

# Bayesian inference & Markov chain Monte Carlo

**Note 1:** Many slides for this lecture were kindly provided by Paul Lewis and Mark Holder

**Note 2:** Paul Lewis has written nice ipad/iphone/Windows software for demonstrating Markov chain Monte Carlo idea. Software is called "MCMCRobot" and is freely available at the itunes store. See also...

<http://www.mcmicrobot.org>

Assume we want to estimate a parameter  $\theta$  with data  $X$ .

Maximum likelihood approach to estimating  $\theta$  finds value of  $\theta$  that maximizes  $\Pr(X | \theta)$ .

Before observing data, we may have some idea of how plausible are values of  $\theta$ . This idea is called our prior distribution of  $\theta$  and we'll denote it  $\Pr(\theta)$ .

Bayesians base estimate of  $\theta$  on the posterior distribution  $\Pr(\theta | X)$ .

$$\begin{aligned}
\Pr(\theta | X) &= \frac{\Pr(\theta, X)}{\Pr(X)} = \frac{\Pr(X | \theta)\Pr(\theta)}{\int_{\theta} \Pr(X, \theta)d\theta} \\
&= \frac{\Pr(X | \theta)\Pr(\theta)}{\int_{\theta} \Pr(X | \theta)\Pr(\theta)d\theta} \\
&= \frac{\text{likelihood} \times \text{prior}}{\text{difficult quantity to calculate}}
\end{aligned}$$

Often, determining the exact value of the above integral is difficult.

# Problems with Bayesian approaches in general:

1. Disagreements about philosophy of inference  
&  
Disagreements over priors
2. Heavy Computational Requirements  
(problem 2 is rapidly becoming less noteworthy)

## Potential advantages of Bayesian phylogeny inference

Interpretation of posterior probabilities of topologies is more straightforward than interpretation of bootstrap support.

If prior distributions for parameters are far from diffuse, very complicated and realistic models can be used and the problem of overparameterization can be simultaneously avoided.

MrBayes software for phylogeny inference is at:

<http:// mrbayes.sourceforge.net/download.php>

RevBayes software is at: <http:// revbayes.github.io>


Let  $p$  be the probability of heads.

Then  $1-p$  is the probability of tails

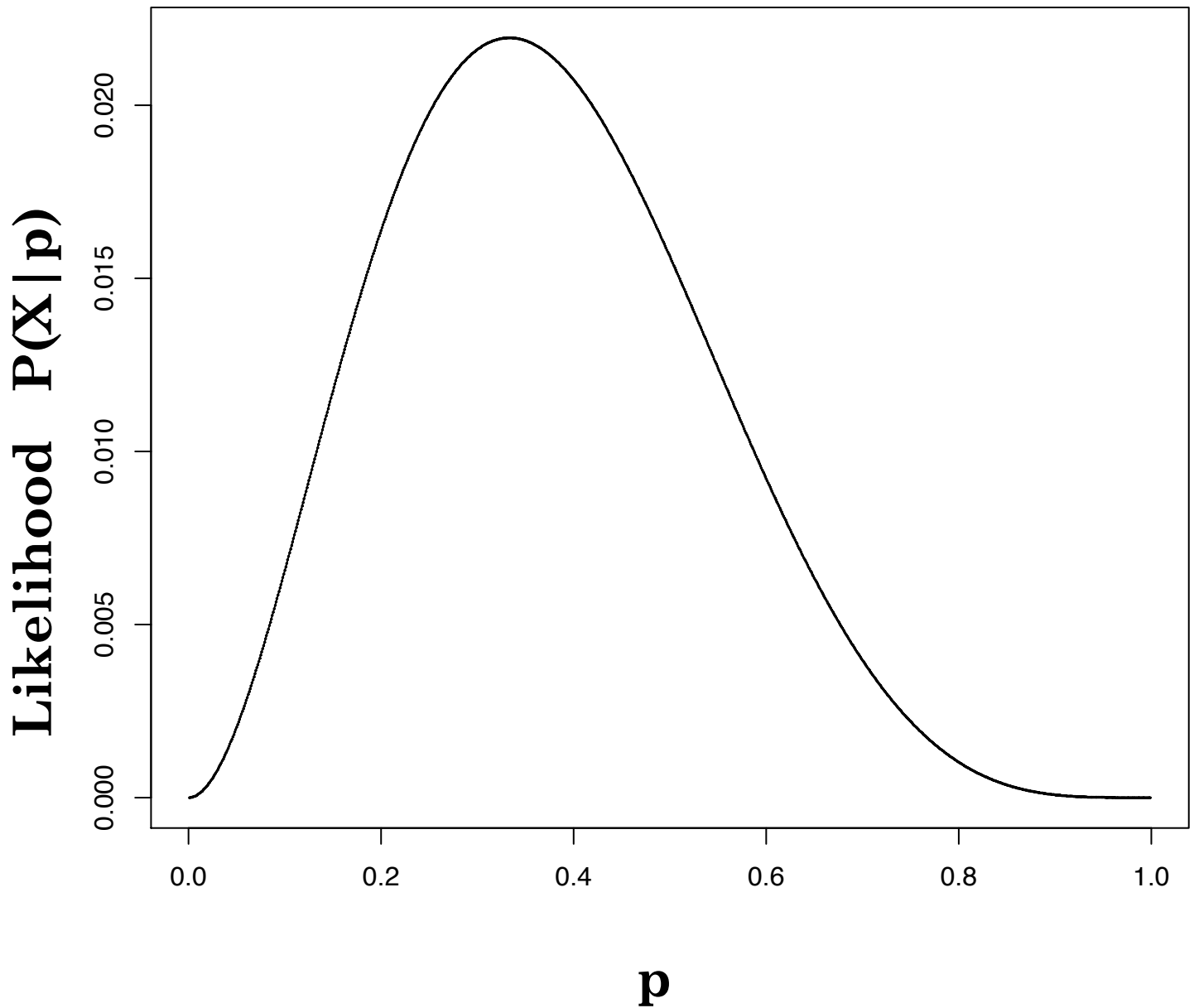
Imagine a data set  $X$  with these results from flipping a coin

Toss	1	2	3	4	5	6
Result	H	T	H	T	T	T
Probability	$p$	$1-p$	$p$	$1-p$	$1-p$	$1-p$

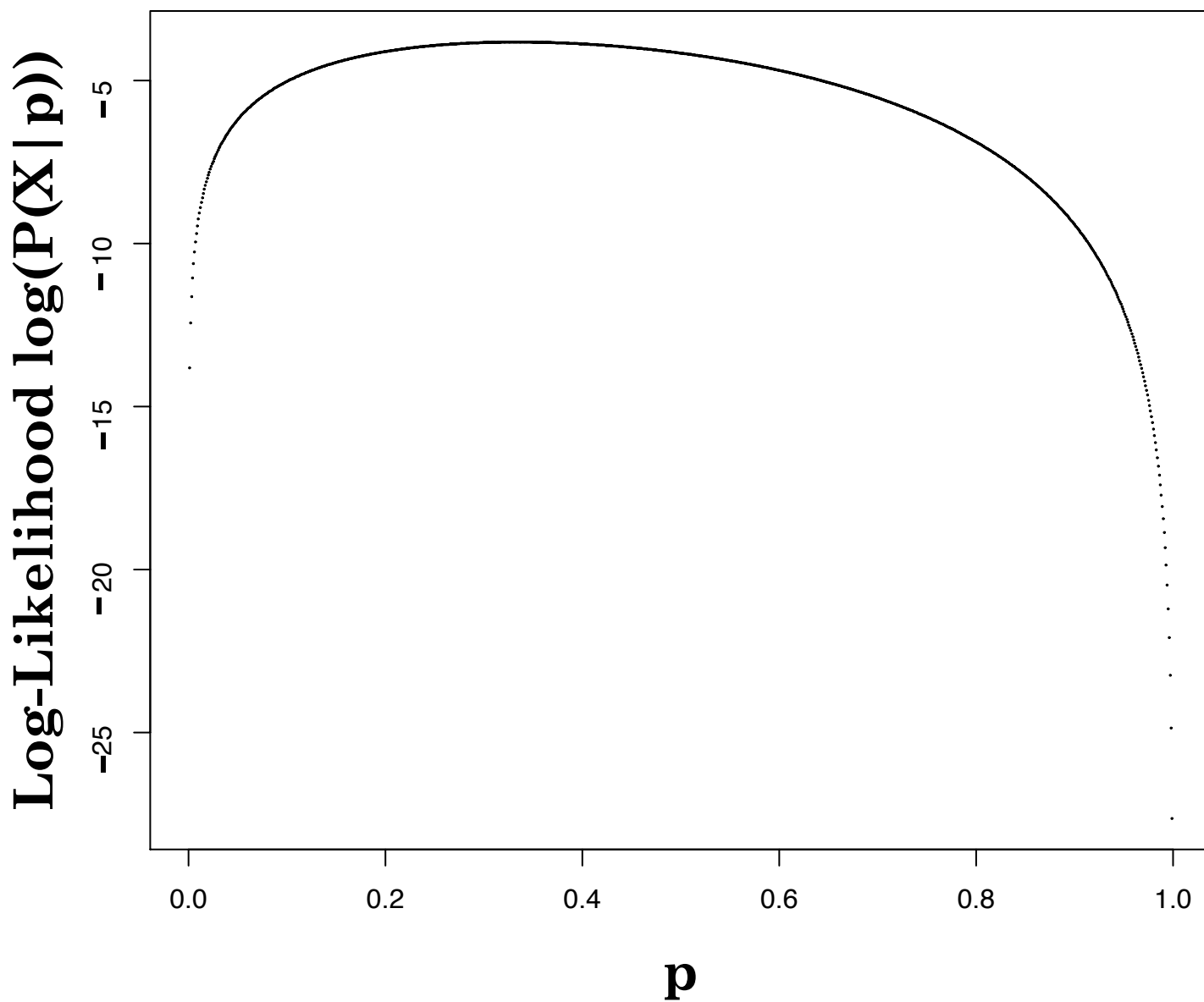
$$P(X | p) = p^2(1-p)^4$$

 **almost binomial distribution form**

# Likelihood with 2 heads and 4 tails



# Log-Likelihood with 2 heads and 4 tails





**For integers a and b, Beta density B(a,b) is**

$$P(p) = \frac{(a+b-1)!}{((a-1)!(b-1)!)} p^{a-1}(1-p)^{b-1}$$

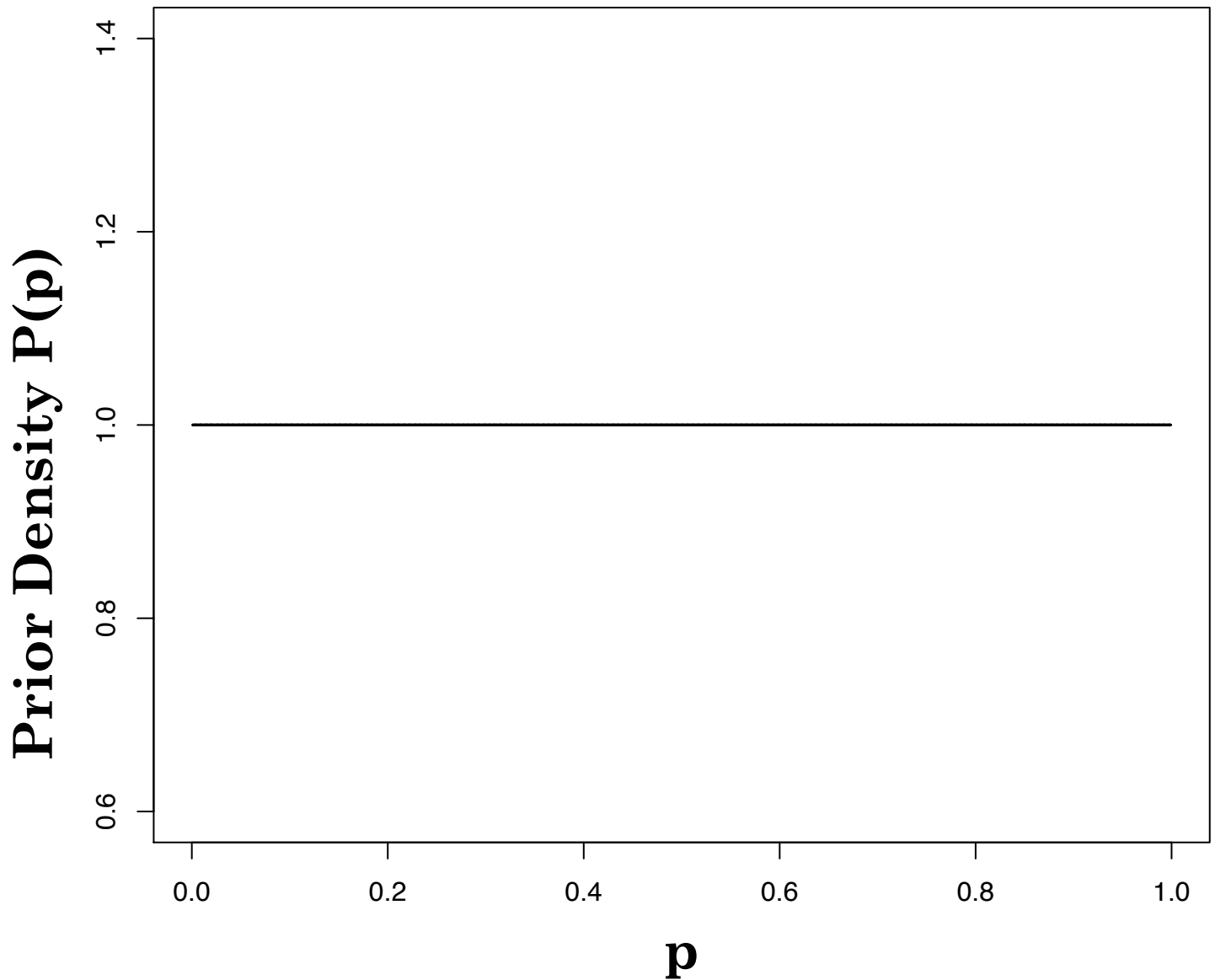
**where p is between 0 and 1.**

**Expected value of p is  $a/(a+b)$**

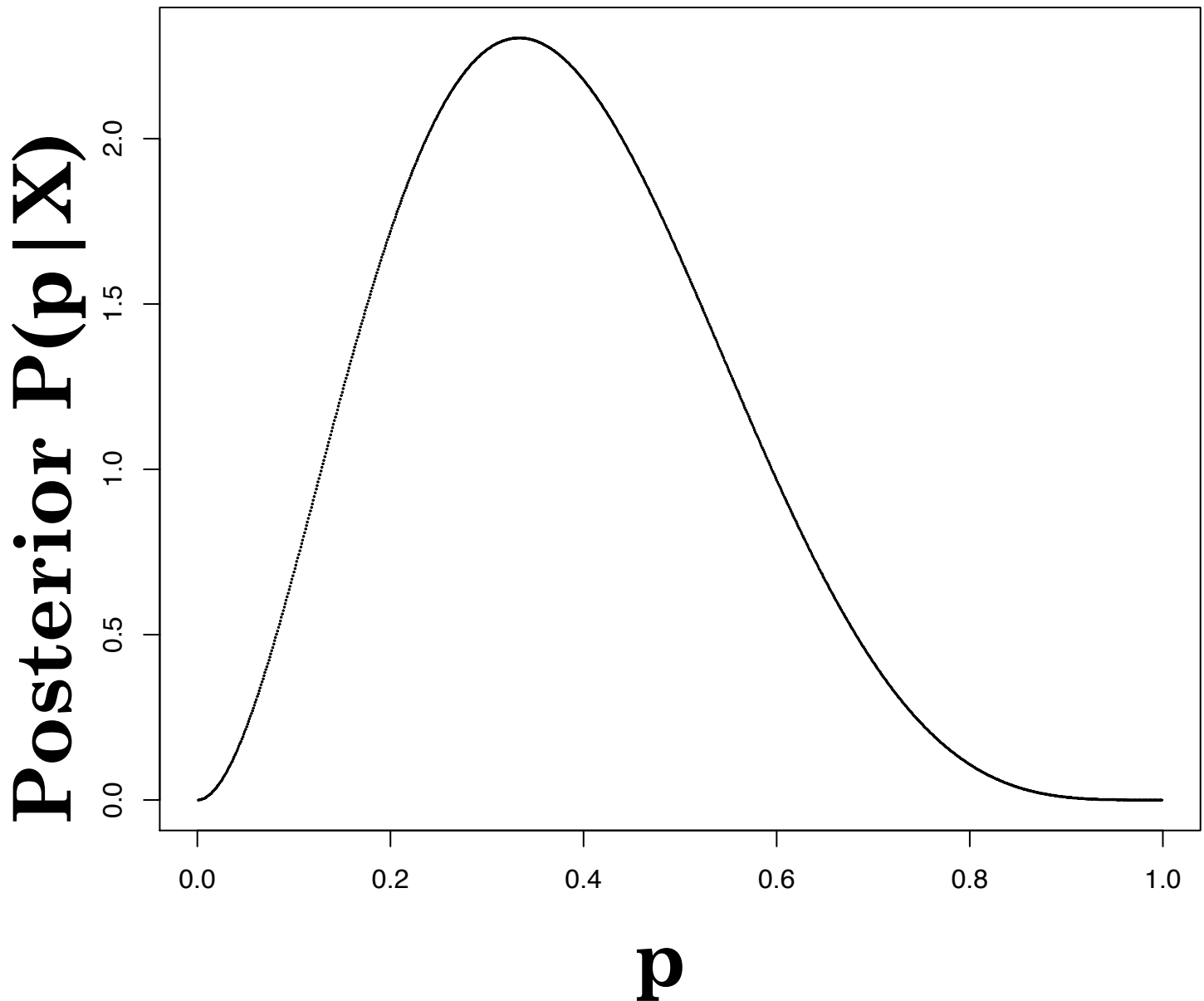
**Variance of p is  $ab/((a+b+1)(a+b)^2)$**

- Beta distribution is conjugate prior for**
- data from binomial distribution**

# Uniform Prior Distribution (i.e., Beta(1,1) distribution)



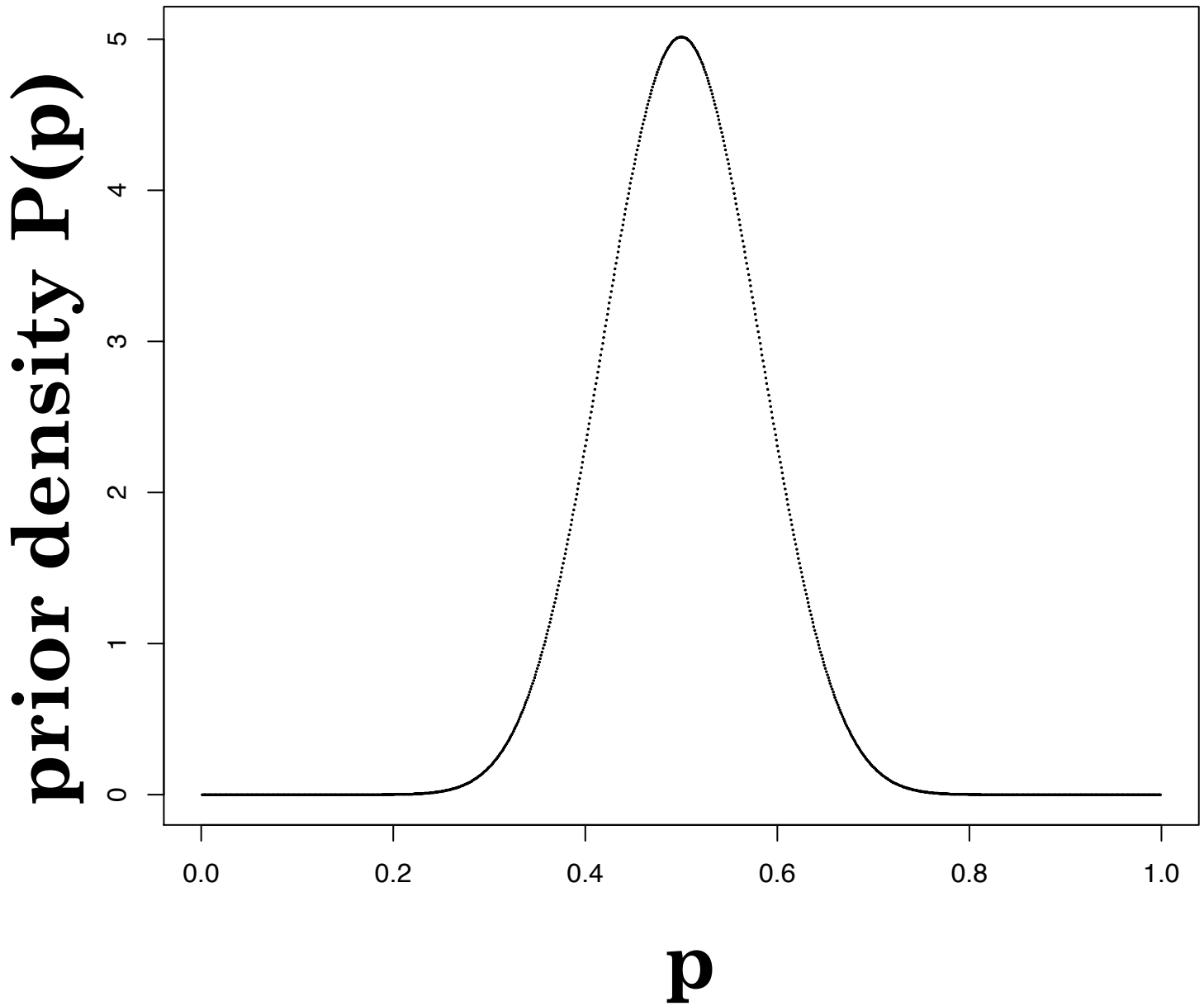
# Beta(3,5) posterior from Uniform prior + data (2 heads and 4 tails)



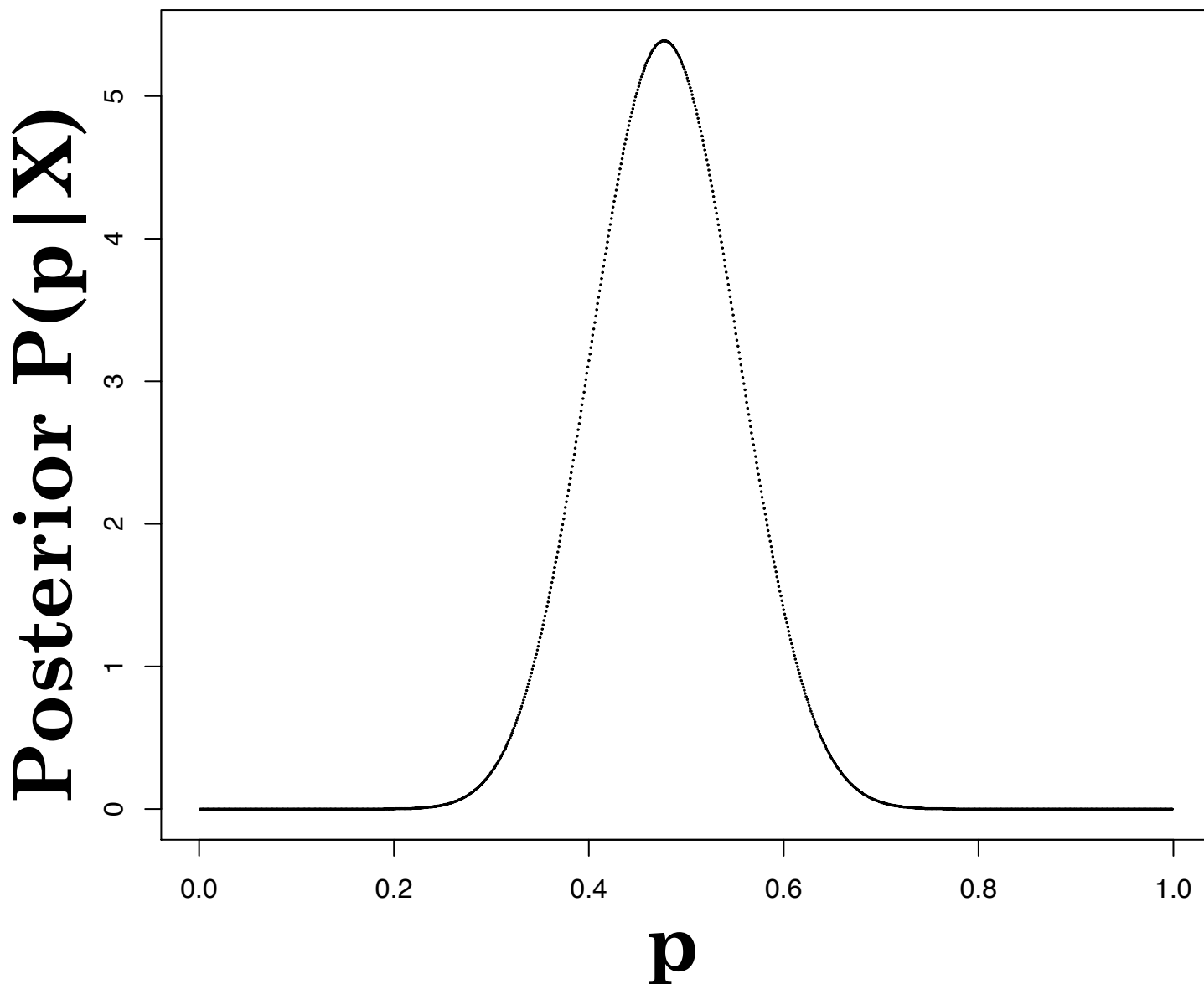
**Posterior Mean =  $3/(3+5)$**

# Beta(20,20) prior distribution

Prior Mean = 0.5

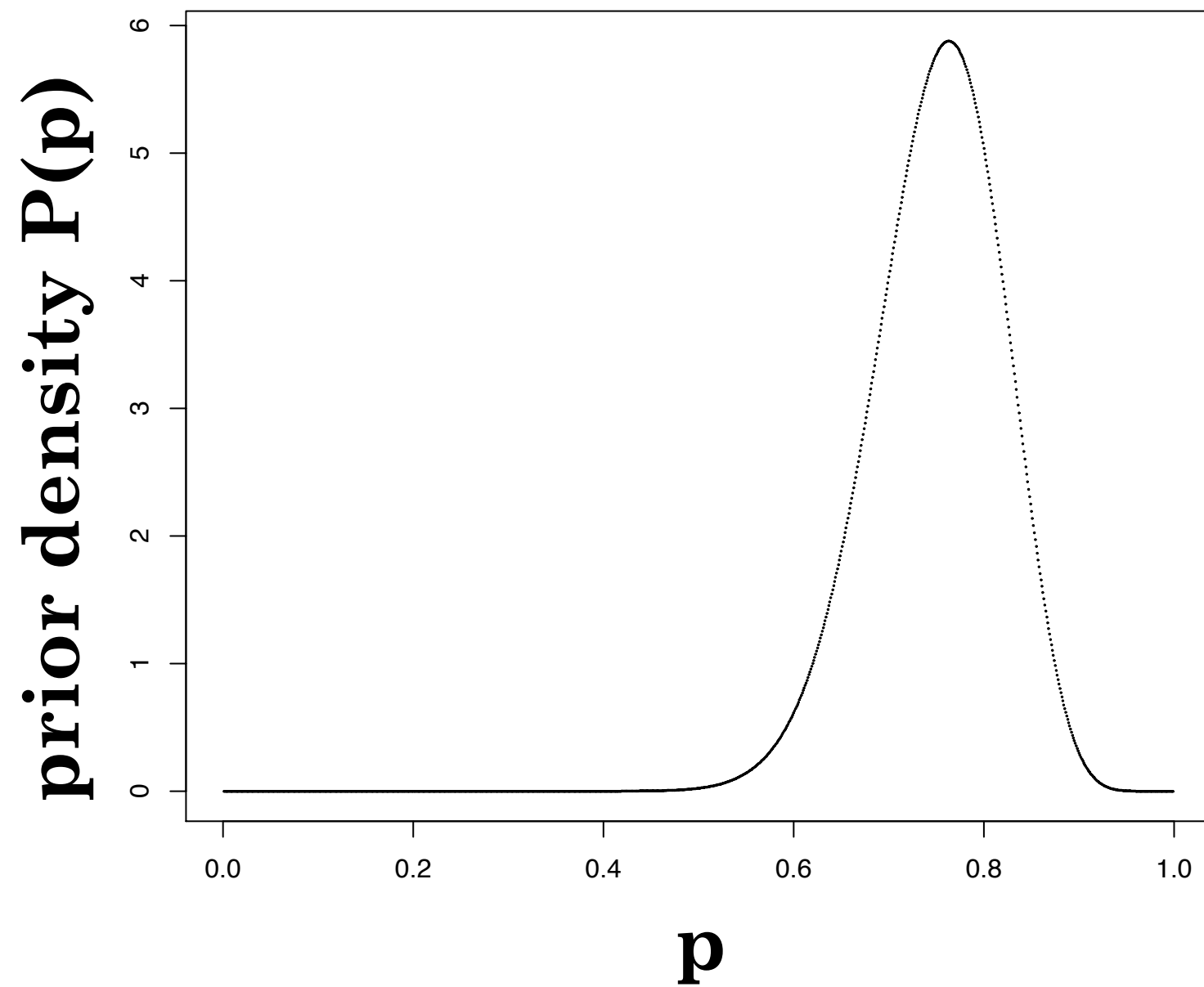


**Beta(22,24) posterior from  
Beta(20,20) prior + data (2  
heads and 4 tails)**

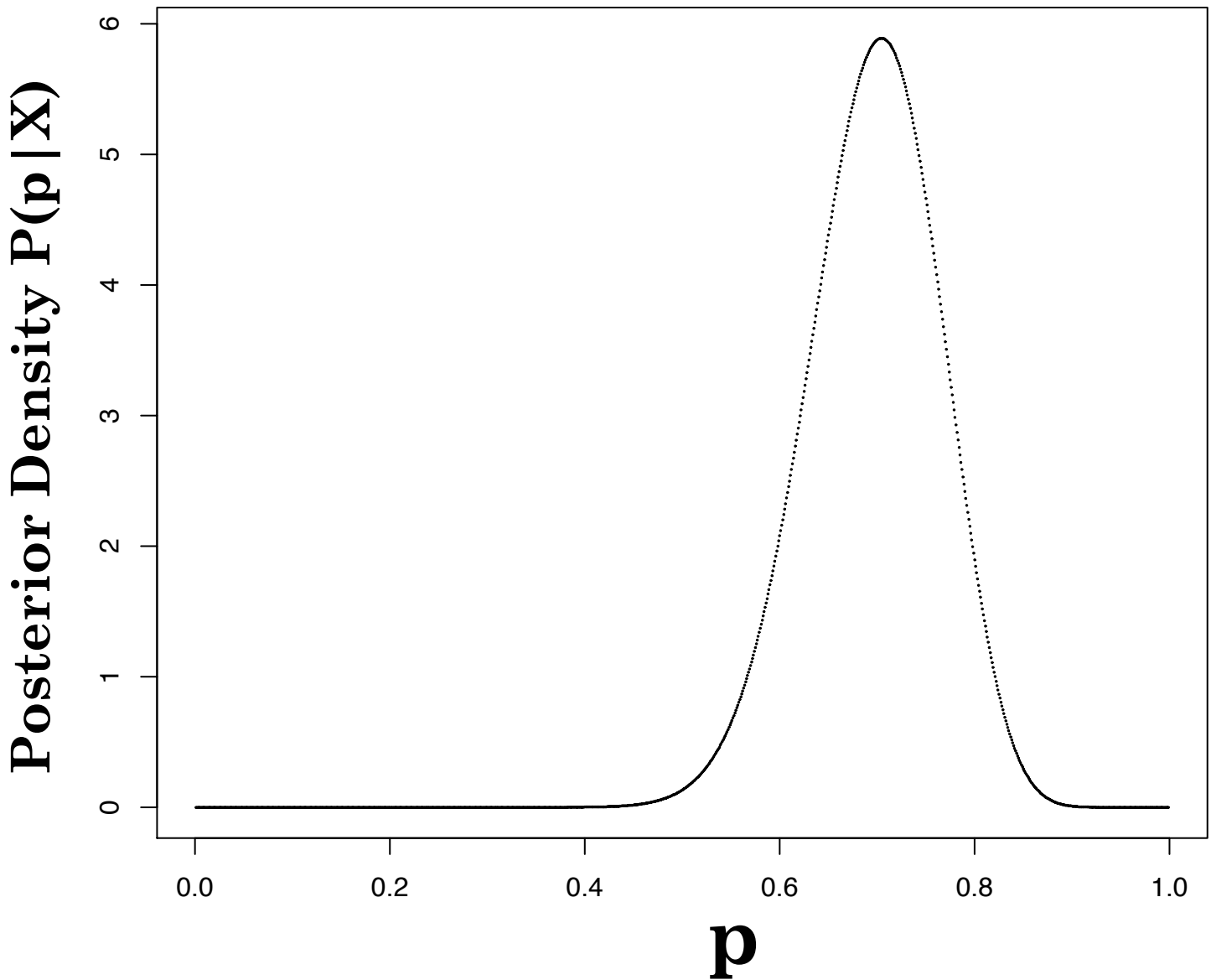


# Beta(30,10) prior distribution

Prior Mean = 0.75

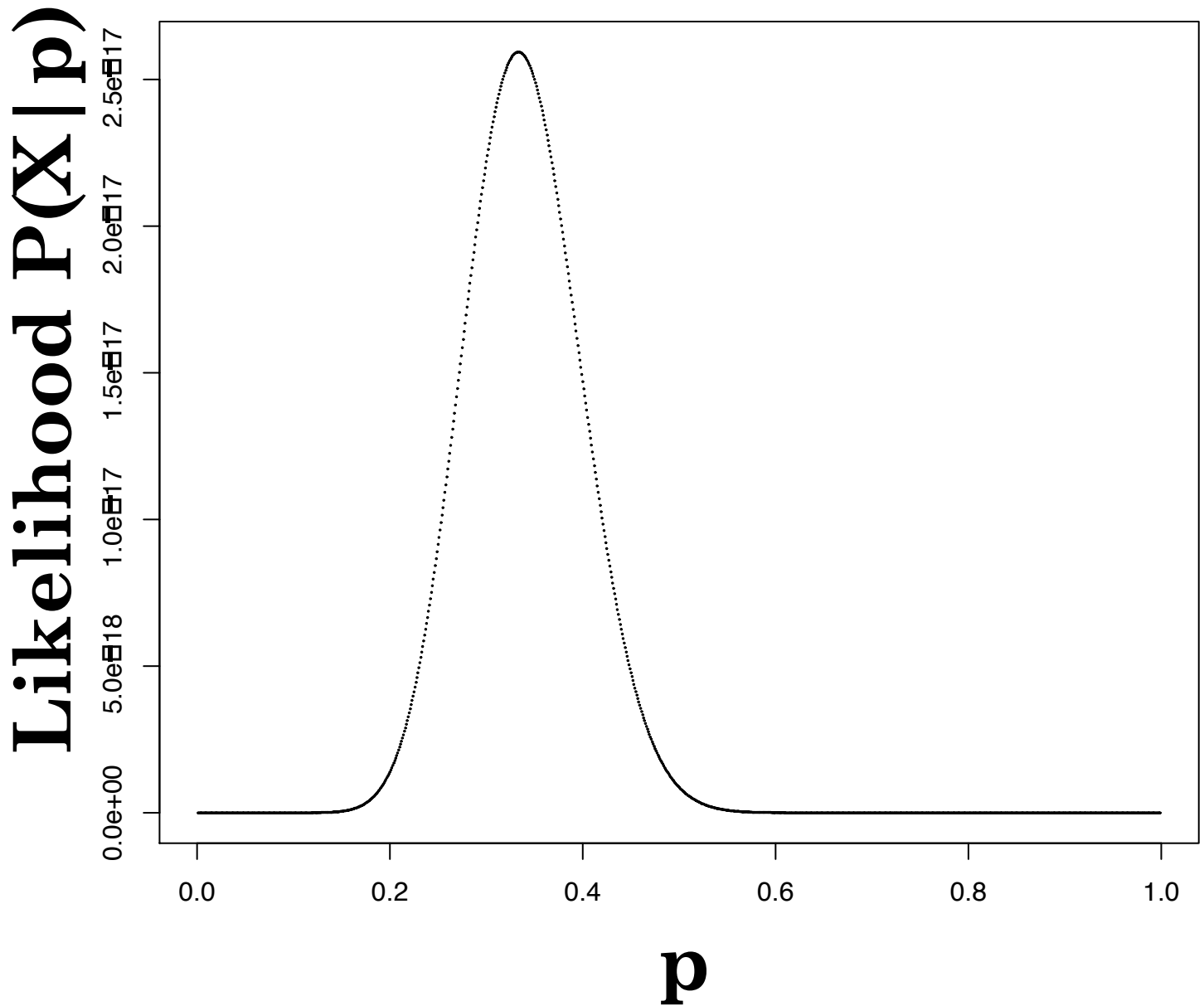


**Beta(32,14) posterior from  
Beta(30,10) prior + data (2  
heads and 4 tails)**



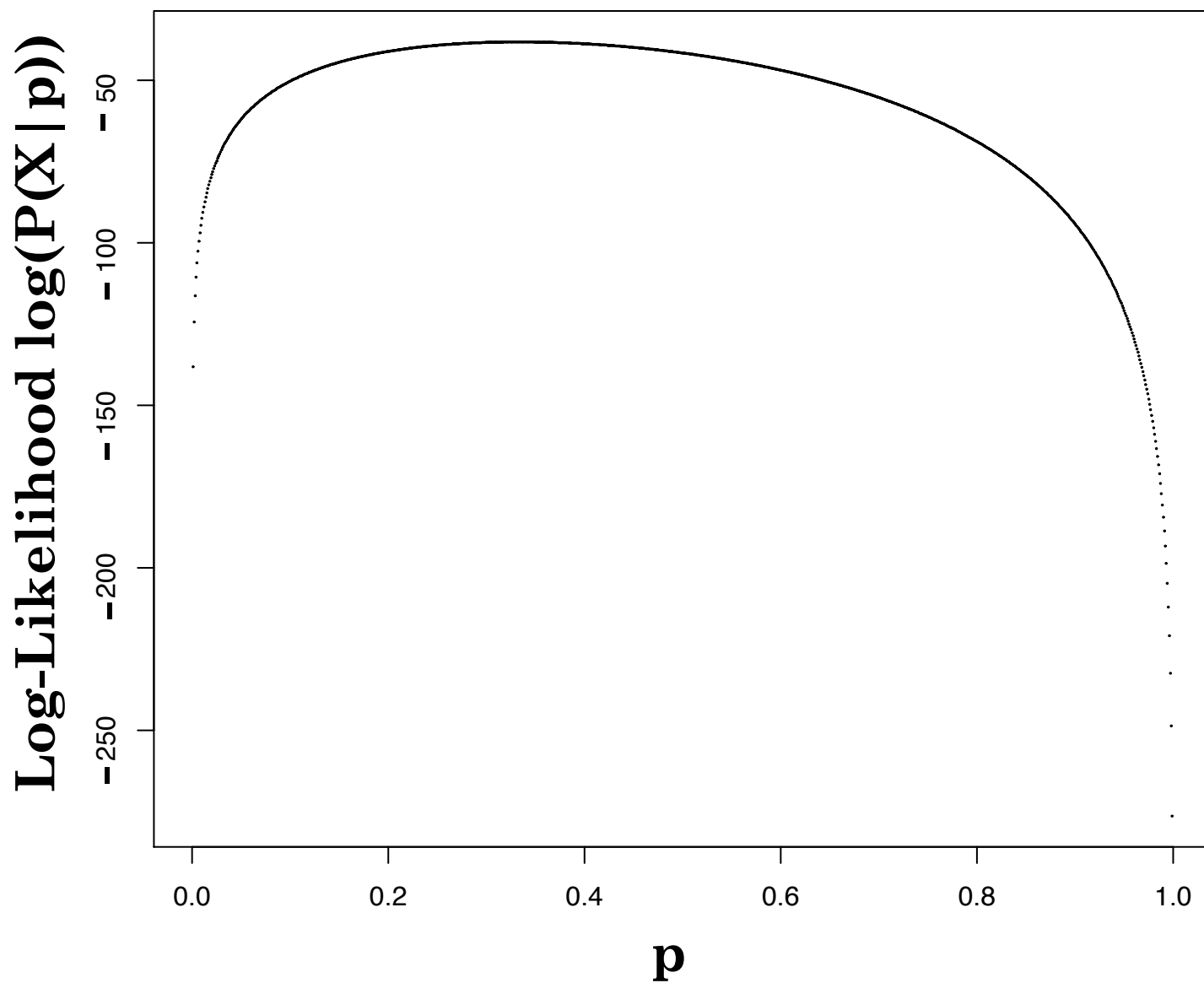
**Posterior Mean =  $32/(32+14)$**

# Likelihood with 20 Heads and 40 Tails

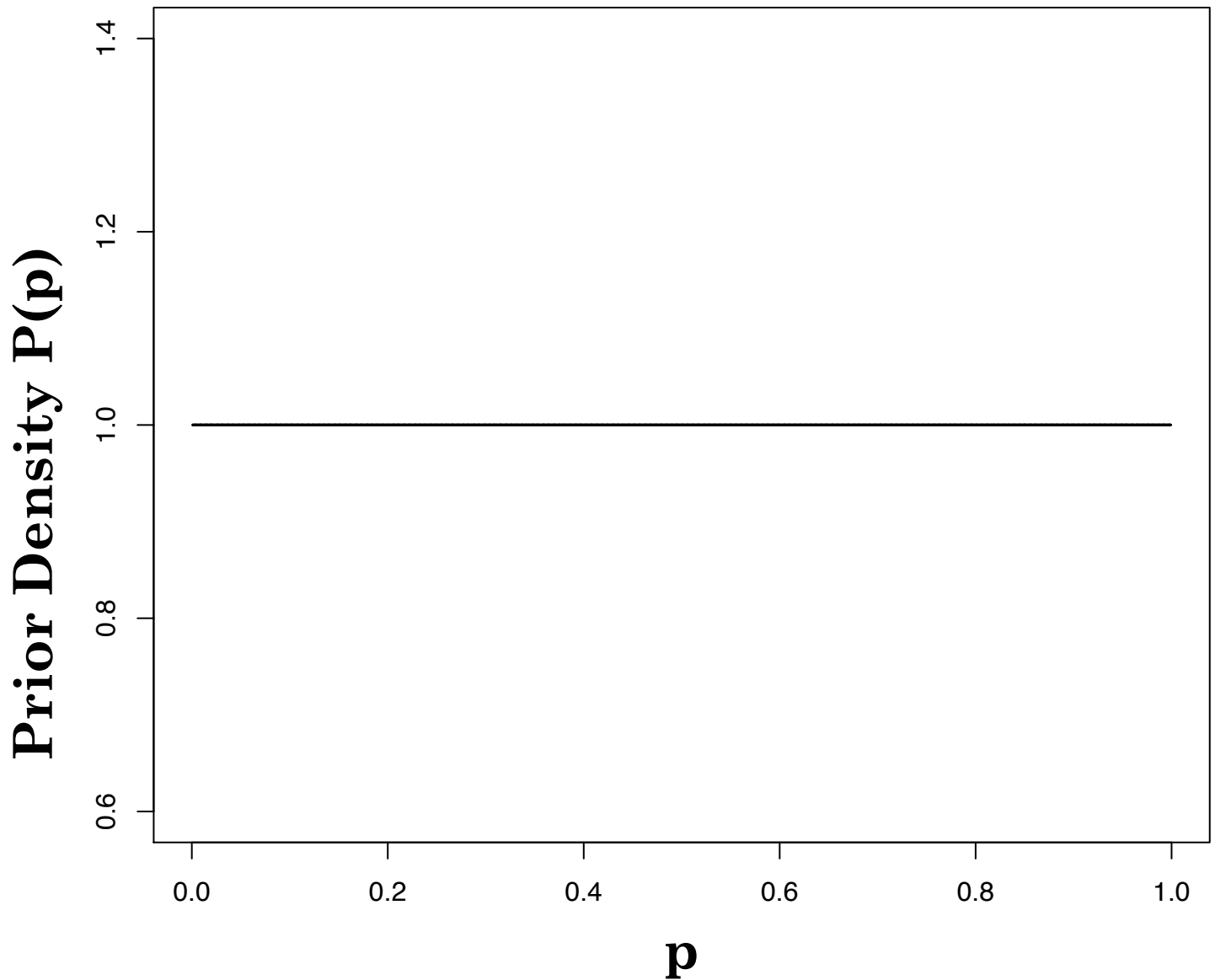




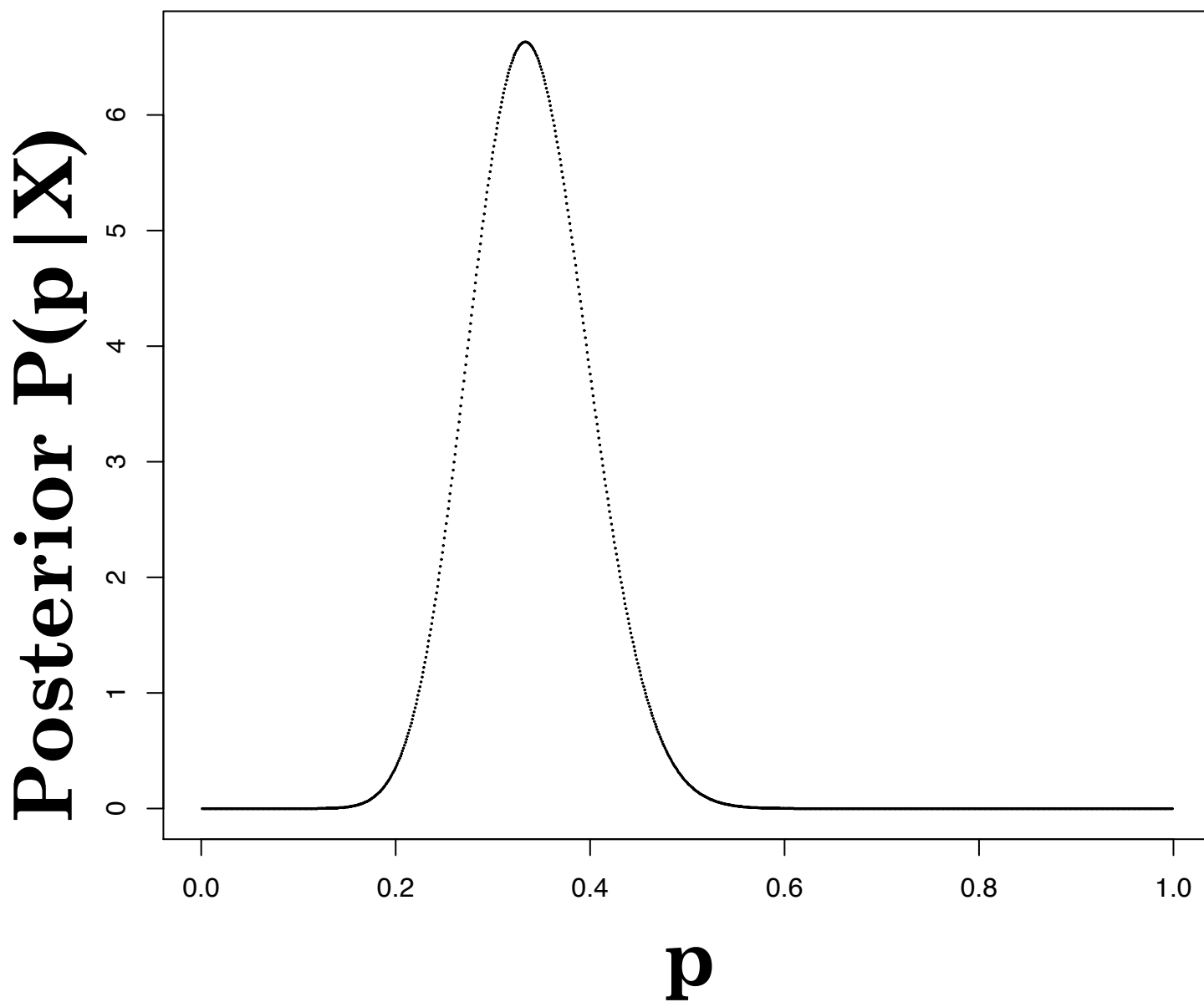
# Log-Likelihood with 20 heads and 40 tails



# Uniform Prior Distribution (i.e., Beta(1,1) distribution)

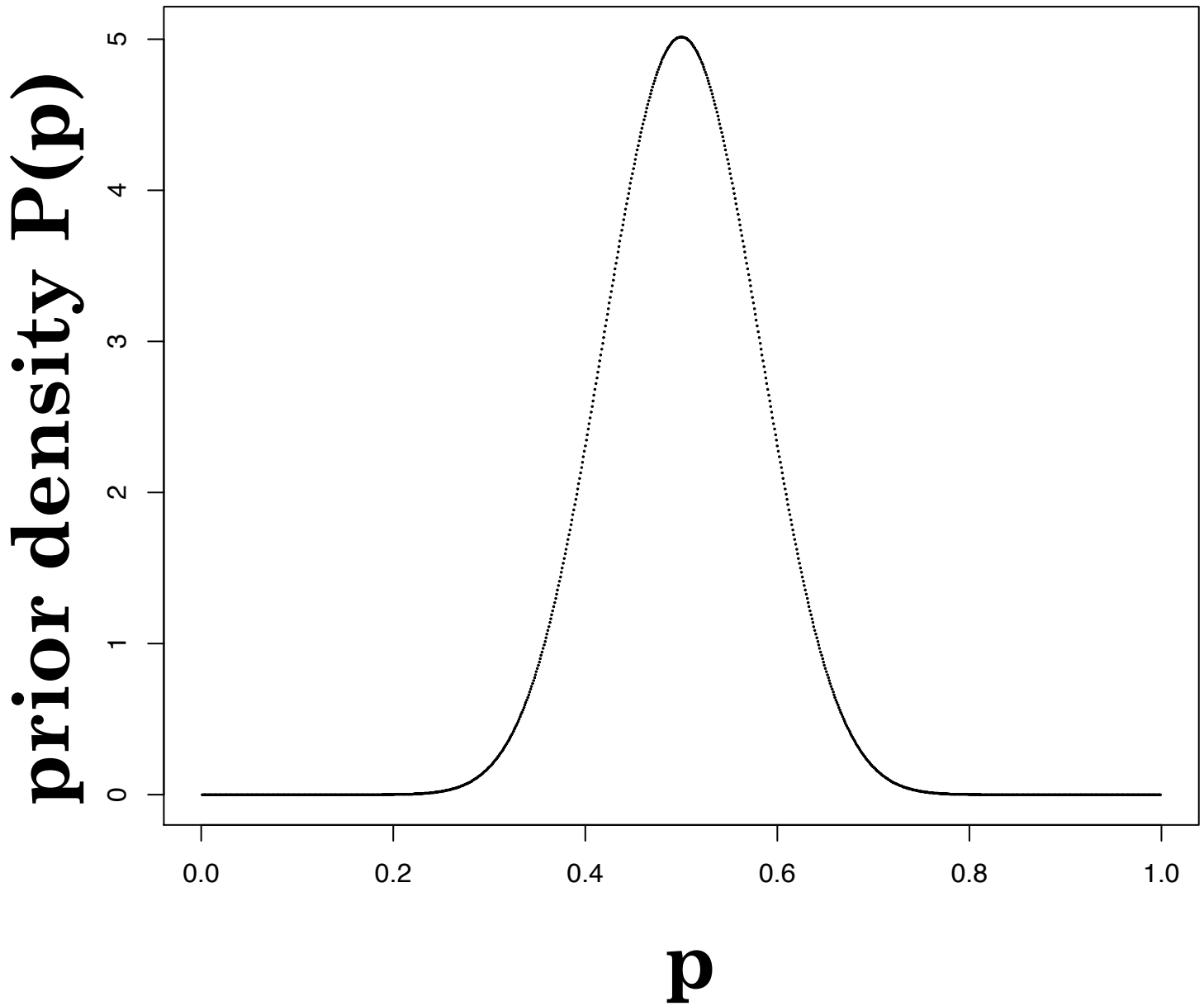


# Beta(21,41) posterior from Uniform prior + data (20 heads and 40 tails)

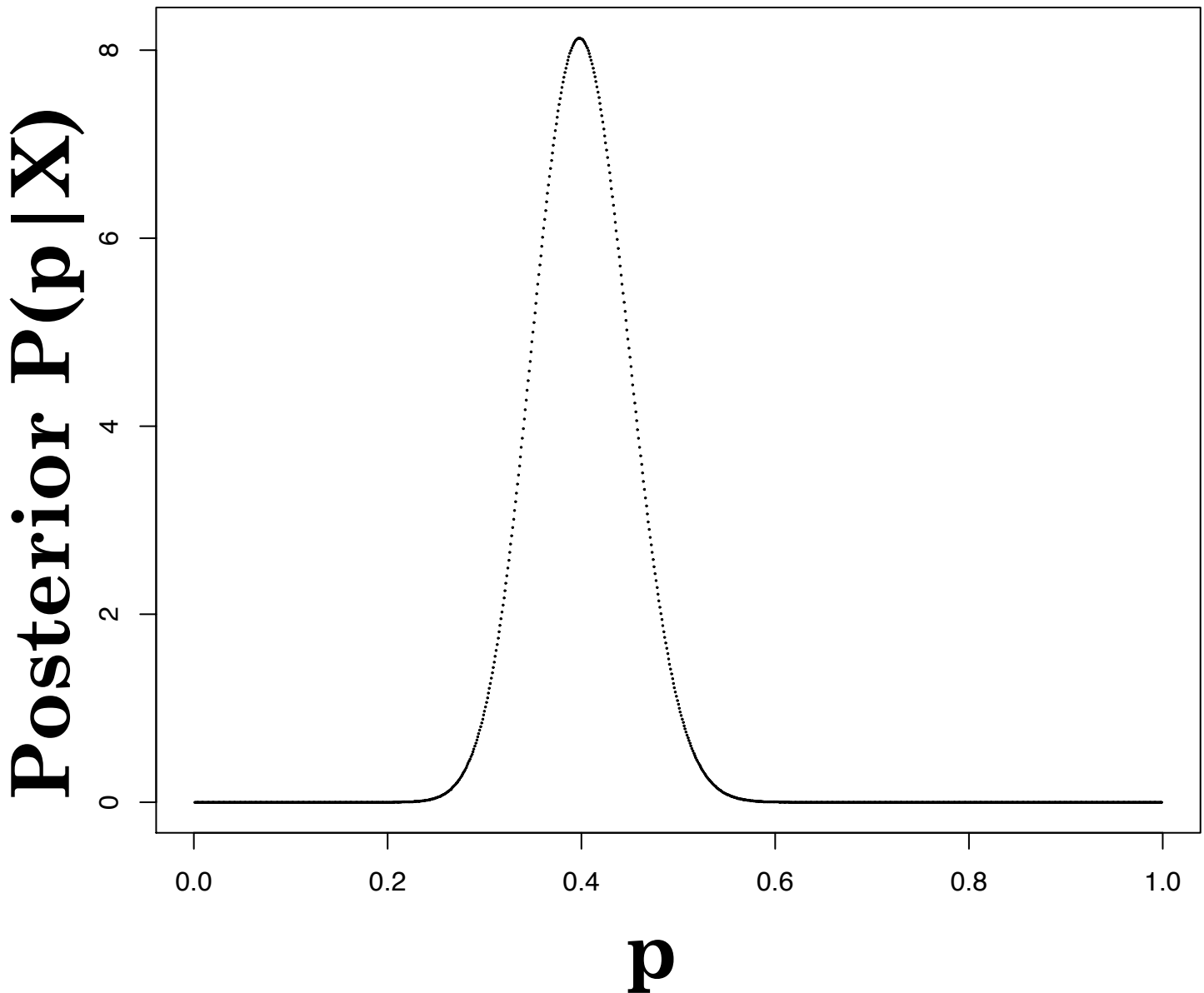


# Beta(20,20) prior distribution

Prior Mean = 0.5

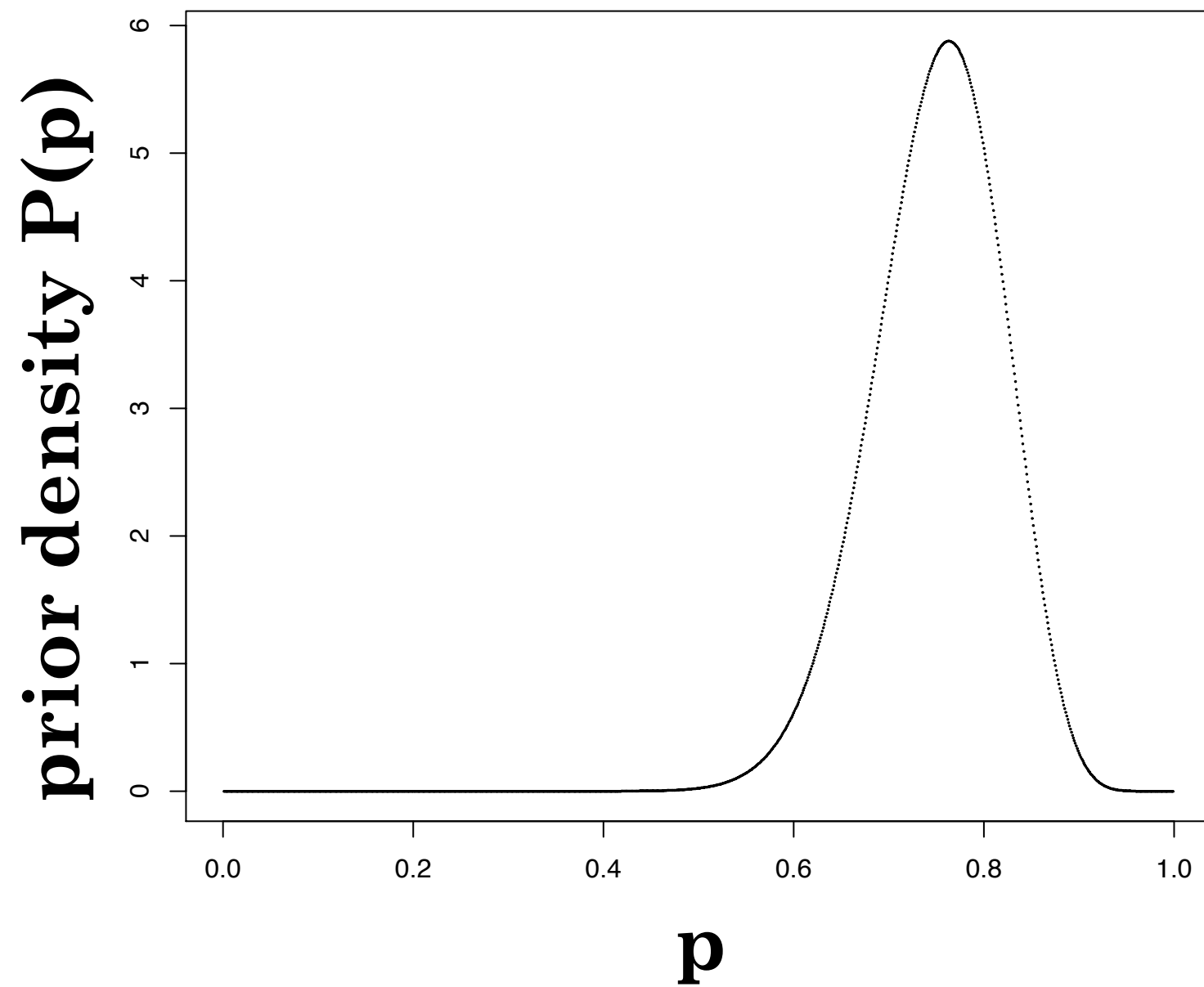


**Beta(40,60) posterior from  
Beta(20,20) prior + data (20  
heads and 40 tails)**

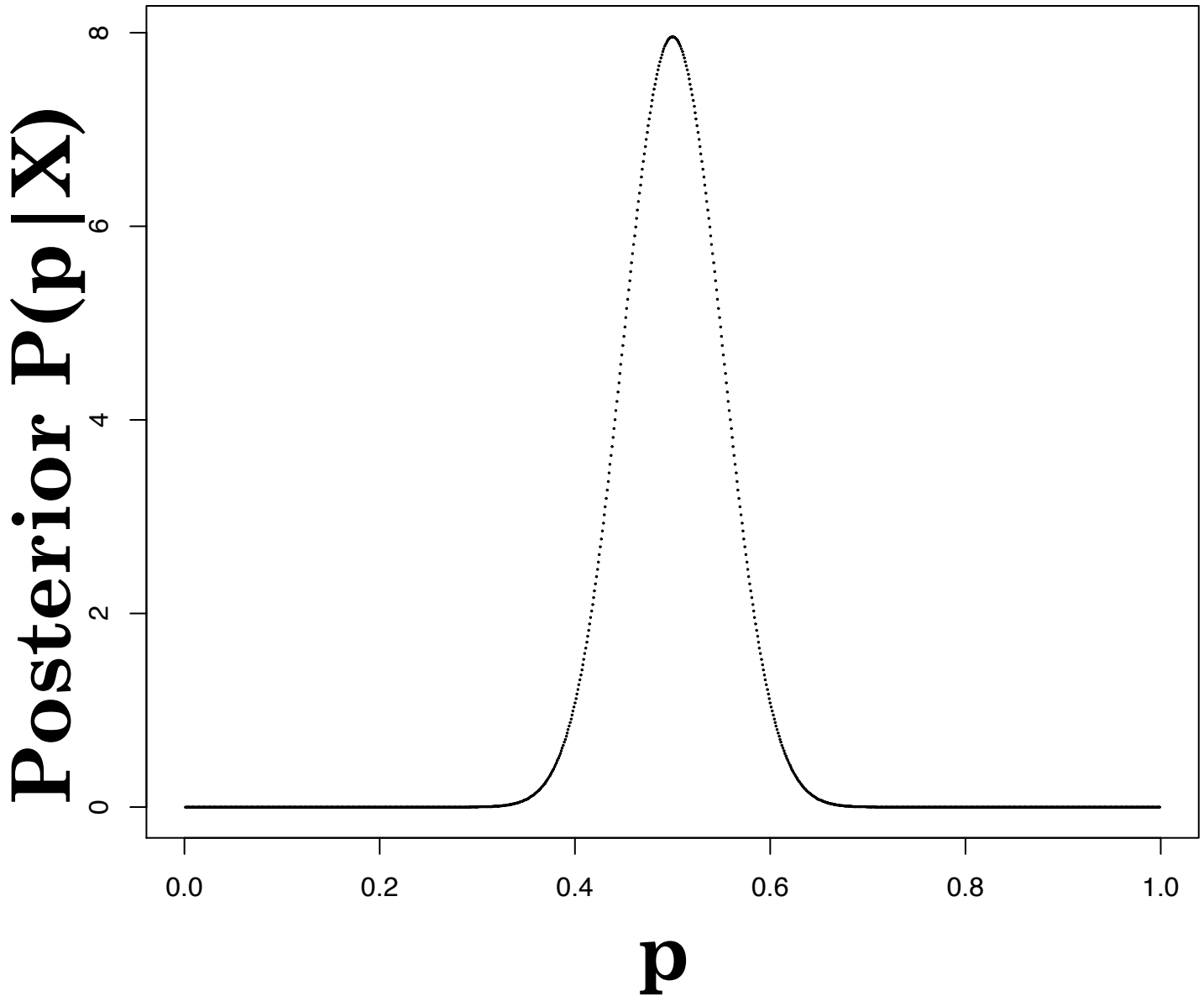


# Beta(30,10) prior distribution

Prior Mean = 0.75



**Beta(50,50) posterior from  
Beta(30,10) prior + data (20  
heads and 40 tails)**



Markov chain Monte Carlo (MCMC) idea approximates  $\Pr(\theta | X)$  by sampling a large number of  $\theta$  values from  $\Pr(\theta | X)$ .

So,  $\theta$  values with higher posterior probability are more likely to be sampled than  $\theta$  values with low posterior probability.



Question: How is this sampling achieved?

Answer: A Markov chain is constructed and simulated. The states of this chain represent values of  $\theta$ . The stationary distribution of this chain is  $\Pr(\theta | X)$ .

In other words, we start chain at some initial value of  $\theta$ . After running chain for a long enough time, the probability of the chain being at some particular state will be approximately equal to the posterior probability of the state.

Let  $\theta^{(t)}$  be the value of  $\theta$  after  $t$  steps of the Markov chain where  $\theta^{(0)}$  is the initial value.

Each step of the Markov chain involves randomly proposing a new value of  $\theta$  based on the current value of  $\theta$ . Call the proposed value  $\theta^*$ .

We decide with some probability to either accept  $\theta^*$  as our new state or to reject the proposed  $\theta^*$  and remain at our current state.

The Hastings (Hastings 1970) algorithm is a way to make this decision and force the stationary distribution of the chain to be  $\Pr(\theta | X)$ .

According to the Hastings algorithm, what state should we adopt at step  $t + 1$  if  $\theta^{(t)}$  is the current state and  $\theta^*$  is the proposed state?

Let  $J(\theta^*|\theta^{(t)})$  be the “jumping” distribution, i.e. the probability of proposing  $\theta^*$  given that the current state is  $\theta^{(t)}$ .

Define  $r$  as

$$r = \frac{\Pr(X | \theta^*)\Pr(\theta^*)J(\theta^{(t)}|\theta^*)}{\Pr(X | \theta^{(t)})\Pr(\theta^{(t)})J(\theta^*|\theta^{(t)})}$$

With probability equal to the minimum of  $r$  and 1, we set

$$\theta^{(t+1)} = \theta^*.$$

Otherwise, we set

$$\theta^{(t+1)} = \theta^{(t)}.$$

For the Hastings algorithm to yield the stationary distribution  $\Pr(\theta | X)$ , there are a few required conditions. The most important condition is that it must be possible to reach each state from any other in a finite number of steps. Also, the Markov chain can't be periodic.

## MCMC implementation details:

The Markov chain should be run as long as possible.

We may have  $T$  total samples after running our Markov chain. They would be  $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(T)}$ . The first  $B$  ( $1 \leq B < T$ ) of these samples are often discarded (i.e. not used to approximate the posterior). The period before the chain has gotten these  $B$  samples that will be discarded is referred to as the “burn-in” period.

The reason for discarding these samples is that the early samples typically are largely dependent on the initial state of the Markov chain and often the initial state of the chain is (either intentionally or unintentionally) atypical with respect to the posterior distribution.

The remaining samples  $\theta^{(B+1)}, \theta^{(B+2)}, \dots, \theta^{(T)}$  are used to approximate the posterior distribution. For example, the average among the sampled values for a parameter might be a good estimate of its posterior mean.

## **Markov Chain Monte Carlo and Relatives (some important papers)**

CARLIN, B.P., and T.A. LOUIS. 1996. Bayes and Empirical Bayes Methods for Data Analysis. Chapman and Hall, London.

GELMAN, A., J.B. CARLIN, H.S. STERN, and D.B. RUBIN. 1995. Bayesian Data Analysis. Chapman and Hall, London.

GEYER, C. 1991. Markov chain Monte Carlo maximum likelihood. Pages 156-163 in Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface. Keramidas, ed. Fairfax Station: Interface Foundation

HASTINGS, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109

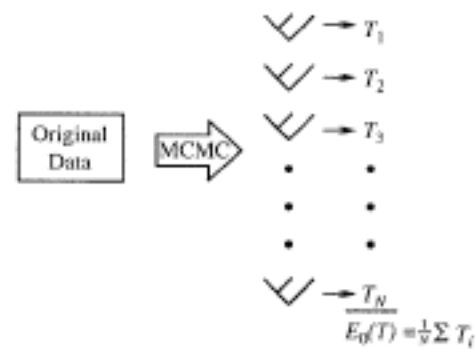
METROPOLIS, N., A.W. ROSENBLUTH, M.N. ROSENBLUTH, A.H. TELLER, and E. TELLER. 1953. Equations of state calculations by fast computing machines. *J. Chem. Phys.* 21: 1087–1092.

**Posterior predictive inference** (notice resemblance to parametric bootstrap)

1. Via MCMC or some other technique, get  $N$  sampled parameter sets  $\theta^{(1)}, \dots, \theta^{(N)}$  from posterior distribution  $p(\theta|X)$
2. For each sampled parameter set  $\theta^{(k)}$ , simulate a new data set  $X^{(k)}$  from  $p(X|\theta^{(k)})$
3. Calculate a test statistic value  $T(X^{(k)})$  from each simulated data set and see where test statistic value for actual data  $T(X)$  is relative to simulated distribution of test statistic values.

From Huelsenbeck et al.  
2003. Syst Biol  
52(2): 131-158

(A) Calculating original value for test statistic



(B) Calculating predicted values for test statistic

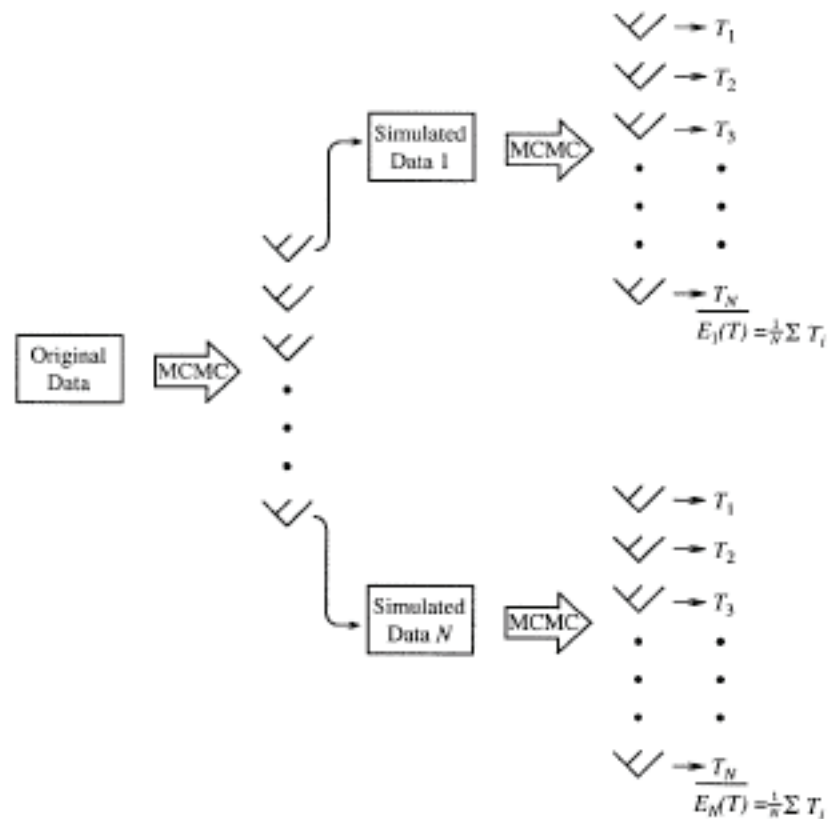


FIGURE 4. An example of how posterior predictive  $P$  values are calculated for a test statistic,  $T$ . The observed value for the test statistic is calculated by averaging over the posterior probability distribution of parameters. We use MCMC to draw parameter values from the posterior probability distribution of parameters. The predictive distribution is calculated by simulating new data using parameter values from the posterior probability distribution of parameters. Each simulated data set is treated exactly as was the original data. The predictive  $P$  value is the proportion of the test statistics from the simulated data that exceed the observed value.

Notation for following pages:

$X$  data

$M_i, M_j$ : Models  $i$  and  $j$

$\theta_i, \theta_j$ : parameters for models  $i$  and  $j$

$p(X|\theta_i, M_i), p(X|\theta_j, M_j)$ : likelihoods



## Bayes factor

$$\begin{aligned}\frac{p(M_i|X)}{p(M_j|X)} &= \frac{p(M_i)p(X|M_i)/p(X)}{p(M_j)p(X|M_j)/p(X)} \\ &= \frac{p(M_i)}{p(M_j)} \times \frac{p(X|M_i)}{p(X|M_j)}\end{aligned}$$

Left factor is called *prior odds* and right factor is called *Bayes factor*.

Bayes factor is ratio of *marginal likelihoods* of the two models.

$$BF_{ij} = \frac{p(X|M_i)}{p(X|M_j)}$$

According to wikipedia, Jeffreys (1961) interpretation of  $BF_{12}$  (1 representing one model and 2 being the other):

$BF_{12}$	Interpretation
$< 1 : 1$	Negative (supports $M_2$ )
$1 : 1$ to $3 : 1$	Barely worth mentioning
$3 : 1$ to $10 : 1$	Substantial
$10 : 1$ to $30 : 1$	Strong
$30 : 1$ to $100 : 1$	Very Strong
$> 100 : 1$	Decisive

$$BF_{ij} = \frac{p(X|M_i)}{p(X|M_j)}$$

Bayes factors hard to compute because marginal likelihoods hard to compute:

$$p(X|M_i) = \int_{\theta_i} p(X|M_i, \theta_i)p(\theta_i|M_i)d\theta_i$$

Important point to note from above: Bayes factors depend on priors  $p(\theta_i|M_i)$  because marginal likelihoods depend on priors!

**How to approximate/compute marginal likelihood?**

$$p(X|M_i) = \int_{\theta_i} p(X|M_i, \theta_i)p(\theta_i|M_i)d\theta_i$$

**Harmonic mean estimator of marginal likelihood (widely used but likely to be terrible and should be avoided):**

$$\frac{1}{p(X|M_i)} \doteq \frac{1}{N} \sum_{k=1}^N \frac{1}{p(X|\theta_i^{(k)}, M_i)}$$

where  $\theta_i^{(k)}$  are sampled from posterior  $p(\theta_i|X, M_i)$ .

## Important papers regarding Bayesian Model Comparison ...

Posterior Predictive Inference in Phylogenetics: J.P. Bollback. 2002. *Molecular Biology and Evolution*. 19:1171-1180

Harmonic Mean and other techniques for estimating Bayes factors: Newton and Raftery. 1994. *Journal of the Royal Statistical Society. Series B*. 56(1):3-48.

Thermodynamic Integration to Approximate Bayes Factors (adapted to molecular evolution data): Lartillot and Philippe. 2006. *Syst. Biol.* 55:195-207

Improving marginal likelihood estimation for Bayesian phylogenetic model selection. W. Xie, P.O. Lewis, Y. Fan, L. Kao, M-H Chen. 2011. *Syst Biol.* 60(2):150-160.

Choosing among partition models in Bayesian phylogenetics. Y. Fan, R. Wu, M-H Chen, L Kuo, P.O. Lewis. 2011. *Mol. Biol. Evol.* 28(1):523-532.

Markov chain Monte Carlo without likelihoods. P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. 2003. *PNAS USA*. 100(26): 15324-15328.

H. Jeffreys. *The Theory of Probability* (3e). Oxford (1961); p. 432

M.A. Beaumont, W. Zhang, D.J. Balding. Approximate Bayesian Computation in Population Genetics. 2002. *Genetics* 162:2025-2035.

more  
reliable  
ways to  
approximate  
marginal  
likelihood



# Paul Lewis' MCMC Robot Demo

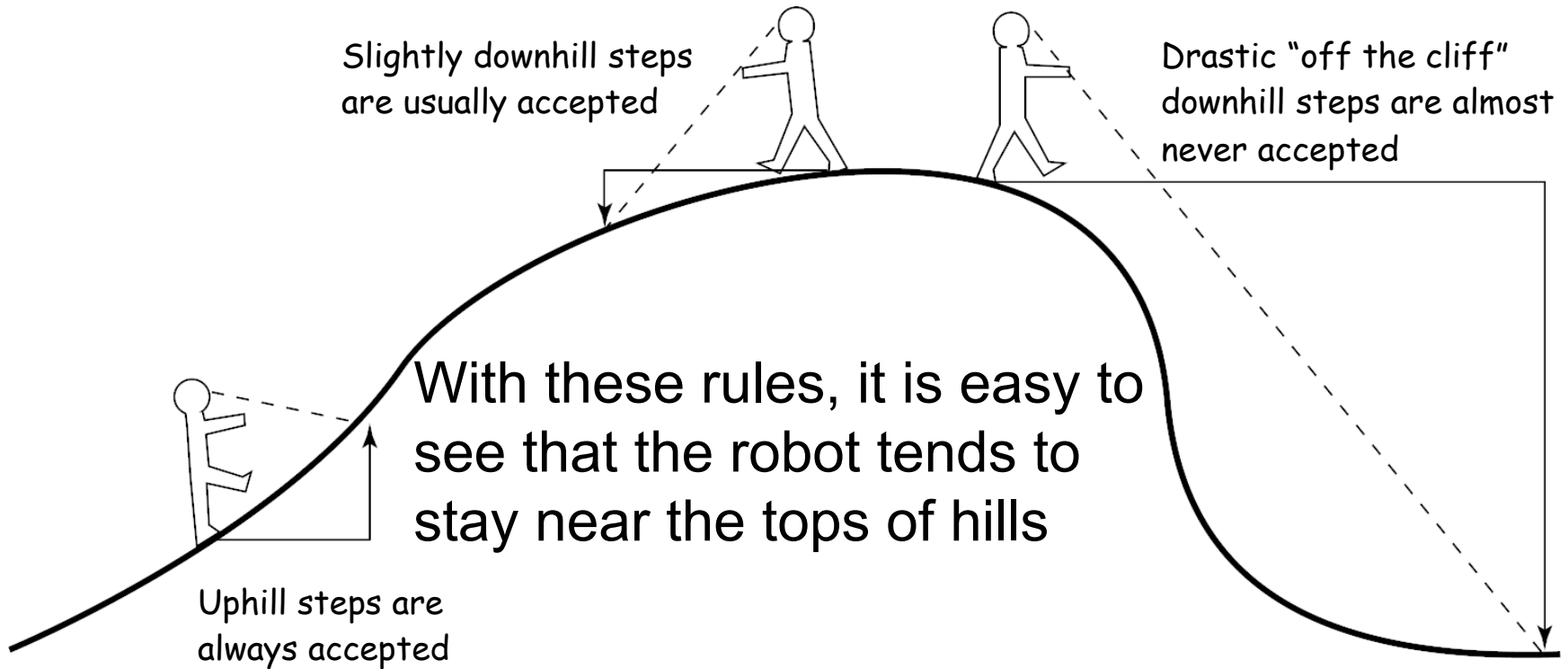
## Target distribution:

- Mixture of bivariate normal "hills"
- inner contours: 50% of the probability
- outer contours: 95%

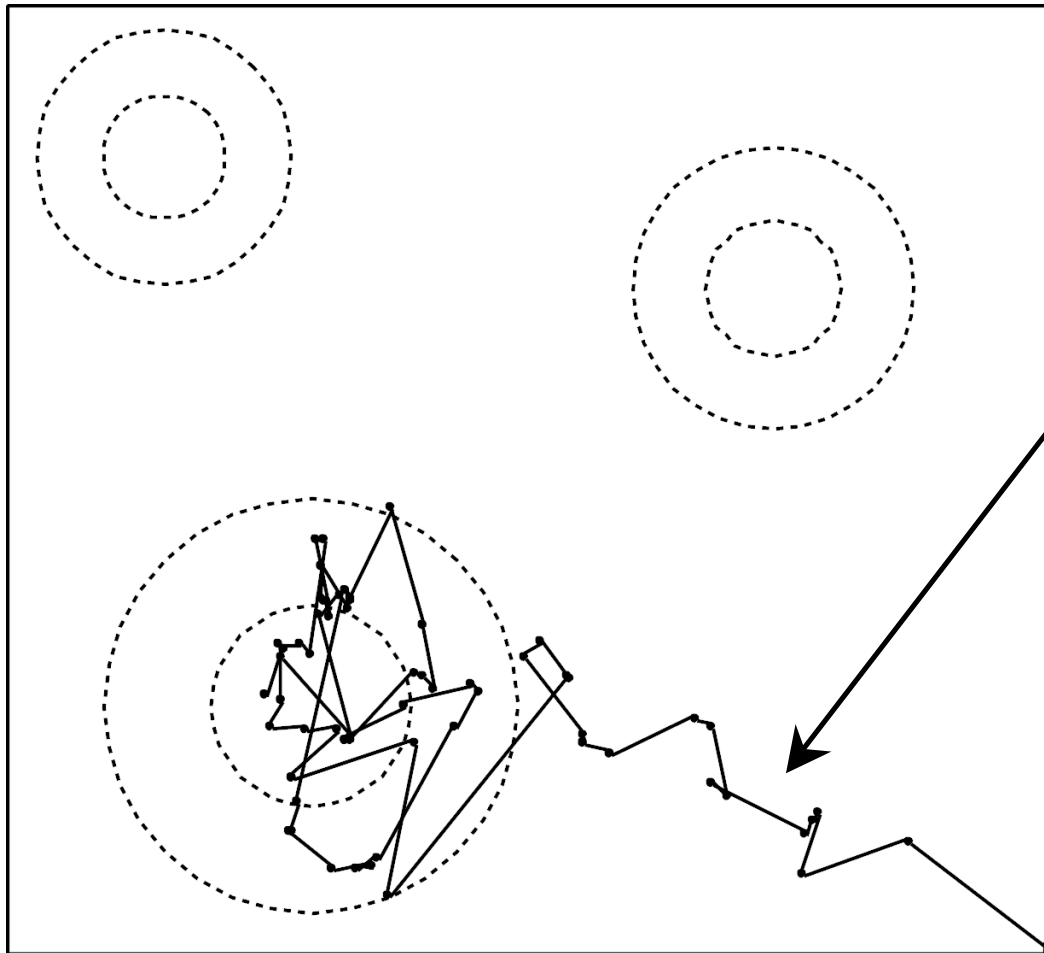
## Proposal scheme:

- random direction
- gamma-distributed step length  
(mean 45 pixels, s.d. 40 pixels)
- reflection at edges

# MCMC robot rules



# Burn-in



First 100 steps

Note that first few steps are not at all representative of the distribution.

Starting point



## Problems with MCMC approaches:

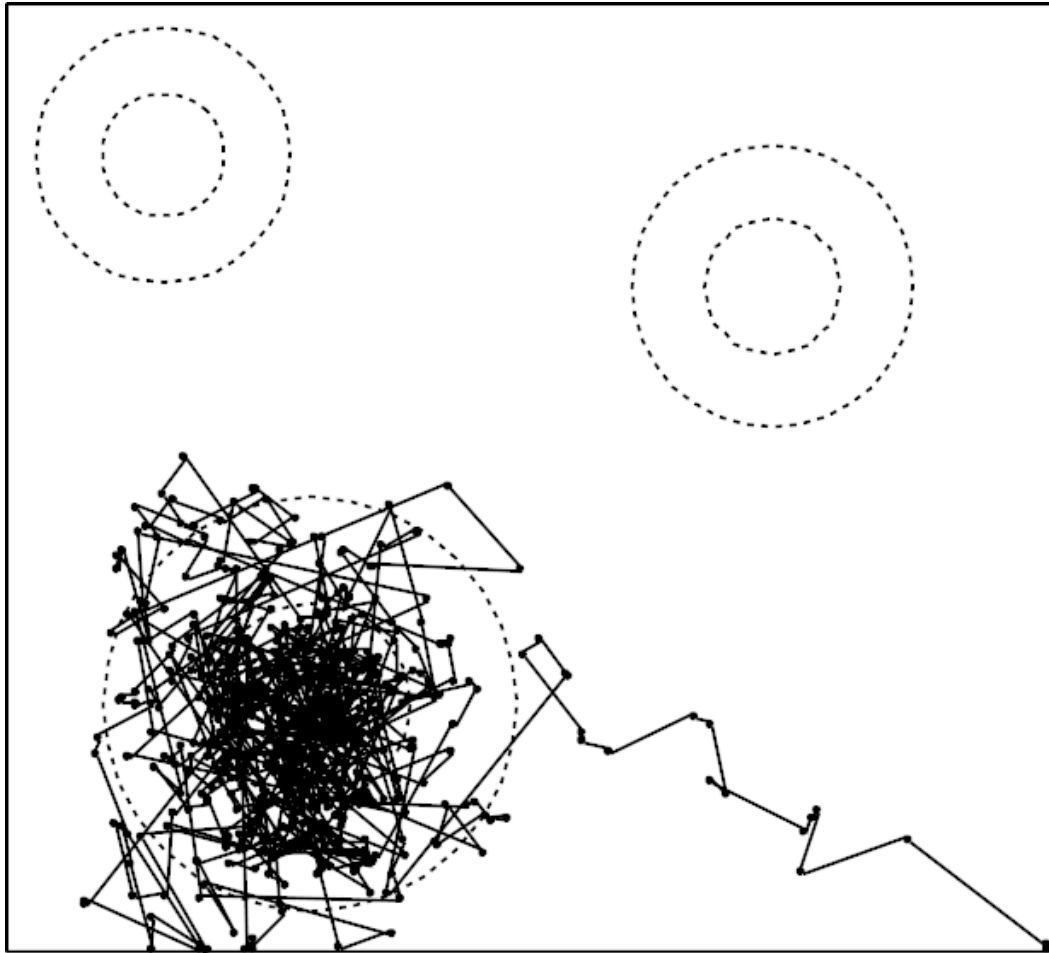
1. They are difficult to implement. Implementation may need to be clever to be computationally tractable and programming bugs are a serious possibility.

2. For the kinds of complicated situations that biologists face, it may be very difficult to know how fast the Markov chain converges to the desired posterior distribution.

There are diagnostics for evaluating whether a chain has converged to the posterior distribution but the diagnostics do not provide a guarantee of convergence.

**A GOOD DIAGNOSTIC : MULTIPLE RUNS !!**

# Just how long is a long run?

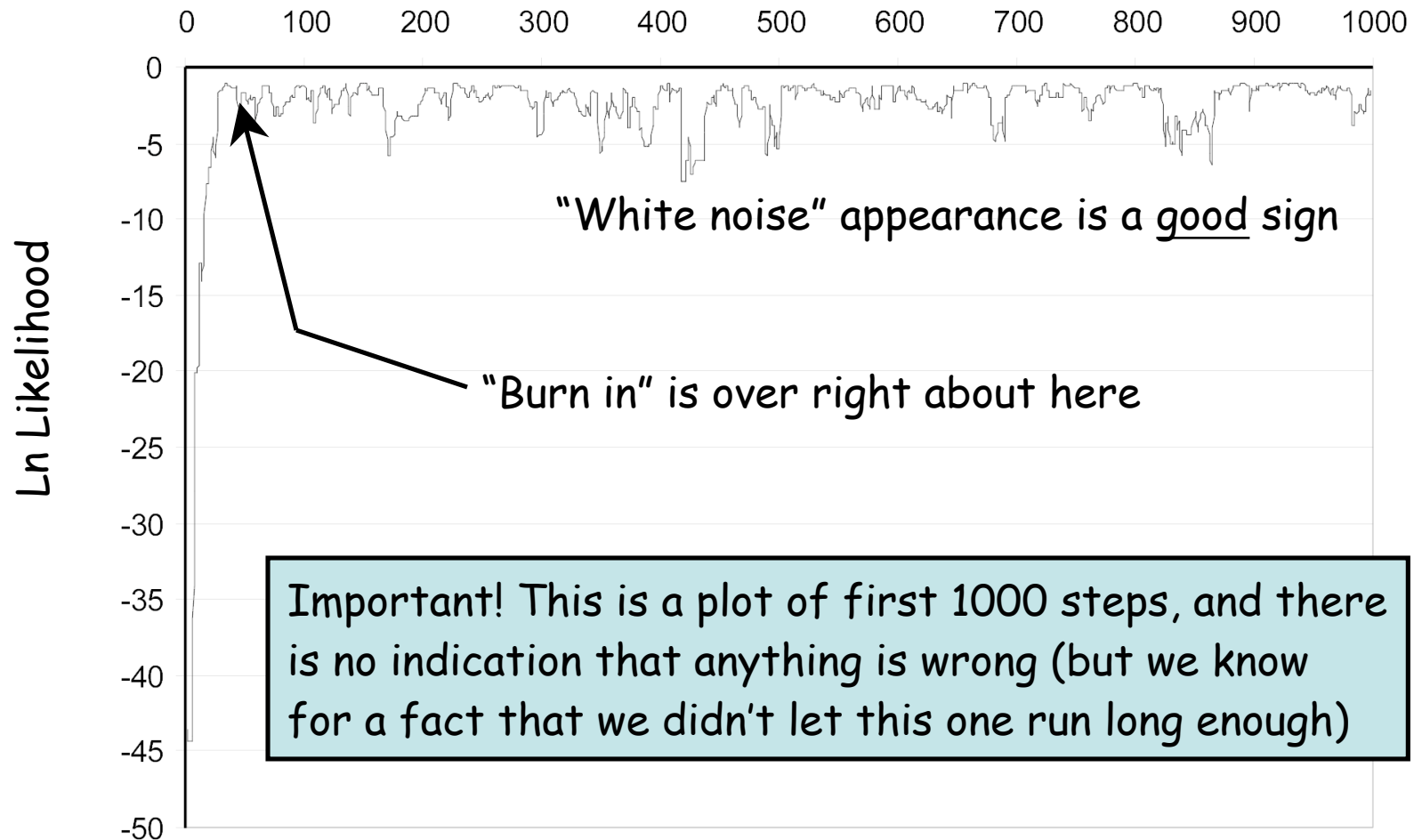


What would you conclude about the target distribution had you stopped the robot at this point?

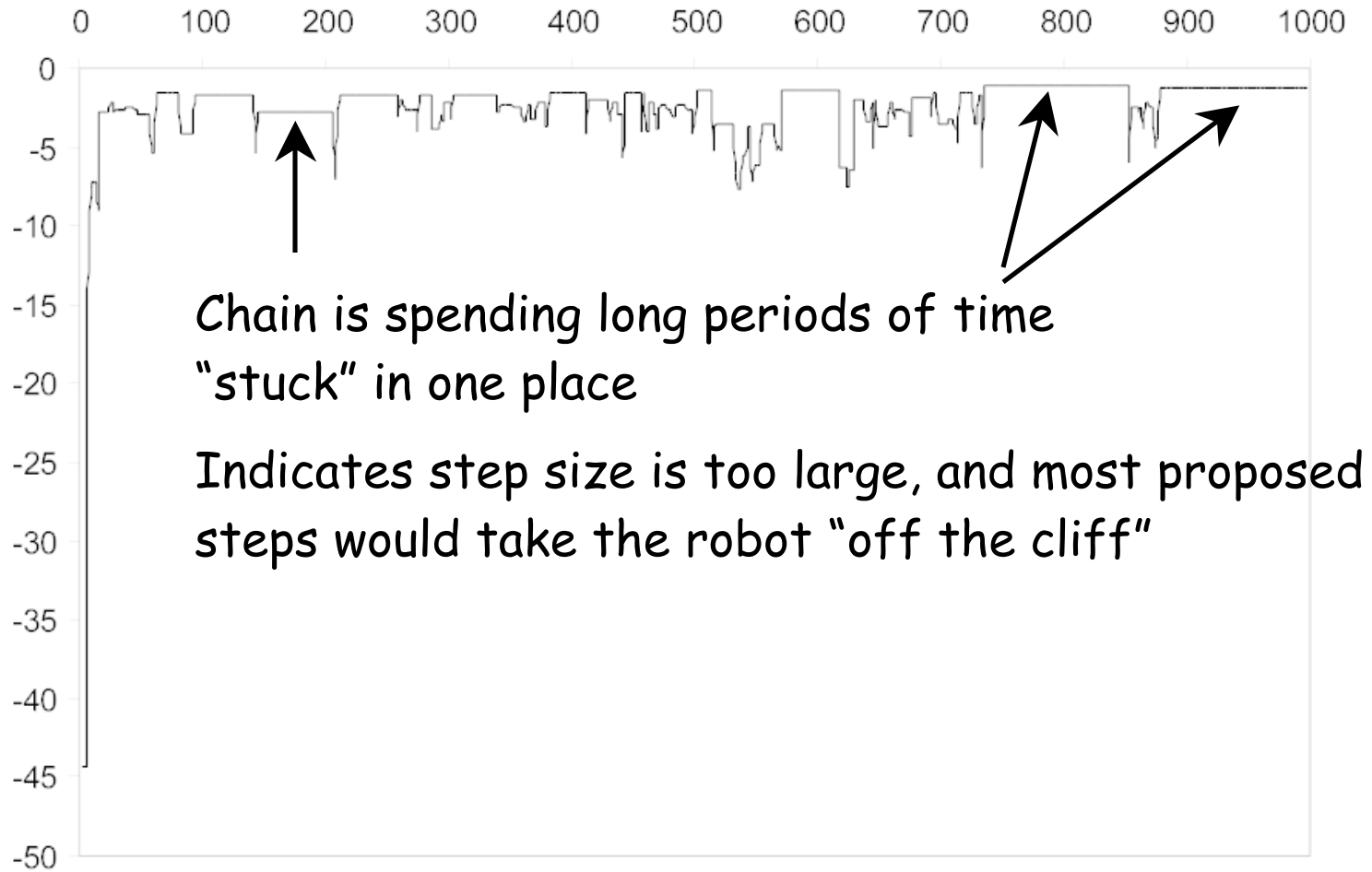
One way to detect this mistake is to perform **several independent runs**.

Results different among runs? Probably none of them were run long enough!

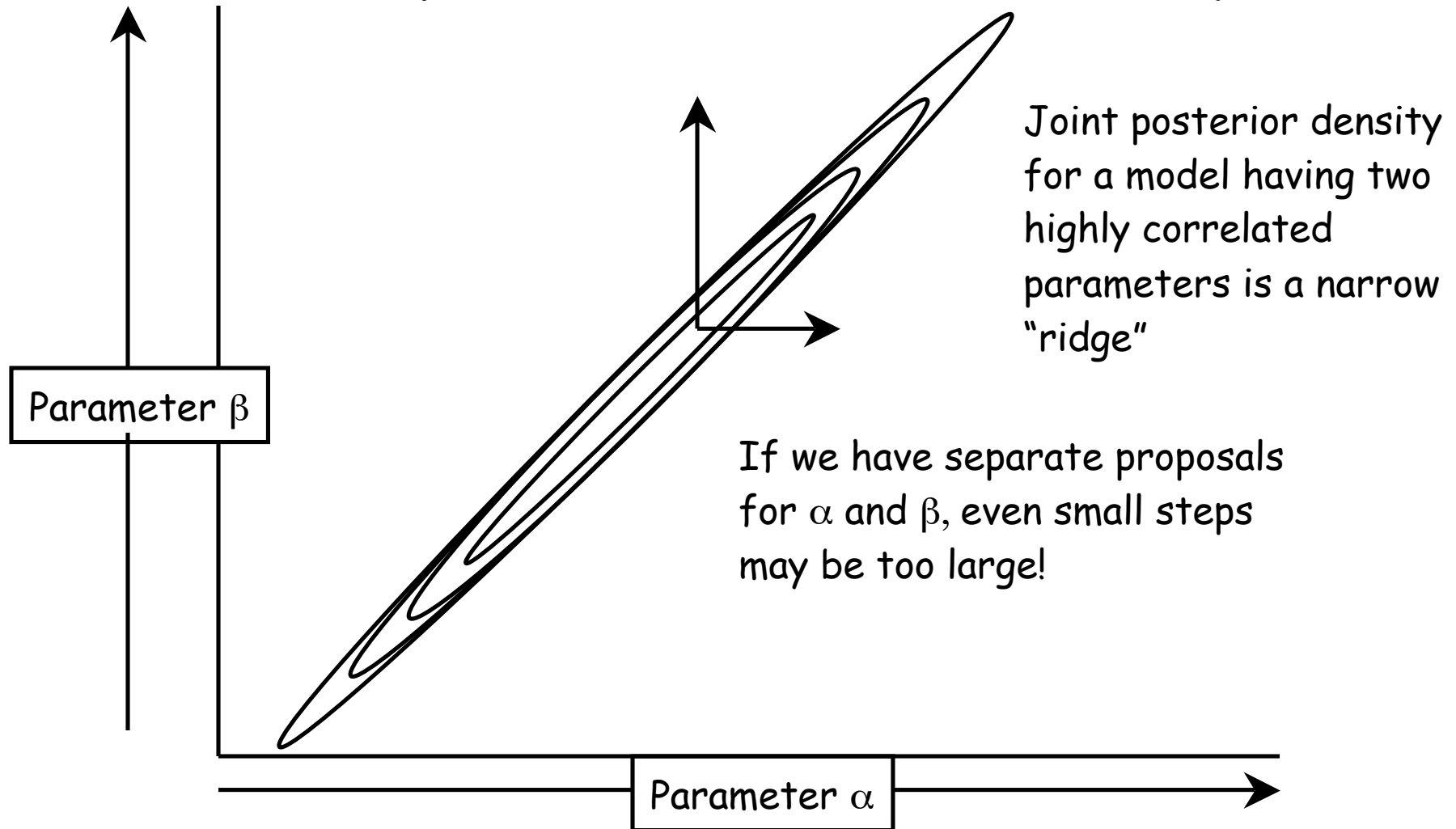
# History plots



# Slow mixing



# The problem of co-linearity



## Some material on Bayesian model comparison and hypothesis testing

1. Some Bayesians dislike much hypothesis testing because null hypotheses often are known *a priori* to be false and *p-value* depends both on “how” wrong null is and on amount of data.
2. Posterior predictive inference for assessing fit of models (see next pages)
3. Bayes factors for comparing models (see next pages)

# The Tradeoff

- *Pro:* Proposing big steps helps in jumping from one “island” in the posterior density to another
- *Con:* Proposing big steps often results in poor mixing
- Solution: Better proposals - MCMCMC

Huelsenbeck has found that a technique called Metropolis-Coupled Markov chain Monte Carlo (i.e., MCMCMC !! or MC<sup>3</sup>) suggested by C.J. Geyer is useful for getting convergence with phylogeny reconstruction.

The idea of MCMCMC is to run multiple Markov chains in parallel.

One chain will have stationary distribution that is the posterior of interest.

The other chains will approximate posterior distributions that are various degrees more smooth than the posterior distribution of interest.

Each chain is run separately, except that occasionally 2 chains are randomly picked and a proposal to switch the states of these two chains is made. This proposal is randomly accepted or reject with the appropriate probability



# Metropolis-coupled Markov chain Monte Carlo (MCMCMC, or MC<sup>3</sup>)

- MC<sup>3</sup> involves running **several chains simultaneously**
- The **cold chain** is the one that counts, the rest are **heated chains**.

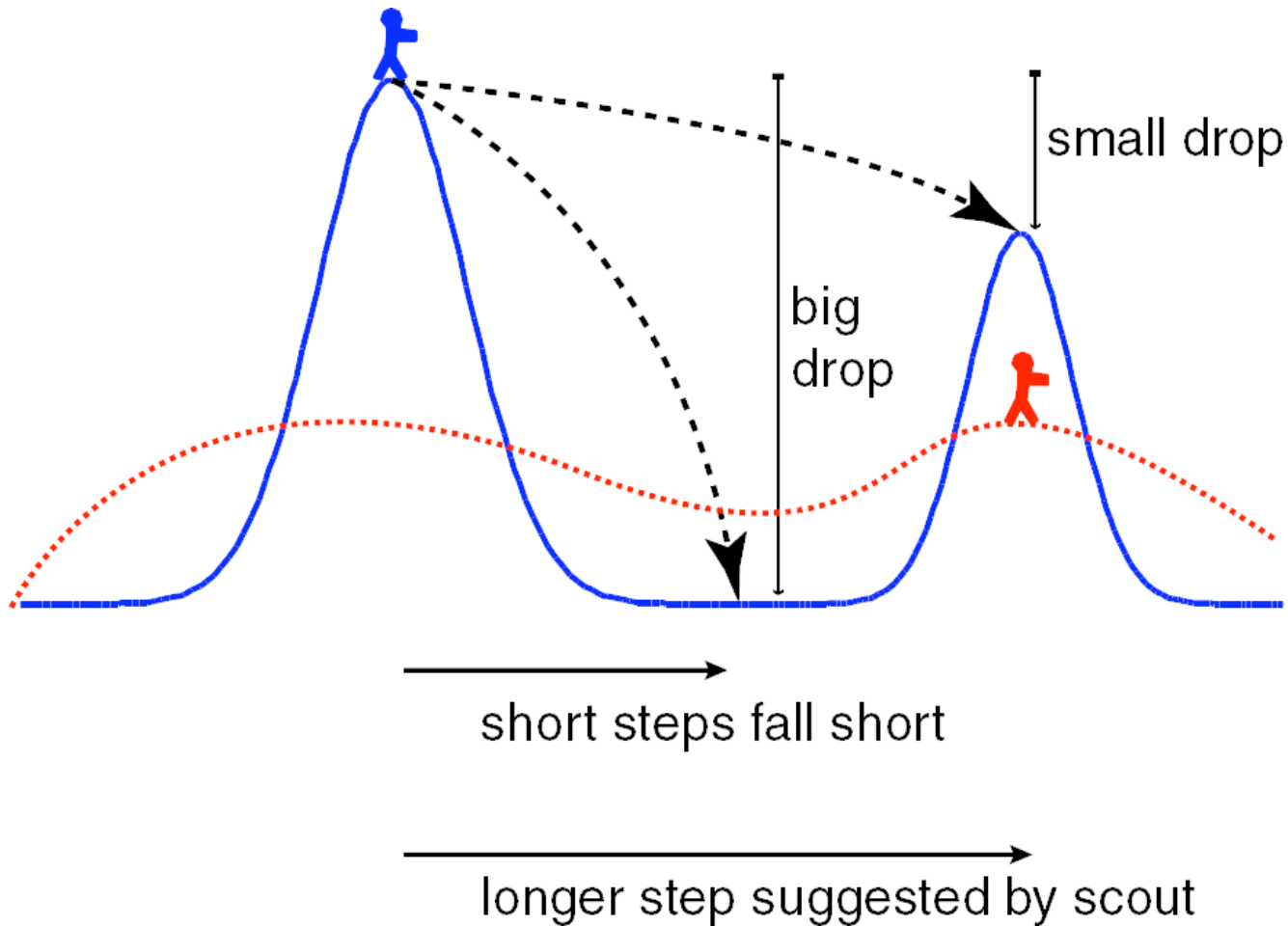
# What is a heated chain?

- Instead of using  $R$ , to make acceptance or rejection decisions, heated chains use:

$$R^{\frac{1}{1+H}}$$

- In MrBayes:  $H = \text{Temperature} * (\text{Chain's index})$
- The cold chain has index 0
- Heated chains explore the surface more freely
- Occasionally, you propose to switch the positions of 2 of the chains

# Heated chains act as scouts



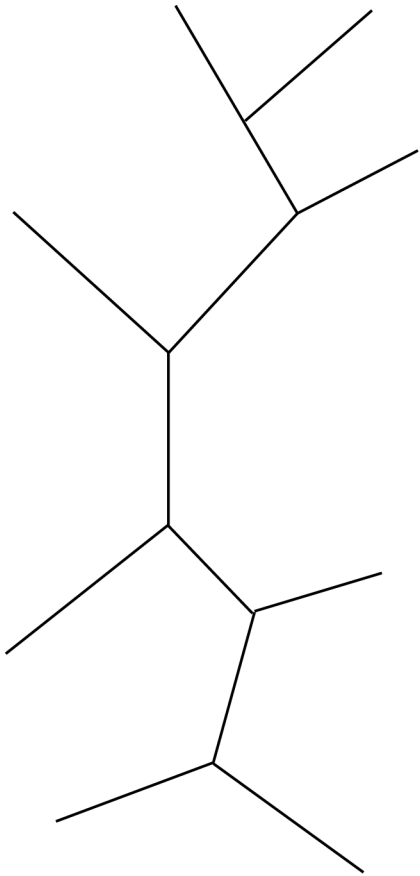
**Phylogeny Priors:** For phylogeny inference, parameters might represent topology, branch lengths, base frequencies, transition-transversion ratio, etc.

Each parameter needs specified prior distribution.  
For example...

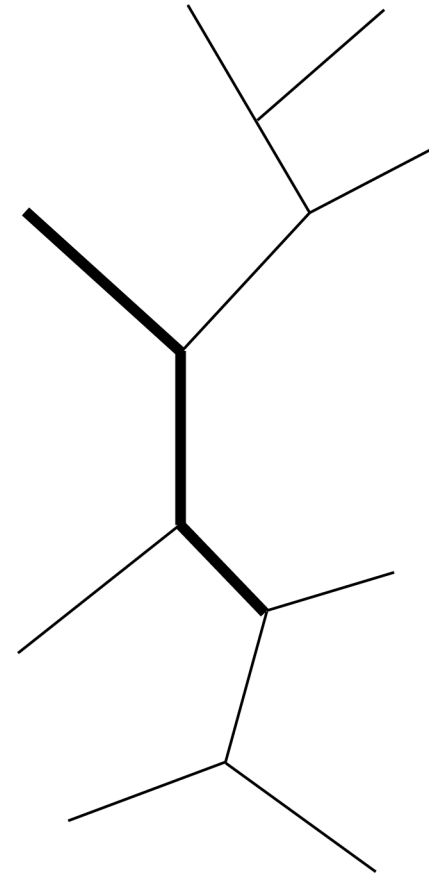
1. All unrooted topologies can be considered equally probable a priori. Given topology, all branch lengths between 0 and some big number could be considered equally likely a priori
  2. All combinations of base frequencies could be considered equally likely a priori
  3. The transition-transversion ratio could have a prior distribution that is uniform between 0 & some big number.
- ... and so on.

# Moving through Tree Space

## Target Simon Local Move

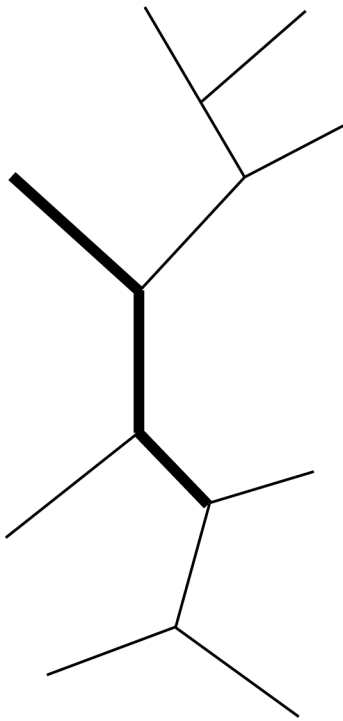


Step 1: Randomly select an internal branch and 2 of its neighbors



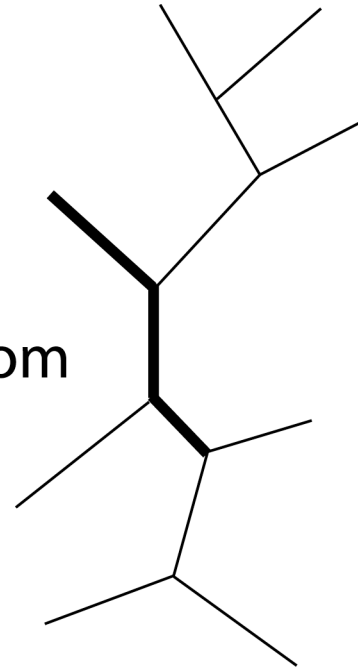
# Moving through Tree Space

## Target Simon Local Move



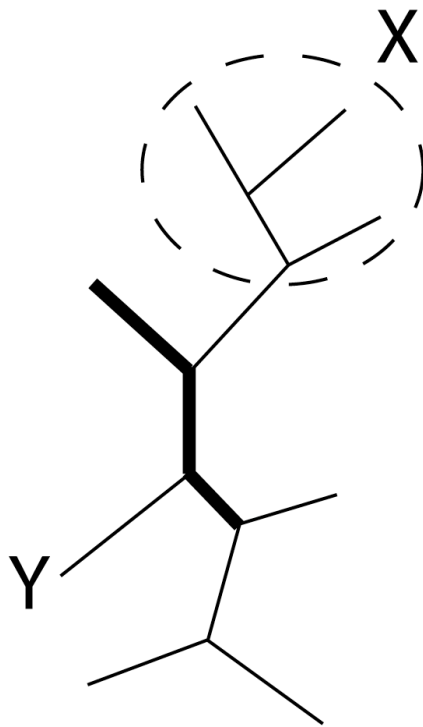
Step 2: Shrink or expand the selected segment by a random amount

$$m^* = m e^{\lambda(u-.5)}$$

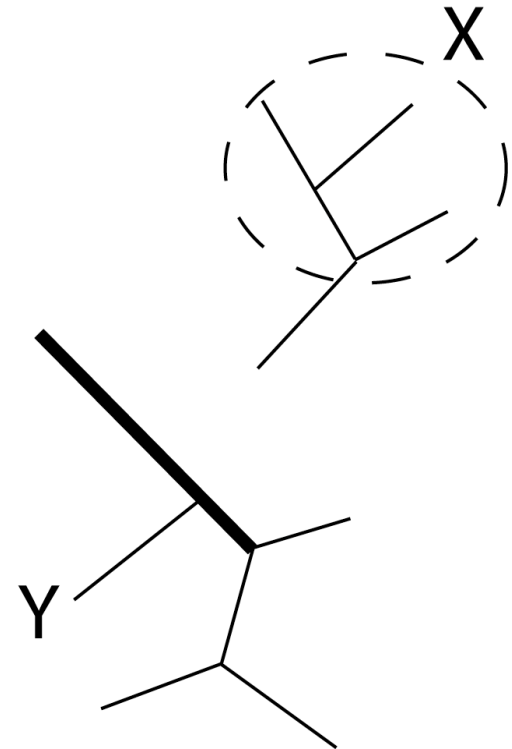


# Moving through Tree Space

## Target Simon Local Move



Step 3: Randomly select 1 of the 2 branches that intersect with the selected segment, and detach it

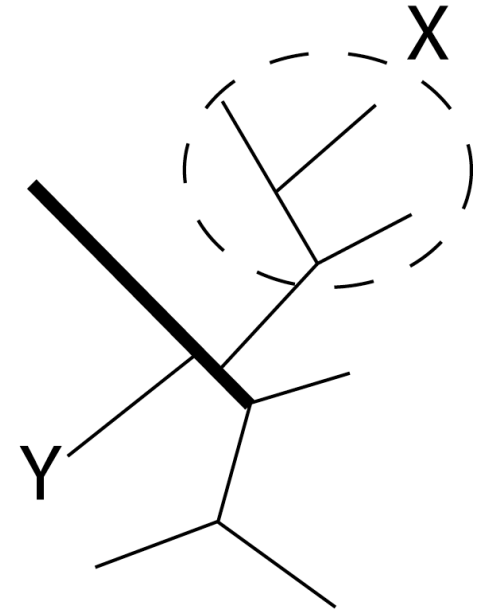


# Moving through Tree Space

## Target Simon Local Move

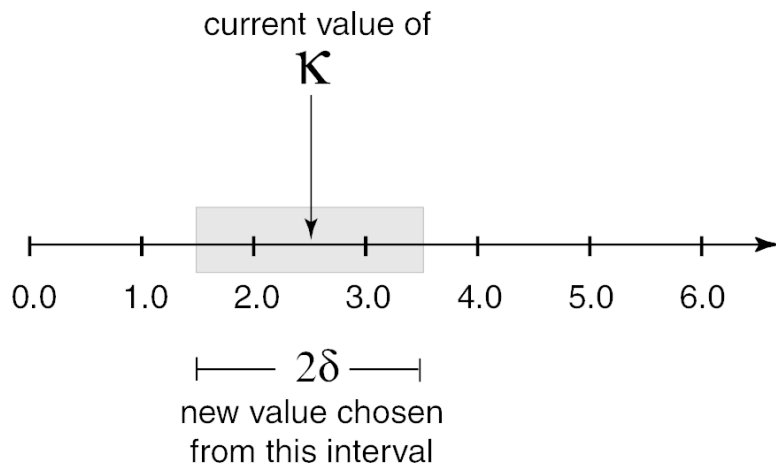


Step 4: Randomly reattach the clade X somewhere along the selected segment. This might result in a new topology.



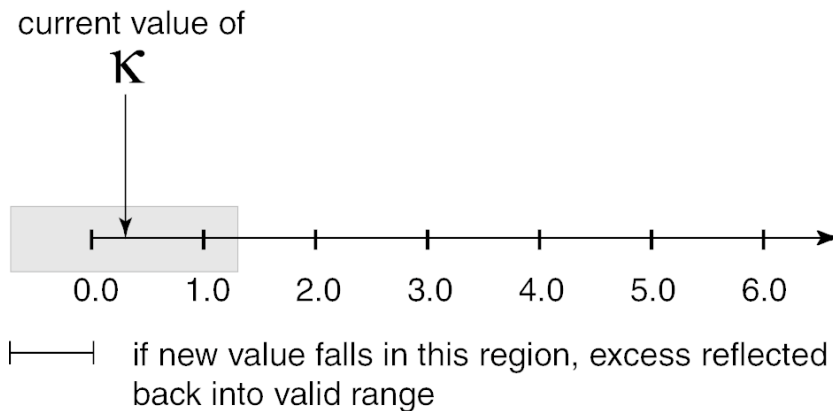


# Moving through parameter space



Using  $\kappa$  (ratio of the transition rate to the transversion rate) as an example of a model parameter.

Proposal distribution is the uniform distribution on the interval  $(\kappa - \delta, \kappa + \delta)$



A larger  $\delta$  means the sampler will attempt to make larger jumps on average.

# Putting it all together

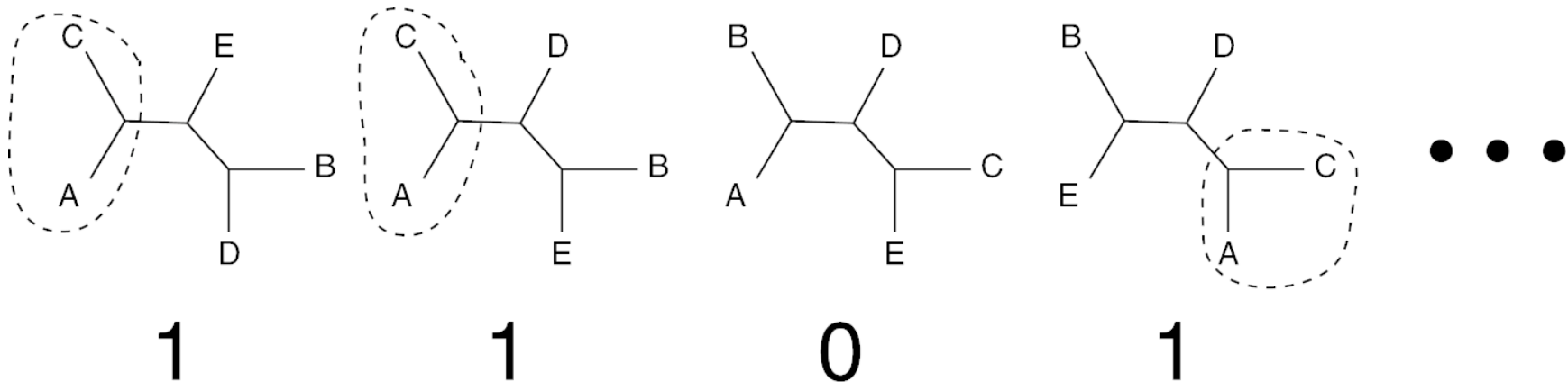
- Start with an initial tree and model parameters (often chosen randomly).
- Propose a new, randomly-selected move. Accept or reject the move (**Walking**).
- Every  $k$  generations, save tree, branch lengths and all model parameters (**Thinning**).
- After  $n$  generations, summarize the sample using histograms, means, credibility intervals, etc. (**Summarizing**).

## Sampling the chain tells us:

- Which tree has the highest posterior probability?
- What is the probability that "tree X" is the true tree?
- What values of the parameters are most probable?

# What if we are only interested in one grouping?

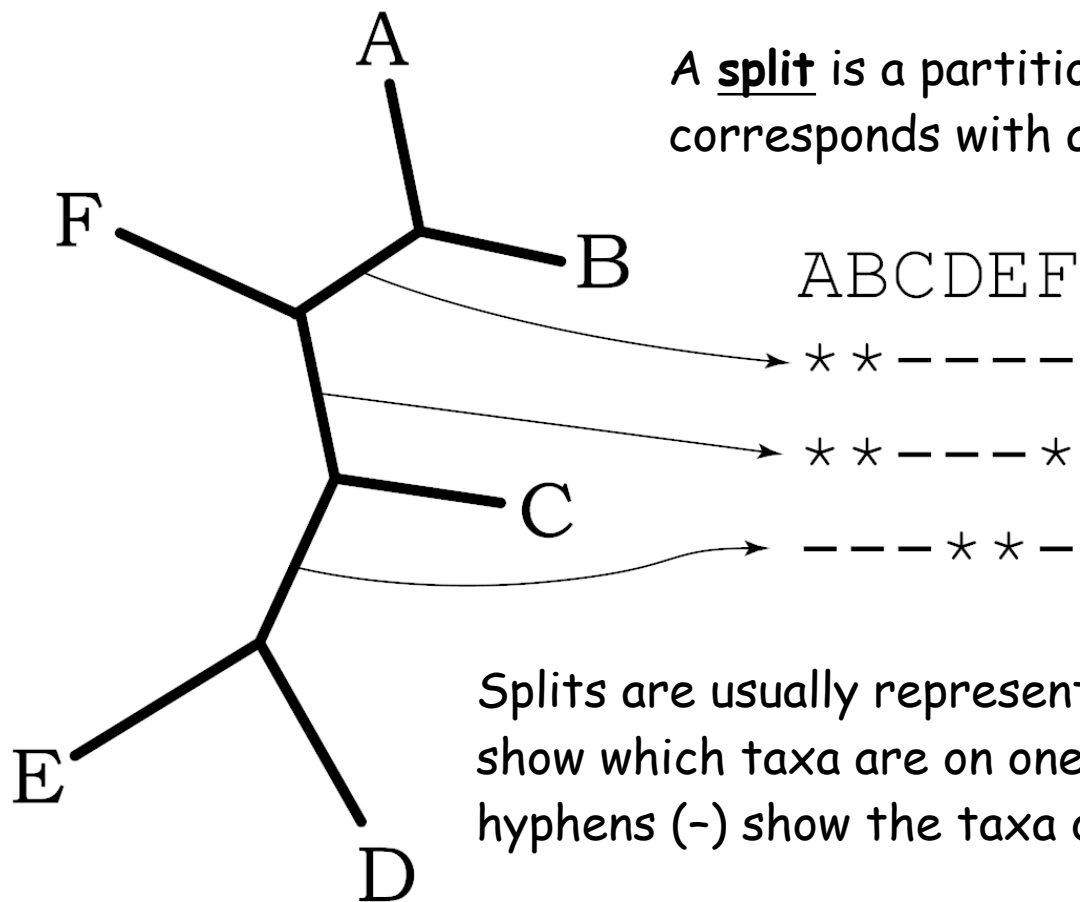
Which of the trees in the MCMC run contained the clade (e.g. A + C)?



The *proportion* of trees with A and C together in our sample approximates the posterior probability that A and C are sister to each other.

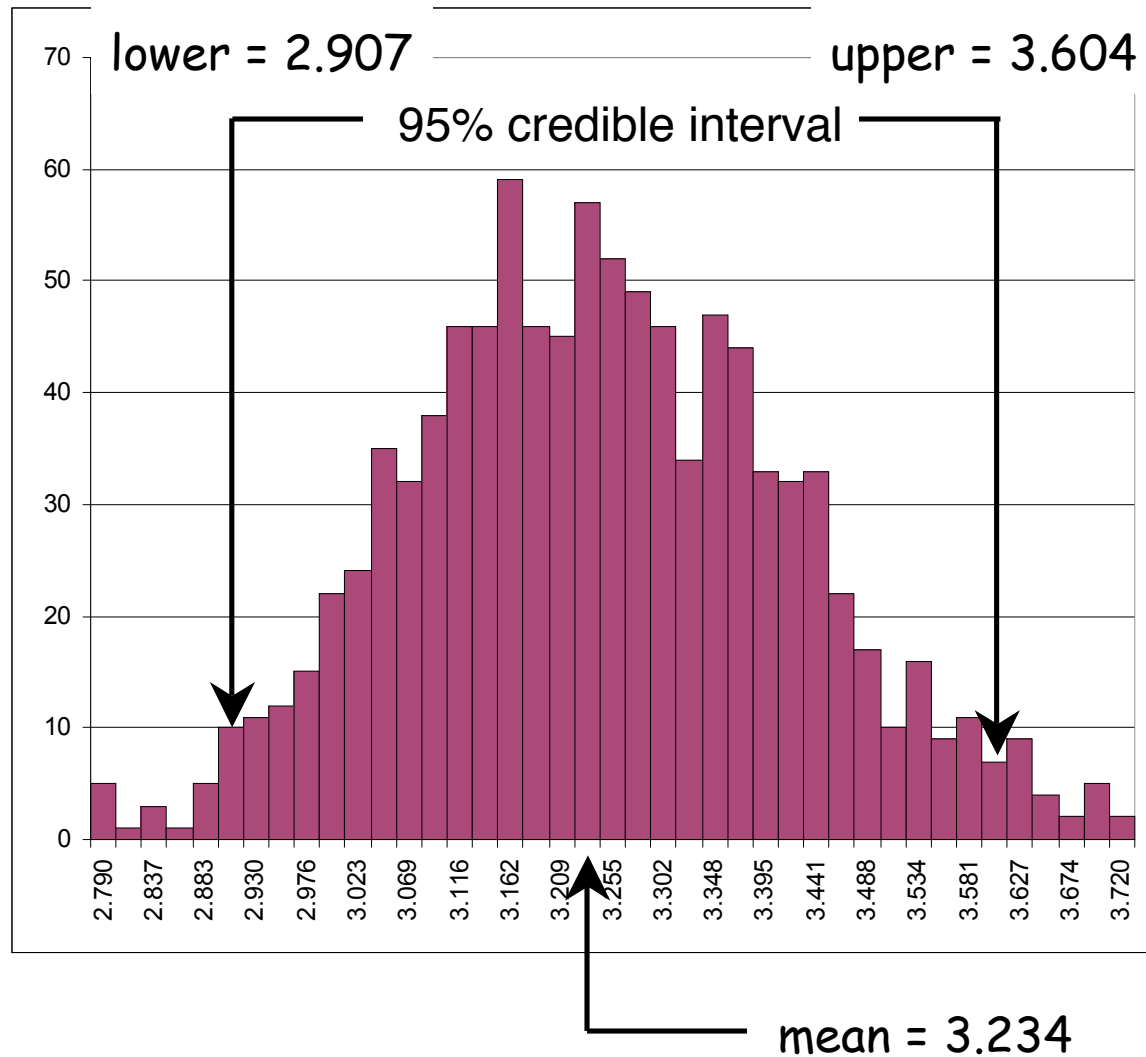
# Split (a.k.a. clade) probabilities

A split is a partitioning of taxa that corresponds with a particular branch.



Splits are usually represented by strings: asterisks (\*) show which taxa are on one side of the branch, and the hyphens (-) show the taxa on the other side.

# Posteriors of model parameters



Histogram created from a sample of 1000 kappa values.

From: Lewis, L., & Flechtner, V. (2002. *Taxon* **51**:443-451)