this value should be can be very controversial. One option is to choose an *uninformative* or *flat prior*, for which every model has equal probability; this is not always as straight forward as it appears. Another option is to choose an *informative* or *strong prior*, this contains information about the world that you have before investigating your data.

To return to our coin example, we have a very strong prior belief that $\alpha=0.50$. We have a lot of experience with coin flipping and they seam to turn up heads about half the time. Furthermore there are good *a priori* reasons based on our intuitive understanding of physics to believe that each side has an equal chance of turning up. Below I show plots of likelihoods and posterior probabilities with both a flat and a strong prior. The data for these plots are derived from a "coin" with $\alpha=0.7$.



There are several important things to note about these plots:

1) Although the posterior probabilities and the likelihoods have different scales, the shape of the likelihood plot and the posterior probability plot with flat priors are identical.

2) With lots of data, all the methods do a good job of estimating α; and they all do a pretty poor job with only a little data.

3) With intermediate amounts of data the maximum likelihood estimate of α is much closer to the actual value of α than the Bayesian estimate with strong priors. This could be seen in two ways. On face value this argues for the ML estimate. However, the Bayesian would argue that you really do have a good reason to believe that α is closer to 0.5, and a hundred coin flips should hardly effect your opinion that much. However, wasn't our prior a little arbitrary. I mean sure we can agree that most coins have a 50% chance of turning up heads; but why a normal distribution and how did you pick the variance? If you went out in the world and sampled millions of coins you could estimate what the prior distribution really is, but who's gonna do that?

We could go on like this forever, but there is an even more fundamental difference between what is called the *Frequentist* and the *Bayesian* perspective. Frequentists believe that there is a real α value out there in the world and we are trying to infer it. To talk about the probability of a particular value of α is silly. It either has that value or it does not. To a *Bayesian* we are always just approximating the world and to state our confidence in any conclusion we draw as a probability makes a lot of sense.

Search Algorithms: Hill Climbing and Metropolis-Hastings

There is also a practical matter of how these statistics are calculated. Maximum Likelihood relies on what are called hill climbing algorithms, These are exactly what they seam. From any point the algorithm modifies the parameter values in such a way that it increases the likelihood. When it reaches the peak it can no longer improve the likelihood and it is finished. There is a chance that one of these algorithms could end up stuck in a local maximum that is not in fact the global maximum. There are several methods to avoid this problem.

Bayesian statistics almost always uses the *Metropolis-Hastings* algorithm. In fact the Metropolis-Hastings algorithm is probably the main reason that anybody actually uses Bayesian statistics. If it weren't for this algorithm Bayesian statistics would be some obscure thing argued about in statistics departments, and no biologist would care.

The Metropolis-Hastings algorithm is a *Markov Chain Monte Carlo Method (MCMC)* that relies on a theoretically simple formula to explore the likelihood space. I won't go through the details here, but it travels around the likelihood space and ends up sampling every value of the parameters in proportion to their posterior probability. This has three great advantages: it calculates the posterior probability without ever having to calculate that ugly denominator; it combines searching for the best parameters values with exploring likelihood space; and it provides a sample distribution of every parameter.

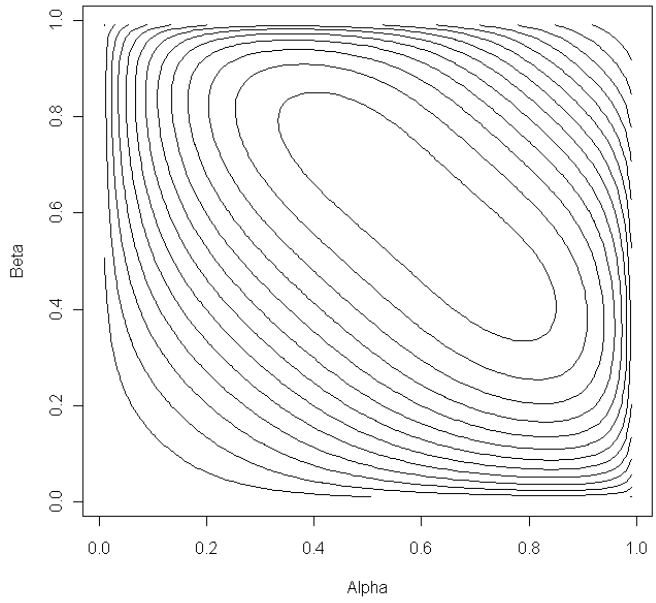
To understand this last point, you must understand what a sample is. The output of an MCMC will be a table. Each row will be a different sample from the likelihood space. It will contain a likelihood and a value for each parameter. The probability of any parameter value appearing in one of those rows is equal to its posterior probability, so if

you made a histogram of that variable it would match the plot of posterior probability.

	Sample	Likelihood	Alpha
Example of MCMC output:	**1**	**-17.058**	**0.4322**
	100	**-54.913**	**0.2196**
	200	**-2.4997**	**0.7151**
	300	**-4.8525**	**0.5942**

Nuisance parameters: Joint vs Marginal Estimation

So far we have only dealt with simple one parameter models. Usually likelihood models are much more complicated and rely on several variables. As a trivial example, imagine that we are now flipping two coins. We know what combination of heads and tails we got, but we do not know which coin got what. In this case to calculate the probability of any coin flip we must know the probability of each coin turning up heads, α and β. We can make a plot of the likelihoods (or for that matter the posterior probabilities) for the particular values of α and β, given that we got 7 HH, 10 HT and 3



TT, where the likelihood is represented by contours.

It is often the case that we are not interested in estimating both of these values, but are instead only concerned with one of those values. However, we must also estimate the other parameter in order to get a good estimate of our important parameter. These other parameters, which we are uninterested in are called *nuisance parameters*. ML and Bayesian inference deal with these parameters in very different ways.

ML uses *joint estimation*, meaning that it maximizes the likelihood for all parameters at once. Your estimate of the maximum likelihood of any one parameter is

based on your maximum likelihood estimate of every other parameter in the model.

$$\max[P(Data \mid \alpha, \beta)]$$

Bayesian inference uses *marginal estimation.* The posterior probability of any one particular value for your parameter of interest is calculated by summing over all possible values of the nuisance parameters. In this way your estimation of any one parameter does not rely on your estimation of any other parameter. Of course it does depend on the priors that you assume for all your parameters, and we have already discussed how that can be fraught with difficulty. The *Metropolis-Hastings* algorithm automatically samples parameters values in proportion to their marginal likelihoods.

$$P(\alpha \mid Data) = \frac{p(\alpha)\int P(Data \mid \alpha, \beta)P(\beta)d\beta}{\int\int P(Data \mid \alpha, \beta)P(\alpha)P(\beta)d\alpha d\beta}$$

Hypothesis Testing

Often we are not actually interested in what the parameter values are, instead we are interested in what model (from this point on by model I mean formula) best fits our data. By asking questions about models instead of parameters we can draw conclusions about what patterns we observe in the world. You may not be surprised to learn that Frequentists and Bayesians have different ideas about how this should be done. I will start with the Frequentist perspective, because it is the more classic perspective and the one you are all more familiar with.

Frequentists ask, "Is the data probable given the null hypothesis?" If you can reject the null hypothesis as extremely improbable then you can tentatively accept your alternative hypothesis. Null hypotheses should be *nested* within the alternative hypothesis.

For one hypothesis to be nested within another, means that it has fewer parameters and that those parameters are a subset of the parameters in the more general model. Several more constrained models are nested within our two coin model. For example you could have the same model but assume that the two coins have the same probability of turning up heads, $\alpha=\beta$, or you could assume that either coin has an even chance of turning up heads or tails, $\alpha=0.5$ or $\beta=0.5$. All three of these models are nested within our more general model, but none of them are nested within each other. The model in which $\alpha=\beta=0.5$ is nested within all three of these models in addition to being nested within our more general model. On the other hand, $\alpha=\beta=0.7$, would only be nested within are most general model and the $\alpha=\beta$ model.

The *Likelihood Ratio Test* can be used to reject a null hypothesis in favor of a alternative hypothesis. Models can not have a higher maximum likelihood than a model that they are nested within. Therefore the issue is not does the more general model have a larger maximum likelihood, but rather is its maximum likelihood larger enough to reject the null hypothesis. The test statistic for this result is two times the natural log of the ratio between the alternative hypothesis and the null hypothesis. This statistic can be compared to the chi-squared distribution, in order to decide whether to reject the null

hypothesis.

Alternative Hypothesis 1: $M_2(\alpha,\beta)$; Null hypothesis: $M_1(\alpha)=M_2(\alpha,\alpha)$

$$LRT = \ln\left(\frac{\max[P(Data \mid M_2(\alpha,\beta))]}{\max[P(Data \mid M_1(\alpha))]}\right)$$

Compare 2*LRT to X^2 to calculate p-value

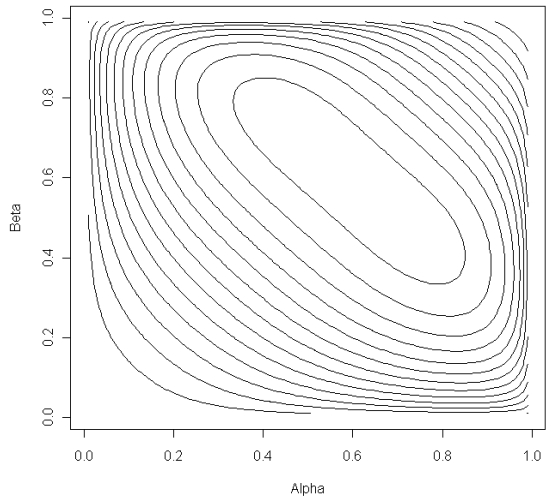
The Bayesian perspective on hypothesis testing is quite a bit different. They are not concerned with rejecting the null hypothesis. Bayesians ask, "To what degree does our data support one hypothesis over the other?" To answer this question they calculate a *Bayes Factor*. The Bayes Factor is the ratio between the *marginal likelihoods* of the two different models.

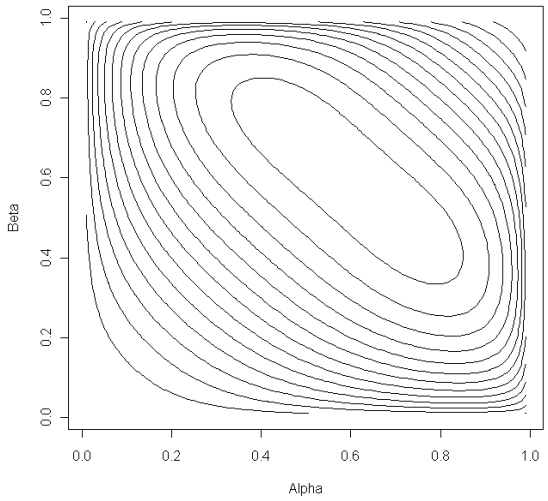
$$BayesFactor = \frac{P(Data \mid M_1)}{P(Data \mid M_2)}$$

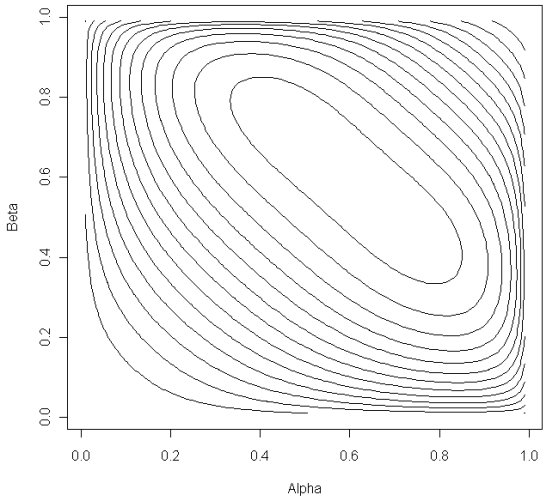
The marginal likelihood is the likelihood of the model summed over all the possible parameter values. In this way it does not depend on any particular parameter value. One way to look at it is that a Likelihood Ratio is a comparison between the heights of the peaks of the likelihood plot, while the Bayes Factor is a comparison between the average heights of the likelihood plots.

Bayes factors can compare any pair of hypotheses, whether or not they are nested. A Bayes factor of 10 is considered strong support for the hypothesis in the numerator, and a Bayes Factor of 100 is considered a slam dunk.

Bayes Factors can be calculated in two ways. A separate MCMC can be run for each model, and the marginal likelihood of each model can be calculated as the harmonic mean of the sample likelihoods (see the lab). Alternatively, one can run a reversible-jump MCMC, which can explore both models as part of its search. Each model will be sampled in proportion to their posterior probability, and a Bayes Factor can be calculated by factoring out the prior probability of the models.


Beta
0.0
0.2
0.4
0.6
0.8
1.0
Alpha


Beta
0.0
0.2
0.4
0.6
0.8
1.0
Alpha


Beta
0.0
0.2
0.4
0.6
0.8
1.0
Alpha