MCMC proposals

Bayesian phylogenetic MCMC analysis in a nutshell

1. Choose starting states
2. Update one or more parameters
   a. Choose parameter(s)
   b. Propose new state(s) for parameter(s)
   c. Calculate ratio R
   d. Draw u from Uniform(0,1) distribution
   e. if $u < R$, accept; otherwise, reject
3. Save tree and parameters
4. Go to 2
5. Summarize posterior samples

loop until sufficient sample points have been saved

Moving through treespace
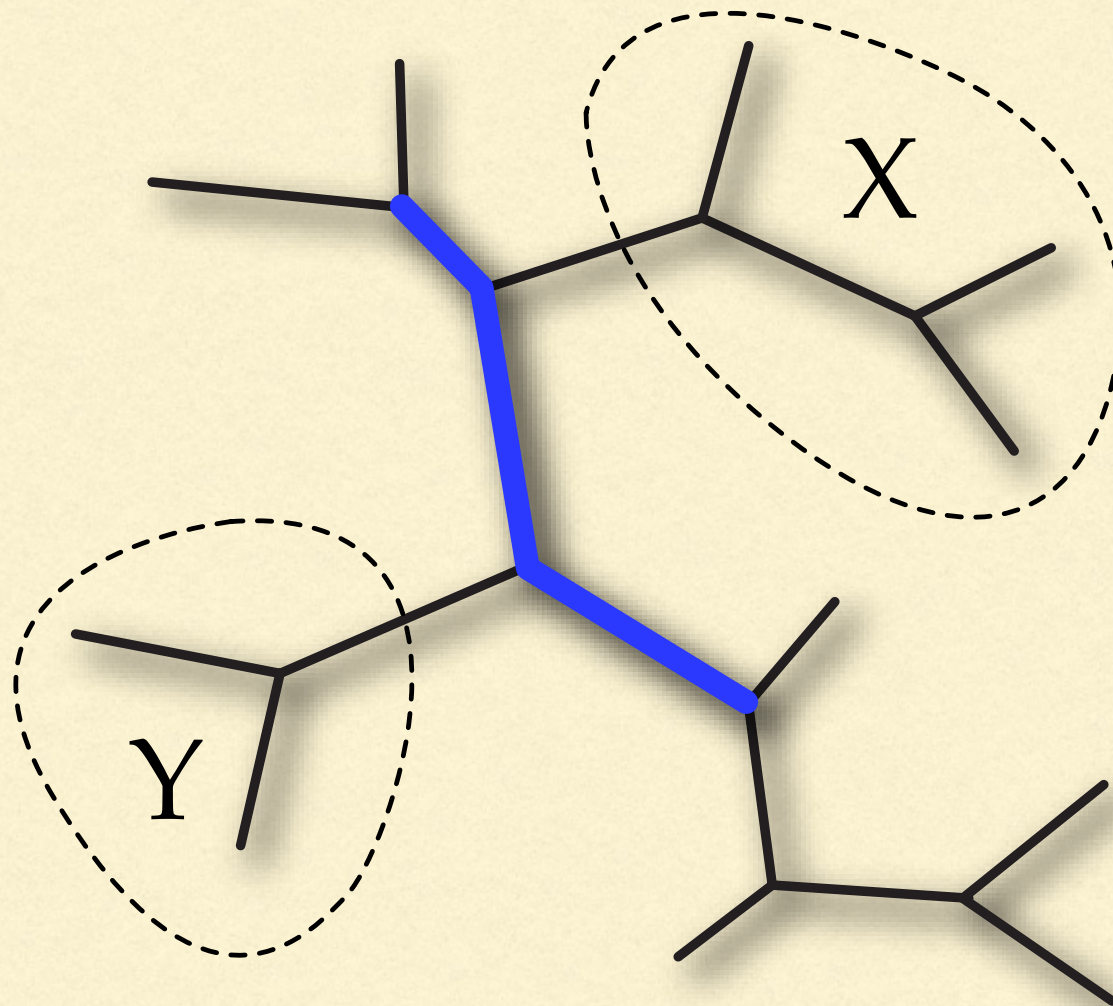
The Larget-Simon move

Step 1:
Pick 3 contiguous edges randomly, defining two subtrees, X and Y

Larget, B., and D. L. Simon. 1999. Mol. Biol. Evol. 16: 750-759.
See also: Holder et al. 2005. Syst. Biol. 54: 961-965.
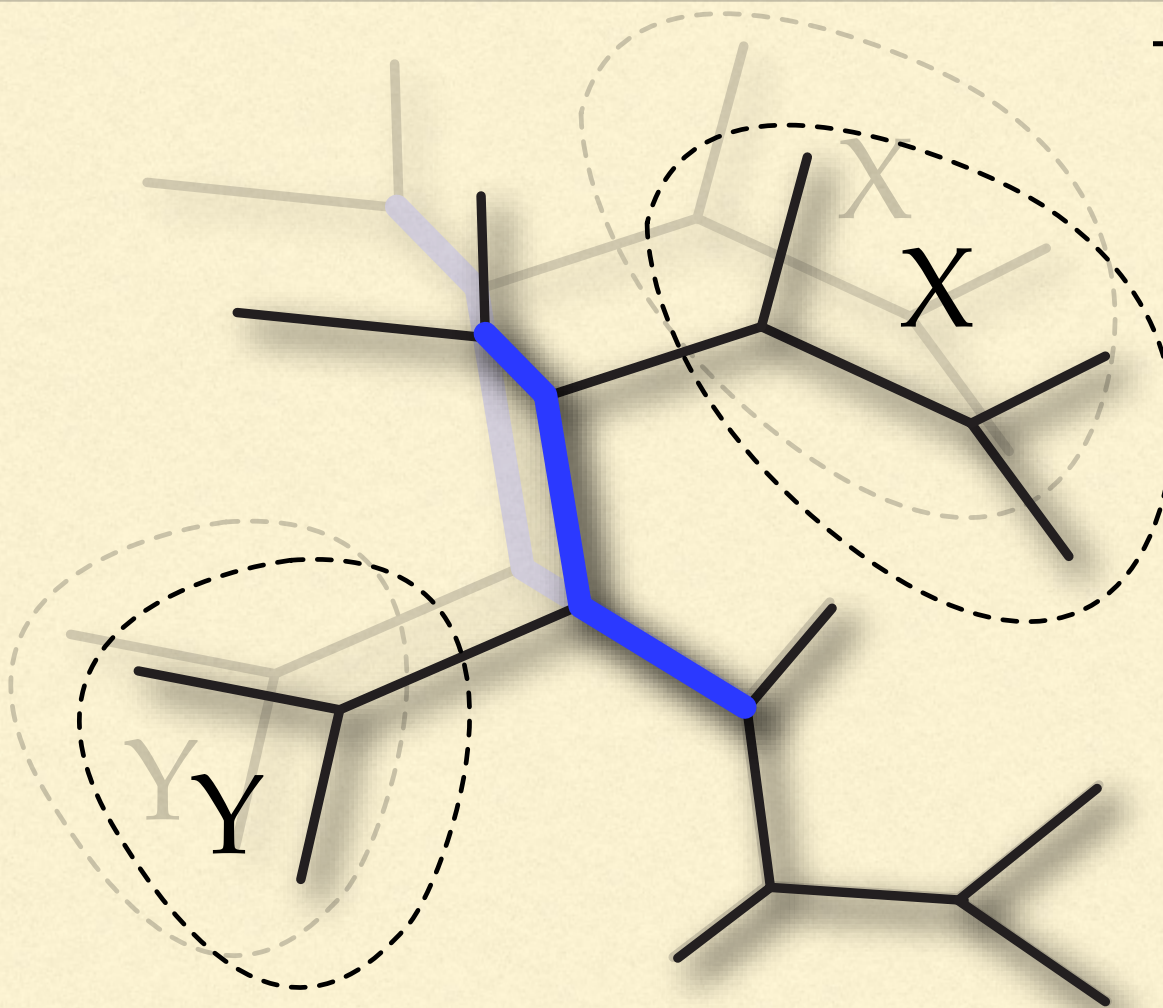
Moving through treespace

The Larget-Simon move

Step 1:
Pick 3 contiguous edges randomly, defining two subtrees, X and Y

Step 2:
Shrink or grow selected 3-edge segment by a random amount

Moving through treespace

The Larget-Simon move

Step 1:
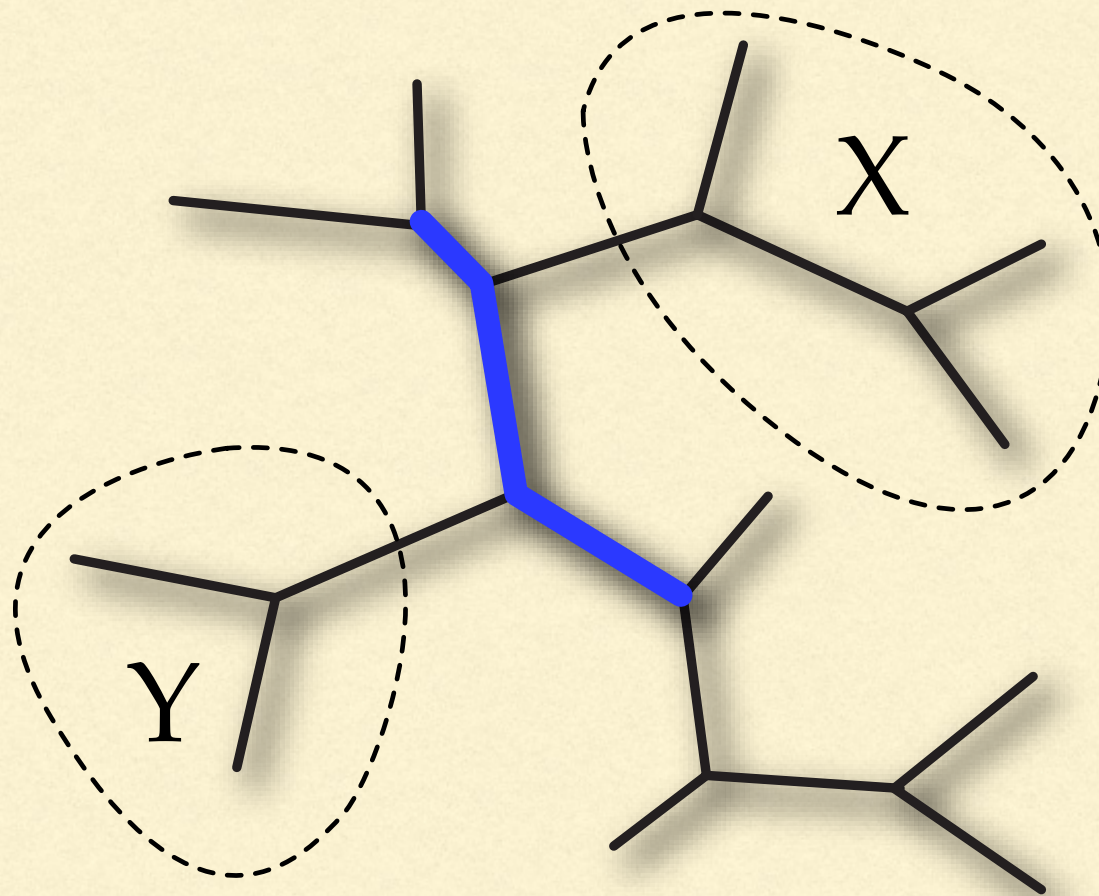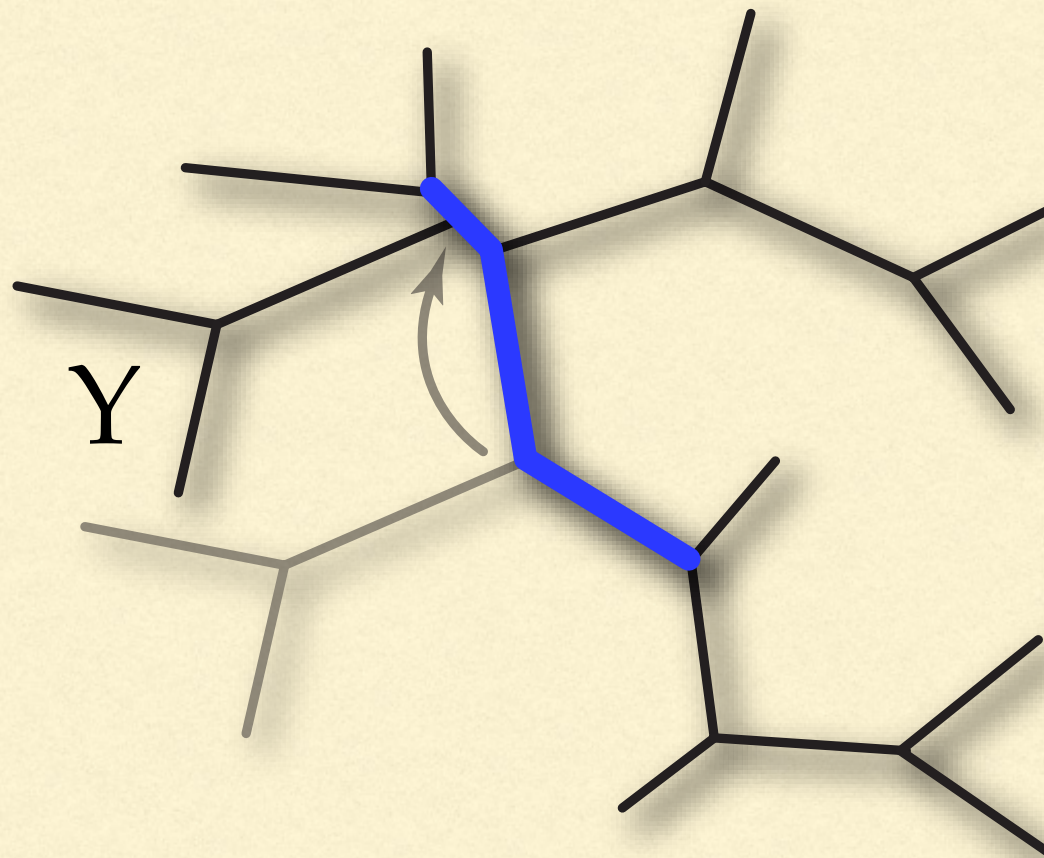Pick 3 contiguous edges randomly, defining two subtrees, X and Y

Step 2:
Shrink or grow selected 3-edge segment by a random amount

Moving through treespace

The Larget-Simon move

Step 1:
Pick 3 contiguous edges randomly, defining two subtrees, X and Y

Step 2:
Shrink or grow selected 3-edge segment by a random amount

Step 3:
Choose X or Y randomly, then reposition randomly

Y

Moving through treespace

The Larget-Simon move

Step 1:
Pick 3 contiguous edges randomly, defining two subtrees, X and Y

Step 2:
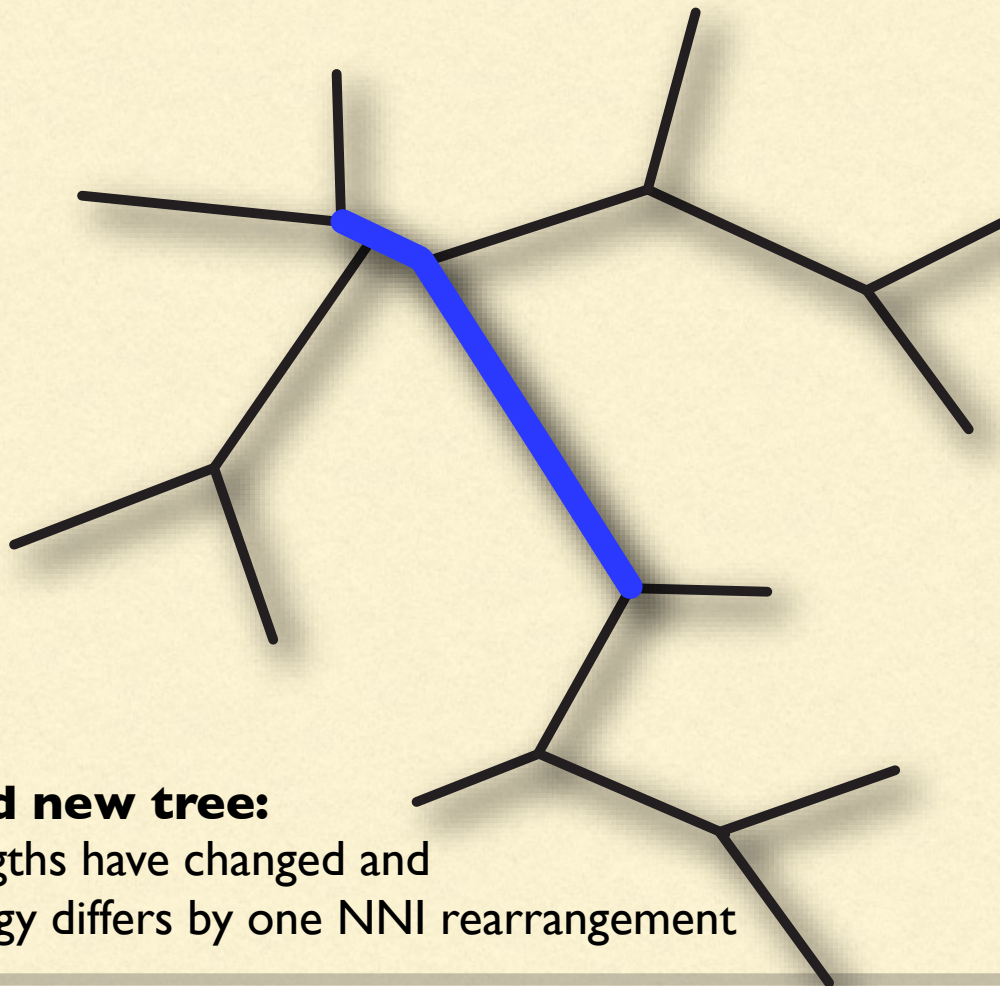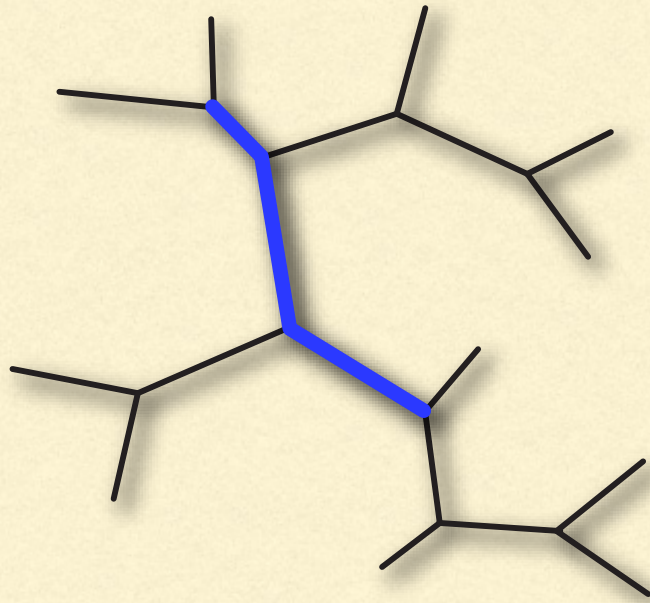Shrink or grow selected 3-edge segment by a random amount

Step 3:
Choose X or Y randomly, then reposition randomly

Proposed new tree:
3 edge lengths have changed and the topology differs by one NNI rearrangement

Moving through treespace

Current tree

log-posterior = -34256
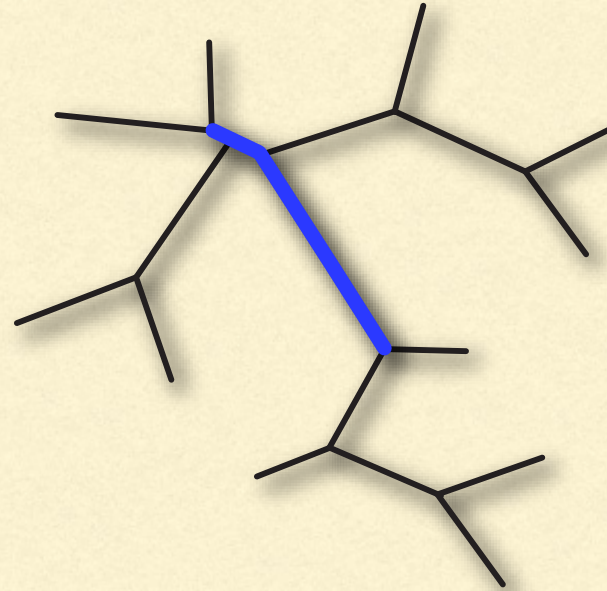
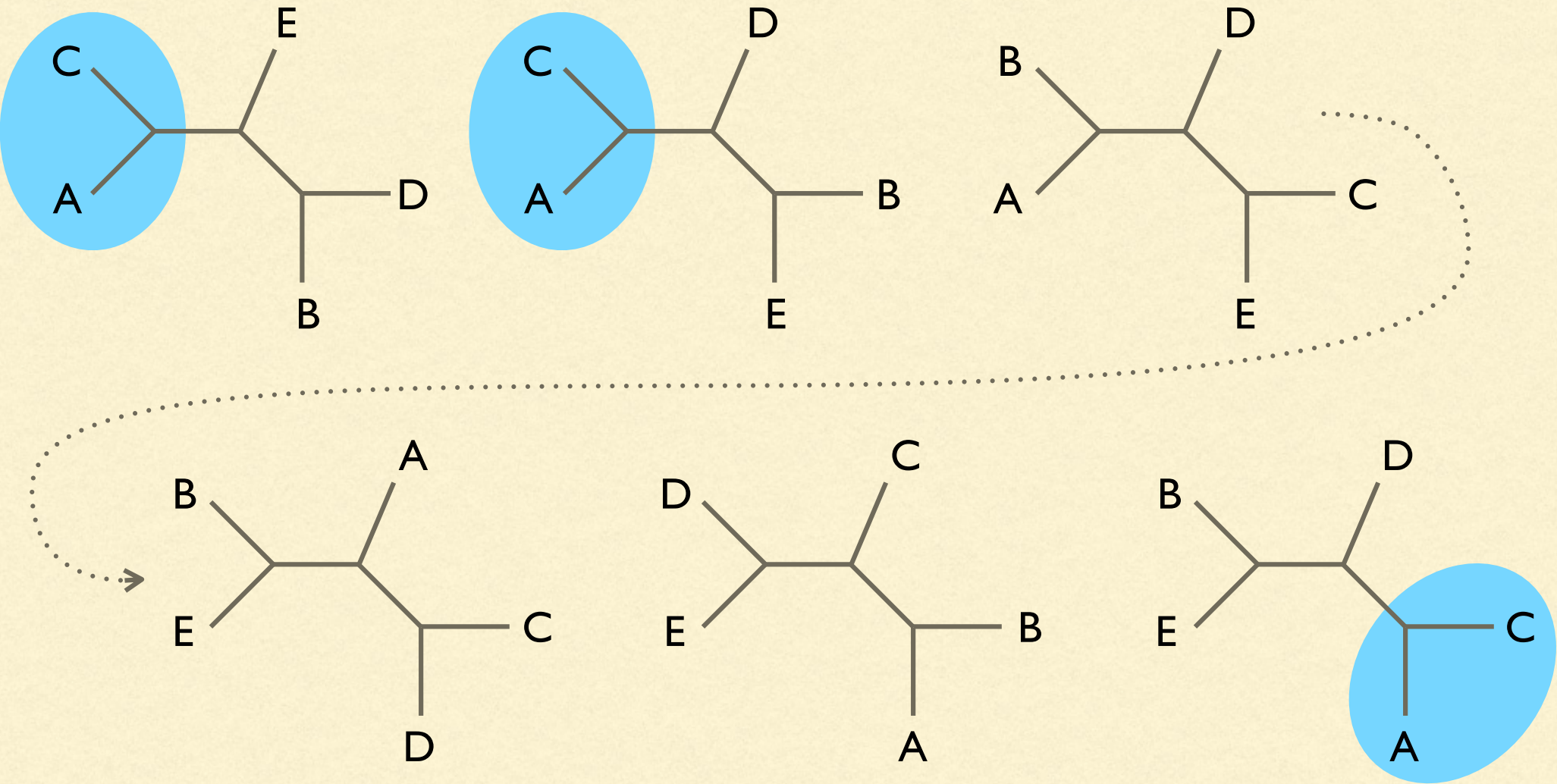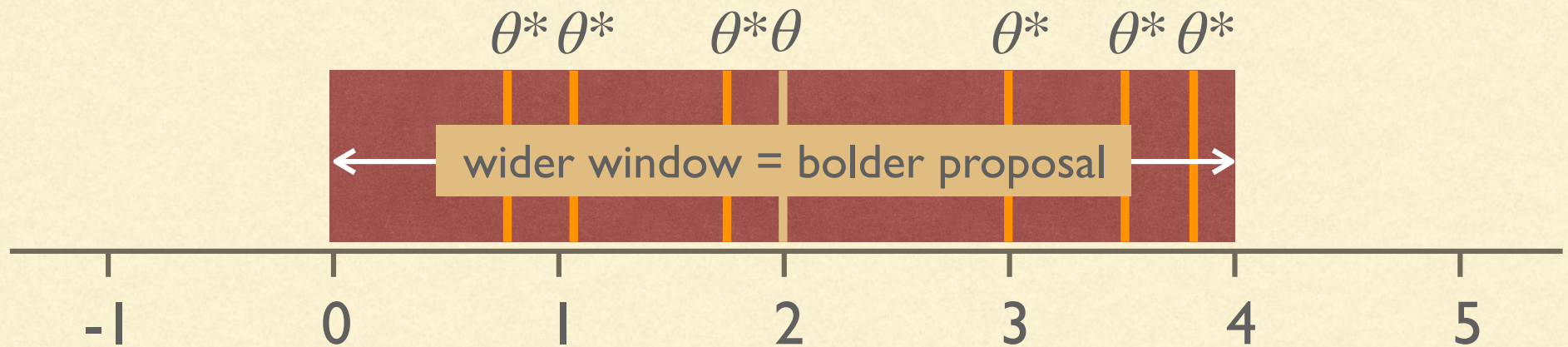**Proposed tree**

log-posterior = -32519
(better, so accept)

Marginal split posterior probabilities

Sliding window proposal

propose new values ($\theta*$) uniformly
within a proposal window
centered at current value (θ)

$\theta* \; \theta*$ $\theta* \, \theta$ $\theta*$ $\theta* \; \theta*$

wider window = bolder proposal

-1   0   1   2   3   4   5

Sliding window proposal

If proposed value is out of bounds,
reflect it back in bounds

$\theta* \qquad \theta* \theta* \qquad \theta* \; \theta \qquad\qquad \theta* \qquad \theta*$

-1   0   1   2   3   4   5
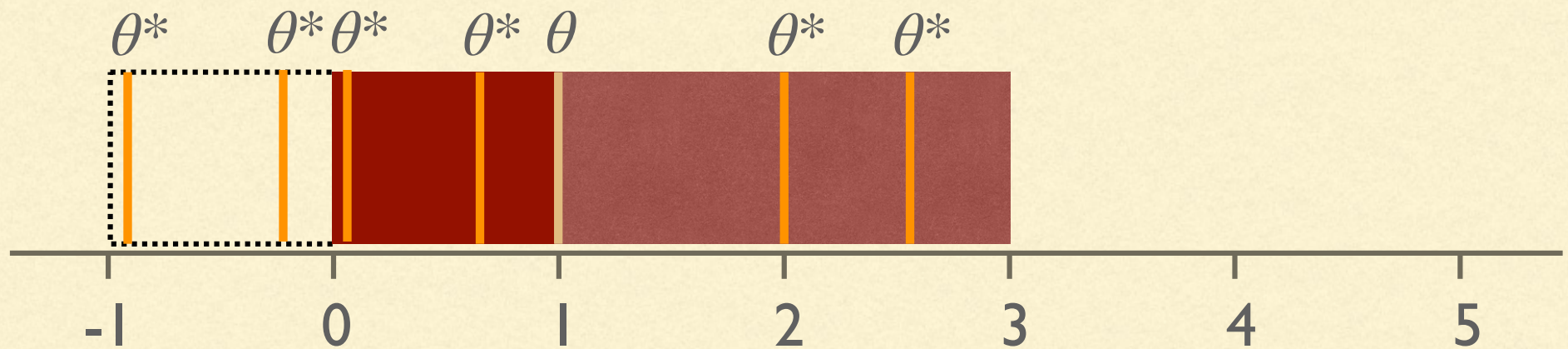
Surprisingly, this is a symmetric
proposal, so the Hastings ratio = 1.0

Marginal distributions and credible intervals

Histogram created from a sample of 10000 κ values using Tracer

95% HPD credible interval
(HPD = Highest Posterior Density)

1.7075

3.4915

mean
2.6029

Tracer downloaded from https://github.com/beast-dev/tracer/releases/tag/v1.7.1
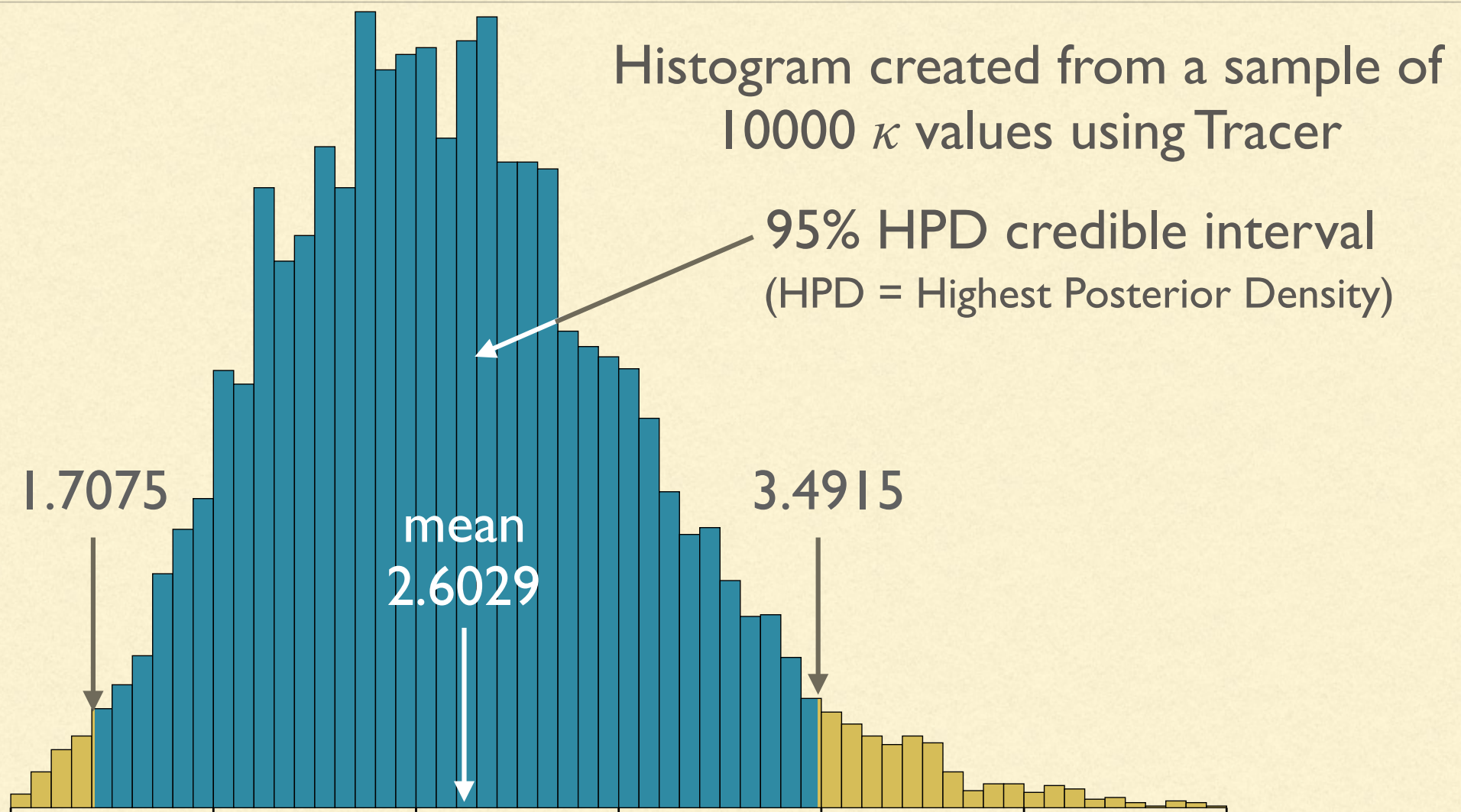
Prior distributions

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{p(D)}$$

Gamma(*a,b*) distribution

Gamma(0.1, 10)
shoots off to infinity if *a* < 1

Gamma(400, 0.01)
mode > 0 if *a*>1

Exponential(1)
= Gamma(1,1)
hits y-axis at 1/*b* if *a*=1
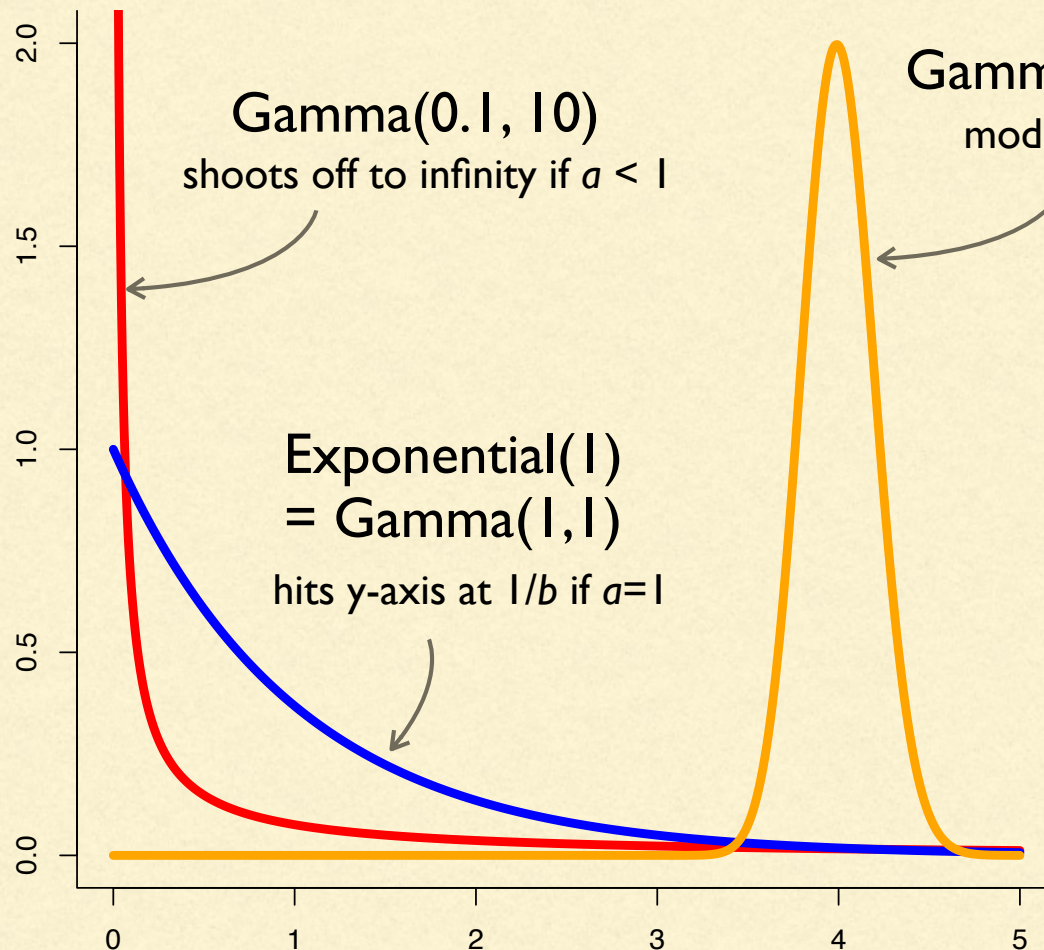
Gamma(*a,b*)
distributions are appropriate for parameters that range from 0 to infinity (e.g. branch lengths)
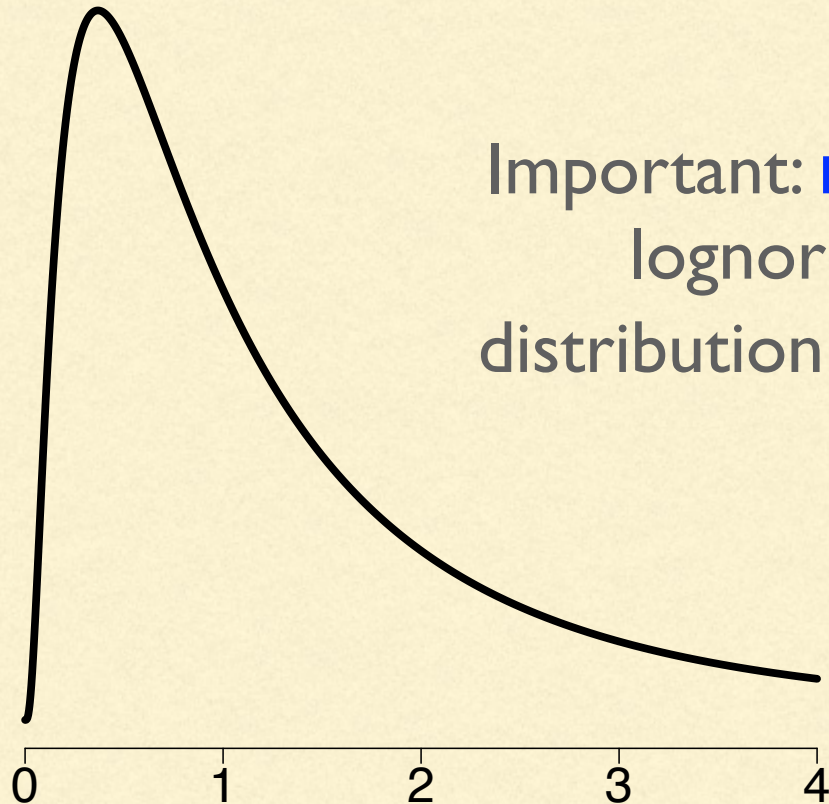
a = shape
b = scale*
mean = *ab*
variance = *ab*²

*Note: be aware that in many papers the Gamma distribution is defined such that the second parameter is the rate (*inverse* of the scale *b* used in this slide) rather than the scale! In this case, the mean and variance would be *a*/*b* and *a*/*b*², respectively.
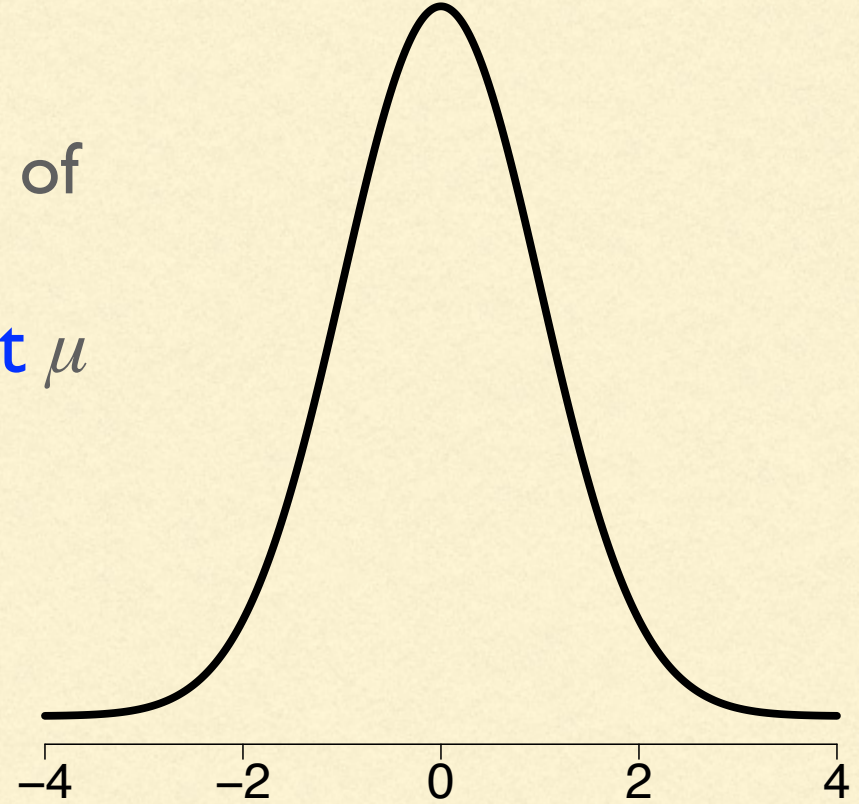
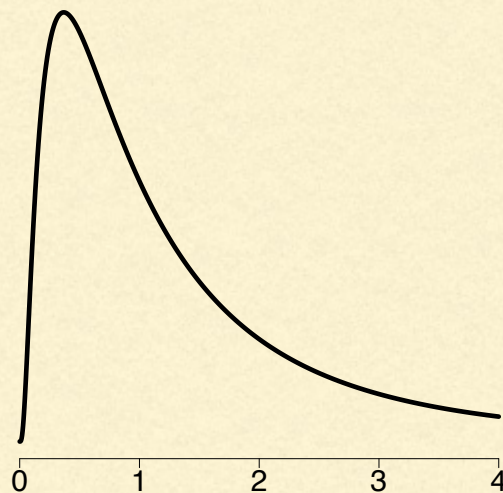# Lognormal(μ,σ) distribution

Lognormal($\mu = 0, \sigma = 1$)

Normal($\mu = 0, \sigma = 1$)

Important: mean of lognormal distribution is not μ

Lognormal(μ,σ) distribution

$\mu = 0, \sigma = 1$

1.65	mean $= e^{\mu + \sigma^2/2}$
4.67	variance $= e^{2\mu + \sigma^2}(e^{\sigma^2} - 1)$
0.37	mode $= e^{\mu - \sigma^2}$
1.0	median $= e^{\mu}$

To choose μ and σ to yield a particular mean (m) and variance (v) for a lognormal prior, use these formulas (log is natural logarithm):
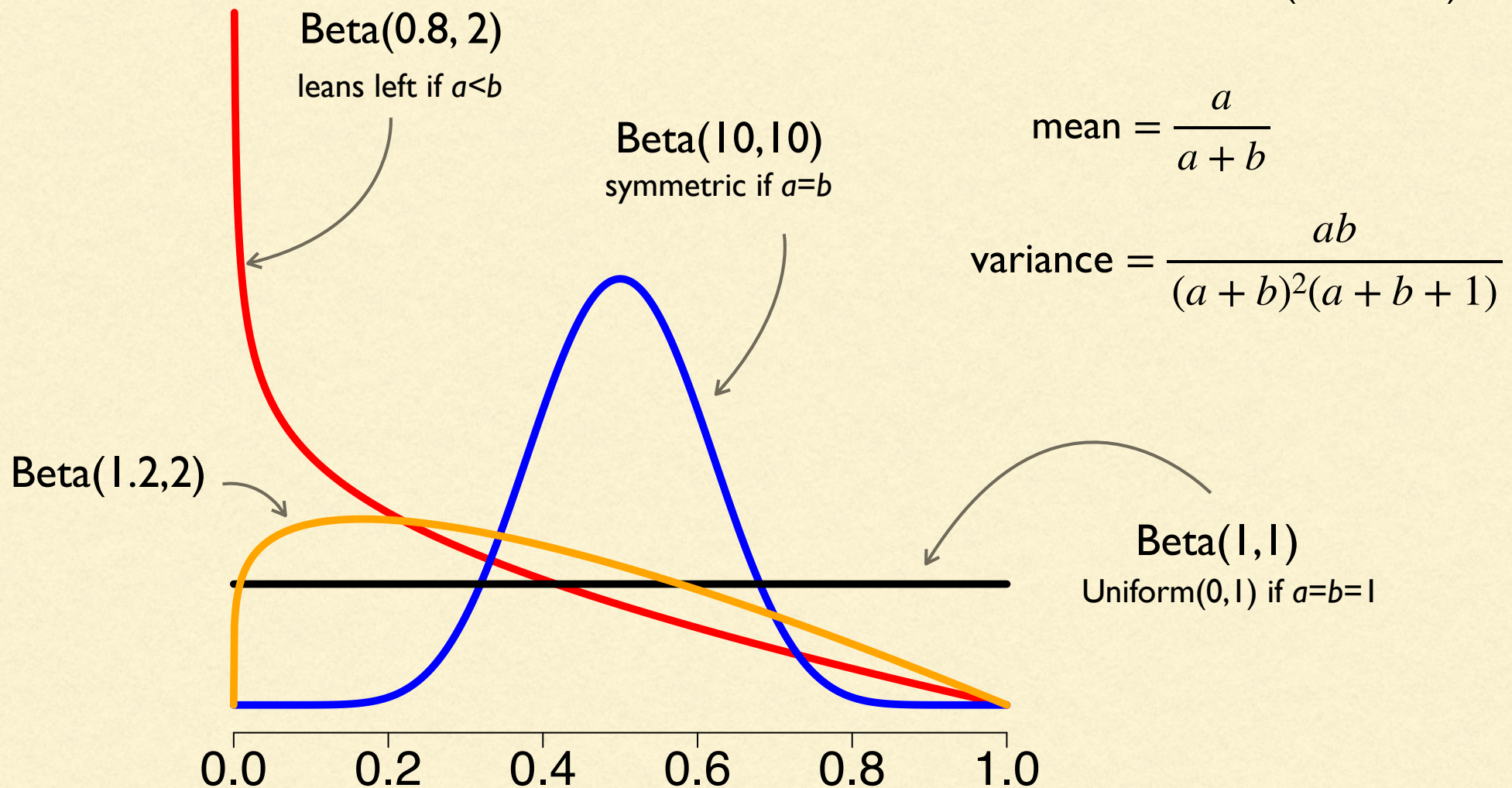
$$\mu = \log(m^2) - \log(m) - \frac{\log(v + m^2) - \log(m^2)}{2}$$

$$\sigma^2 = \log(v + m^2) - \log(m^2)$$

Beta(*a,b*) distribution

Beta(*a,b*) distributions are appropriate for proportions, which must lie between 0 and 1 (inclusive).

Beta(0.8, 2)

leans left if *a<b*

Beta(10,10)

symmetric if *a=b*

$$\text{mean} = \frac{a}{a+b}$$

$$\text{variance} = \frac{ab}{(a+b)^2(a+b+1)}$$

Beta(1.2,2)

Beta(1,1)

Uniform(0,1) if *a=b=*1

0.0   0.2   0.4   0.6   0.8   1.0

Dirichlet(*a*,*b*,*c*,*d*) distribution
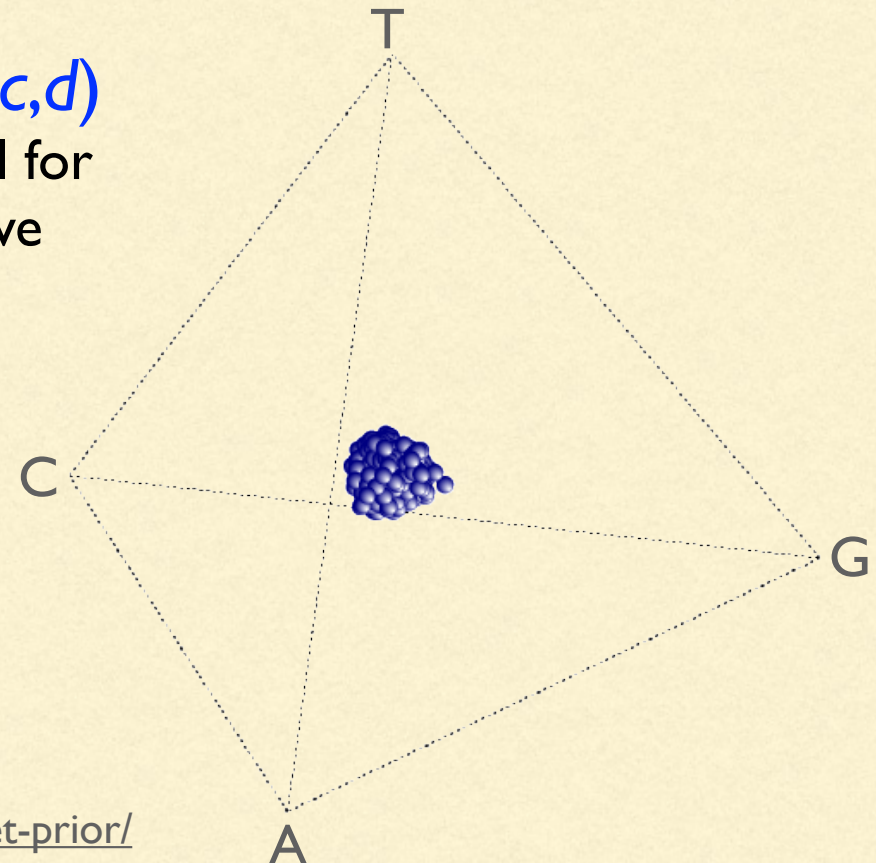
Flat: *a* = *b* = *c* = *d* = 1
(every combination equally probable)
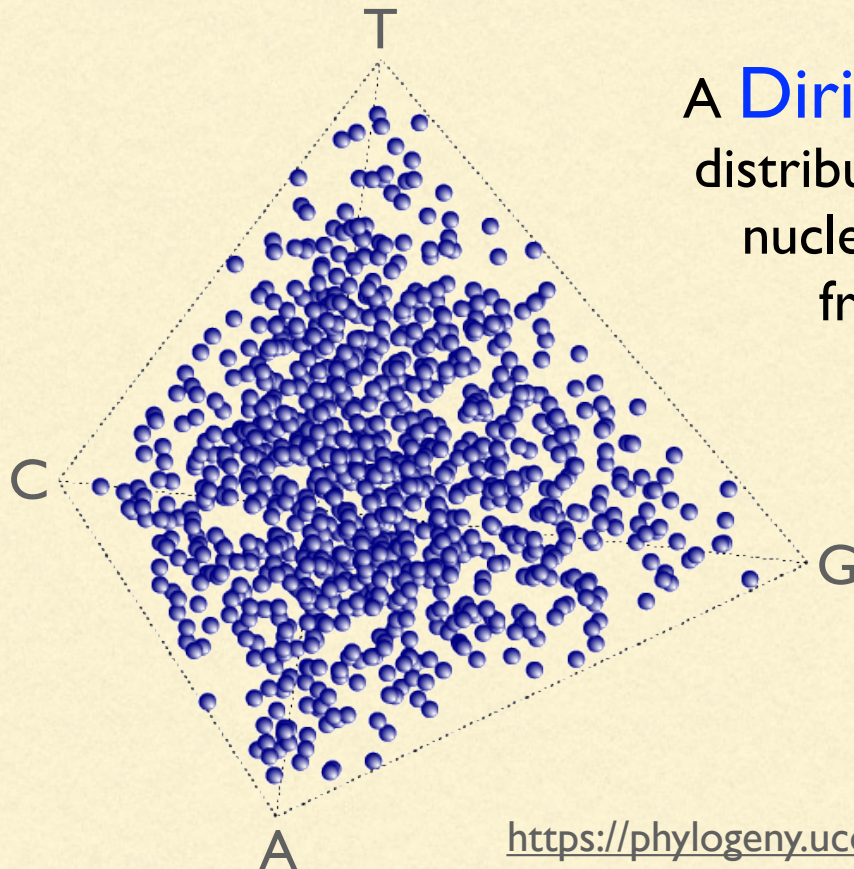
Informative: *a* = *b* = *c* = *d* = 100
(frequencies tend to be nearly equal)
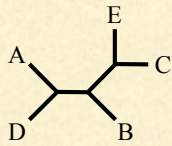
A Dirichlet(*a*,*b*,*c*,*d*) distribution is ideal for nucleotide relative frequencies.

https://phylogeny.uconn.edu/dirichlet-prior/

Topology: discrete uniform
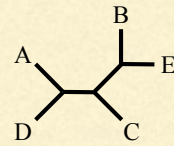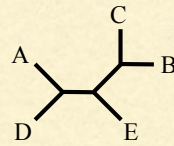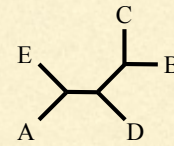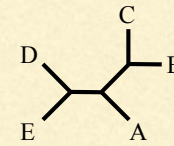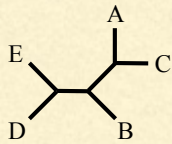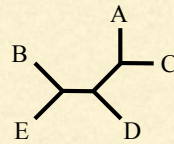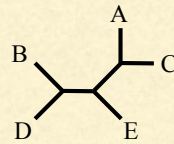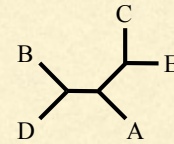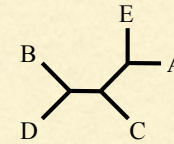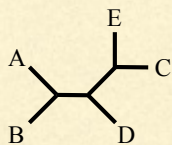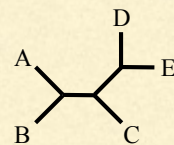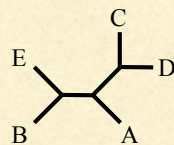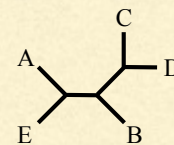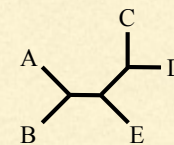
Edge lengths: Gamma-Dirichlet

Tree length (TL)
Gamma(2,2)

Edge lengths: Gamma-Dirichlet

Tree length (TL)
Gamma(2,2)

Dir(1,1,1,1,1,1,1)   Dir(10,10,10,10,10,10,10)

Edge length proportions

Yule (pure birth) prior

		Time to next
t_1	6λ	speciation
t_2	5λ	event is
t_3	4λ	exponentially
		distributed
		with rate
t_4	3λ	proportional
		to the
		number of
t_5	2λ	lineages.

It is possible to specify a prior distribution that jointly specifies both tree topology and edge lengths.
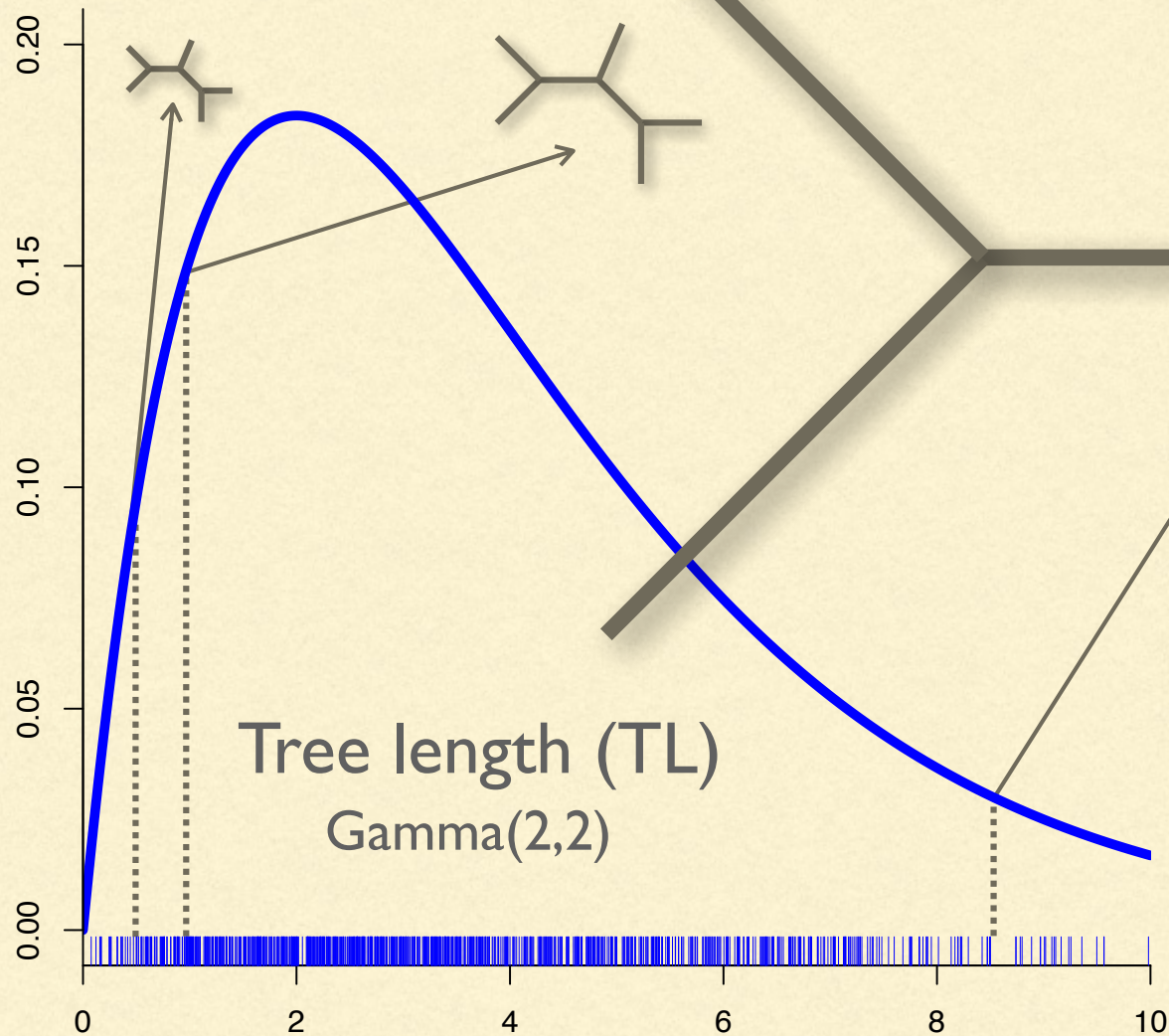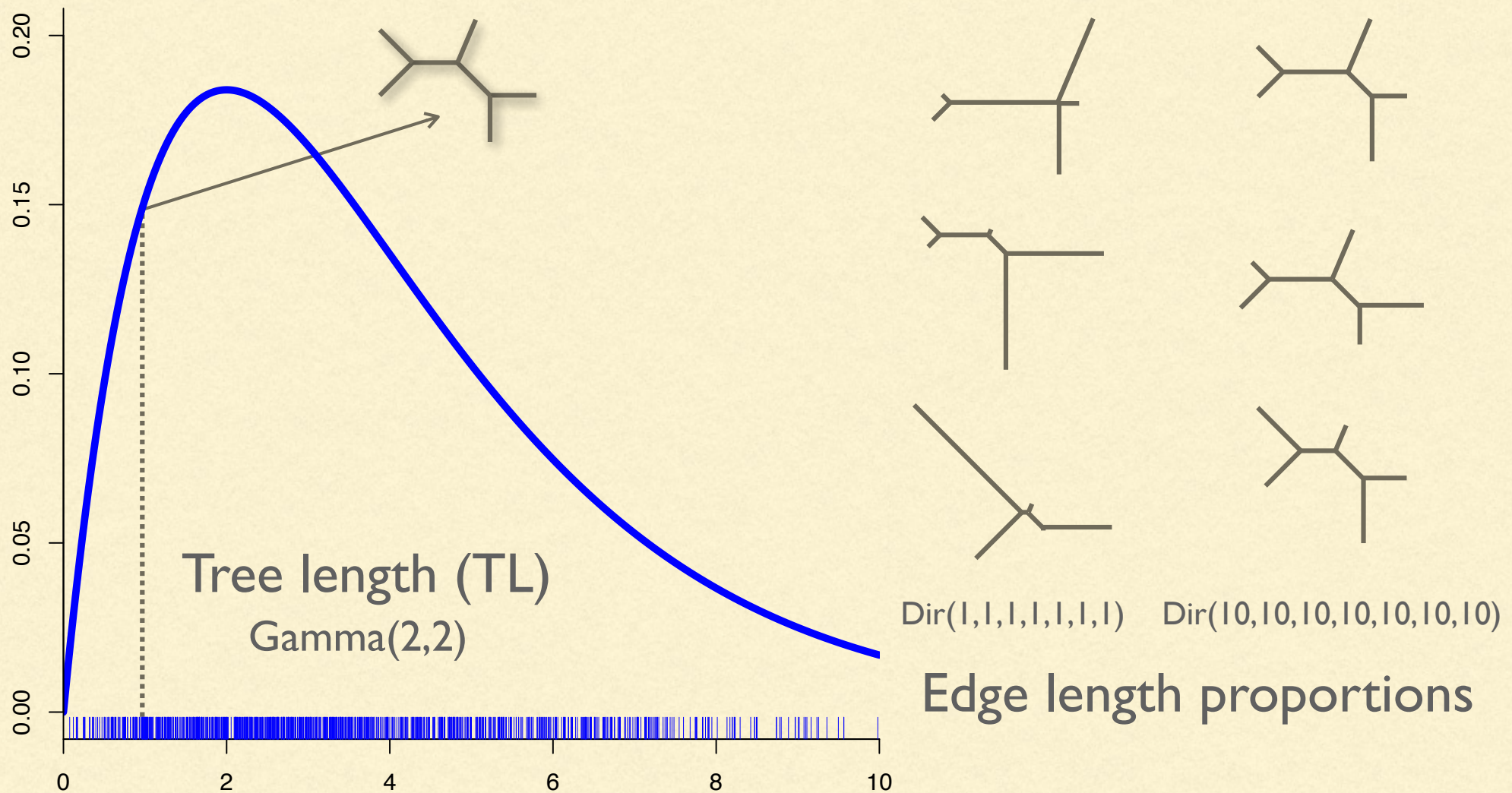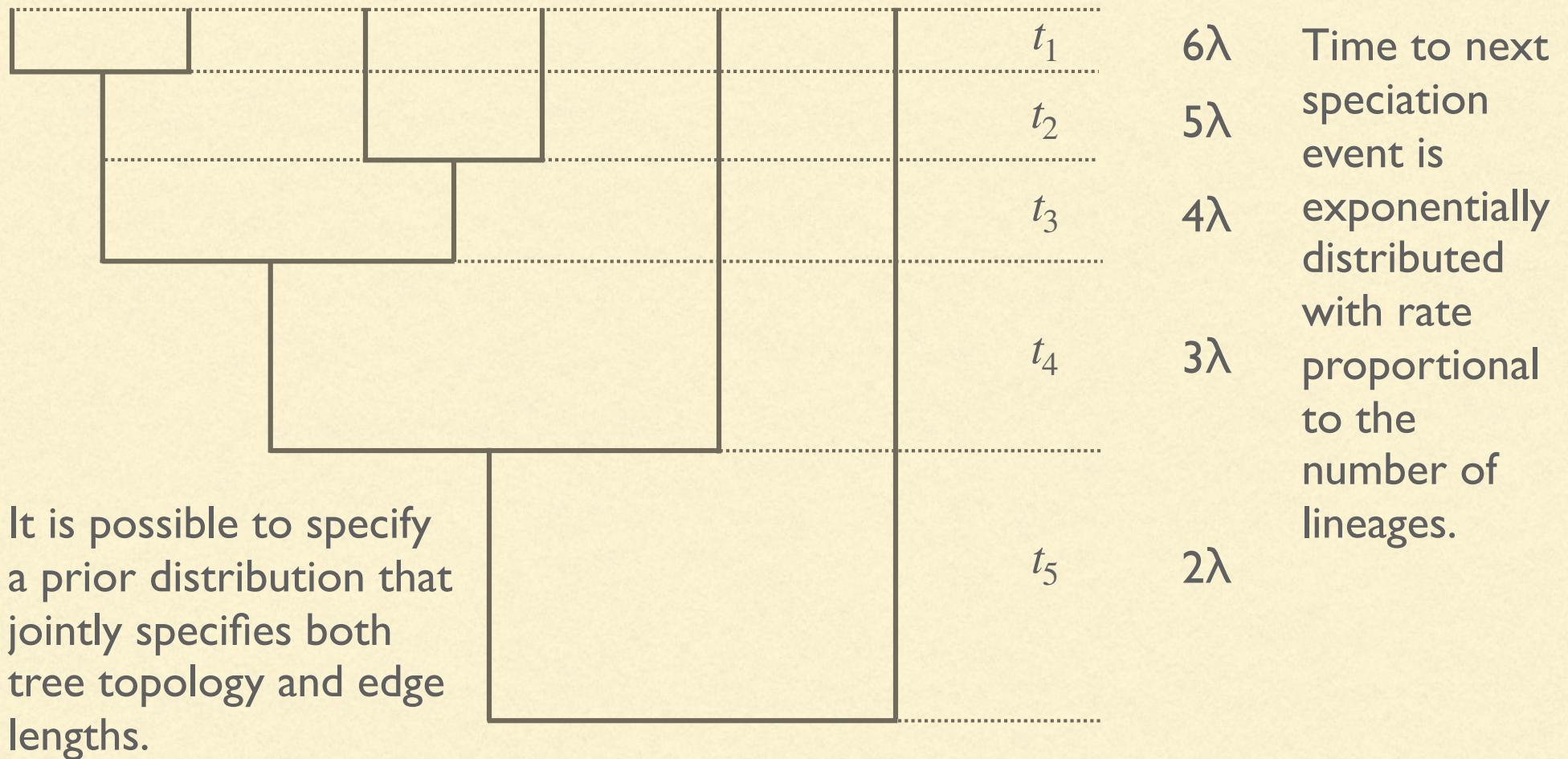
Hierarchical models

Non-hierarchical model

All model parameters can be found in the likelihood function
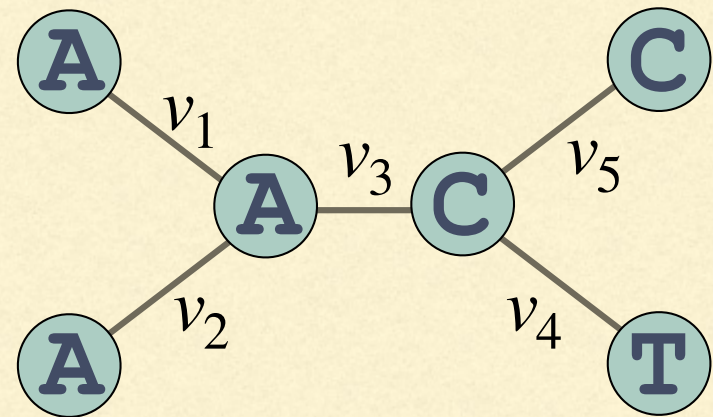
Prior mean 0.1

$$L_k = \frac{1}{4}\left(\frac{1}{4} + \frac{3}{4}e^{-4v_1/3}\right)\left(\frac{1}{4} + \frac{3}{4}e^{-4v_2/3}\right)\left(\frac{1}{4} - \frac{1}{4}e^{-4v_3/3}\right)\left(\frac{1}{4} - \frac{1}{4}e^{-4v_4/3}\right)\left(\frac{1}{4} + \frac{3}{4}e^{-4v_5/3}\right)$$

JC69 likelihood for site k

Hierarchical model

A prior for a
hyperparameter is
a hyperprior
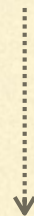
Hyperprior

Some model parameters
(hyperparameters) are
outside the likelihood
function

μ is a
hyperparameter

Prior mean μ

$$L_k = \frac{1}{4}\left(\frac{1}{4} + \frac{3}{4}e^{-4v_1/3}\right)\left(\frac{1}{4} + \frac{3}{4}e^{-4v_2/3}\right)\left(\frac{1}{4} - \frac{1}{4}e^{-4v_3/3}\right)\left(\frac{1}{4} - \frac{1}{4}e^{-4v_4/3}\right)\left(\frac{1}{4} + \frac{3}{4}e^{-4v_5/3}\right)$$

JC69 likelihood for site k

Empirical Bayes

Empirical Bayes approach

Average edge length MLE used as the mean of the prior.

$0.07 = (0.12+0.02+0.07+0.04+0.10)/5$

Prior mean = 0.07

Maximum likelihood estimates (MLEs) of edge lengths

$v_1 = 0.12$

$v_2 = 0.02$

$v_3 = 0.07$

$v_4 = 0.10$

$v_5 = 0.04$

$$L_k = \frac{1}{4}\left(\frac{1}{4}+\frac{3}{4}e^{-4v_1/3}\right)\left(\frac{1}{4}+\frac{3}{4}e^{-4v_2/3}\right)\left(\frac{1}{4}-\frac{1}{4}e^{-4v_3/3}\right)\left(\frac{1}{4}-\frac{1}{4}e^{-4v_4/3}\right)\left(\frac{1}{4}+\frac{3}{4}e^{-4v_5/3}\right)$$

JC69 likelihood for site k

rjMCMC
(reversible-jump MCMC)

Examples of rjMCMC analyses

a.

b.

$$
\begin{array}{c}
\begin{array}{cccc} A & C & G & T \end{array} \\
\begin{array}{c} A \\ C \\ G \\ T \end{array}
\begin{bmatrix}
- & \beta & \beta\kappa & \beta \\
\beta & - & \beta & \beta\kappa \\
\beta\kappa & \beta & - & \beta \\
\beta & \beta\kappa & \beta & -
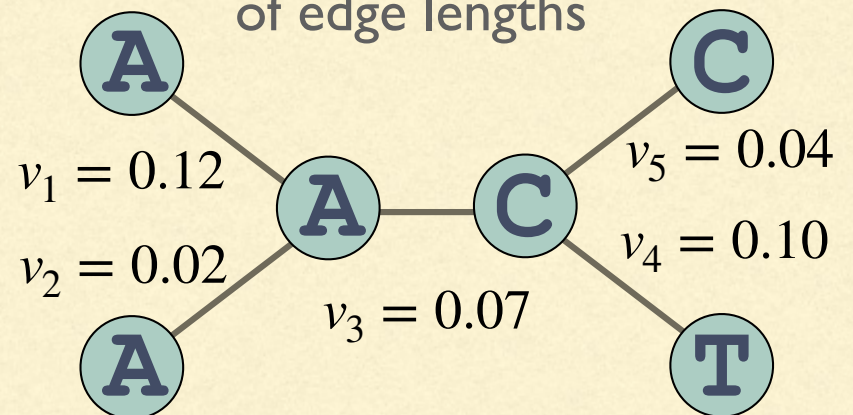\end{bmatrix}
\end{array}
\quad \cdots \quad
\begin{array}{c}
\begin{array}{cccc} A & C & G & T \end{array} \\
\begin{array}{c} A \\ C \\ G \\ T \end{array}
\begin{bmatrix}
- & \pi_C\beta & \pi_G\beta\kappa & \pi_T\beta \\
\pi_A\beta & - & \pi_G\beta & \pi_T\beta\kappa \\
\pi_A\beta\kappa & \pi_C\beta & - & \pi_T\beta \\
\pi_A\beta & \pi_C\beta\kappa & \pi_G\beta & -
\end{bmatrix}
\end{array}
$$

K80 HKY85

$$
\begin{array}{c}
\begin{array}{cccc} A & C & G & T \end{array} \\
\begin{array}{c} A \\ C \\ G \\ T \end{array}
\begin{bmatrix}
- & a\pi_C\beta & b\pi_G\beta & c\pi_T\beta \\
a\pi_A\beta & - & d\pi_G\beta & e\pi_T\beta \\
b\pi_A\beta & d\pi_C\beta & - & f\pi_T\beta \\
c\pi_A\beta & e\pi_C\beta & f\pi_G\beta & -
\end{bmatrix}
\end{array}
\quad \cdots \quad
\begin{array}{c}
\begin{array}{cccc} A & C & G & T \end{array} \\
\begin{array}{c} A \\ C \\ G \\ T \end{array}
\begin{bmatrix}
- & \beta & \beta & \beta \\
\beta & - & \beta & \beta \\
\beta & \beta & - & \beta \\
\beta & \beta & \beta & -
\end{bmatrix}
\end{array}
$$

GTR JC69

c.

5 species   4 species   3 species   2 species   3 species
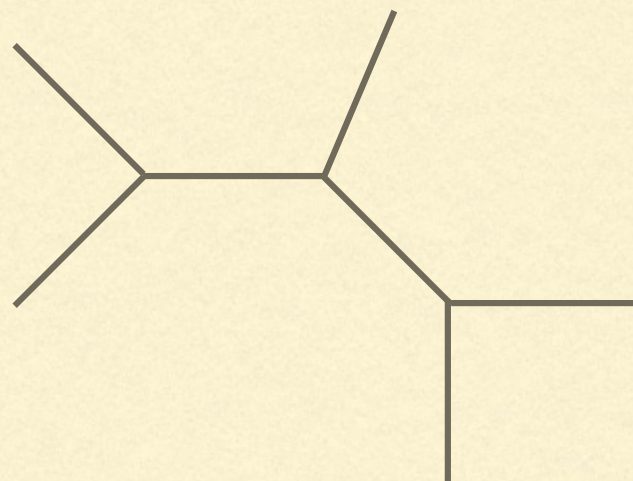
a. Lewis, Holder, & Holsinger. 2005.
   Systematic Biology 54:241–253.

b. Huelsenbeck, Larget, & Alfaro. 2004.
   Molecular Biology and Evolution 21:1123–1133.

c. Rannala & Yang. 2013.
   Genetics 194:245–253.

rjMCMC polytomy model

a. choose 1 of 2 edges
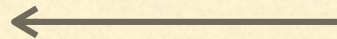   to delete

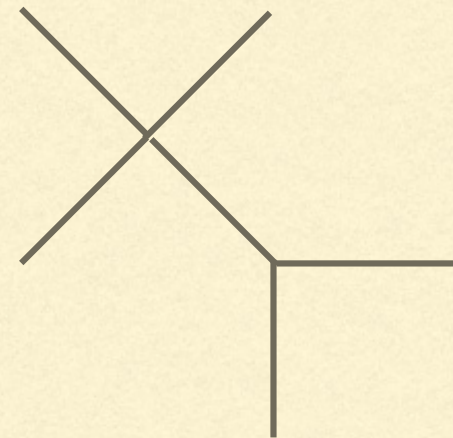**forward**

reverse

a. choose 1 of 1 polytomies
   in which to insert an edge
b. invent a new edge length
   to go with the new edge

7 edge lengths

6 edge lengths

The probability of accepting the forward proposal must take into account the proposal that exactly reverses it to ensure that at equilbrium the chain visits each model with the appropriate frequency

Bayes' factors

Likelihood surface when K80 true

Simulated data for 2 taxa

sequence length = 500 sites
true branch length = 0.15
true kappa = **5.0**

Assume joint prior density
is uniform over the area
shown and zero elsewhere.

K80 model
(entire 2d
space)

**K80 has better
average fit
than JC69**

JC69 model (just this line)

15

10

5

κ

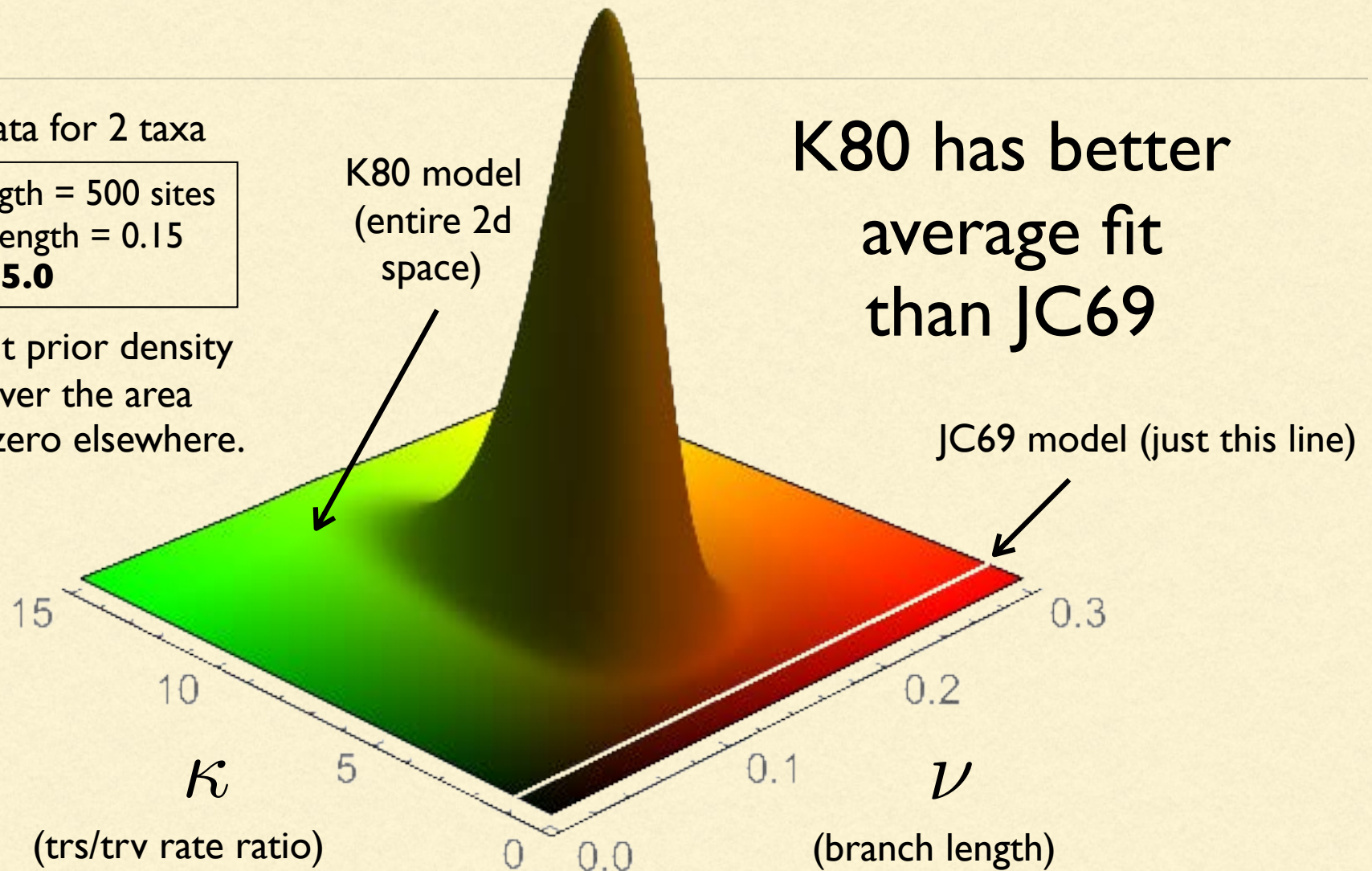(trs/trv rate ratio)

0

0.0

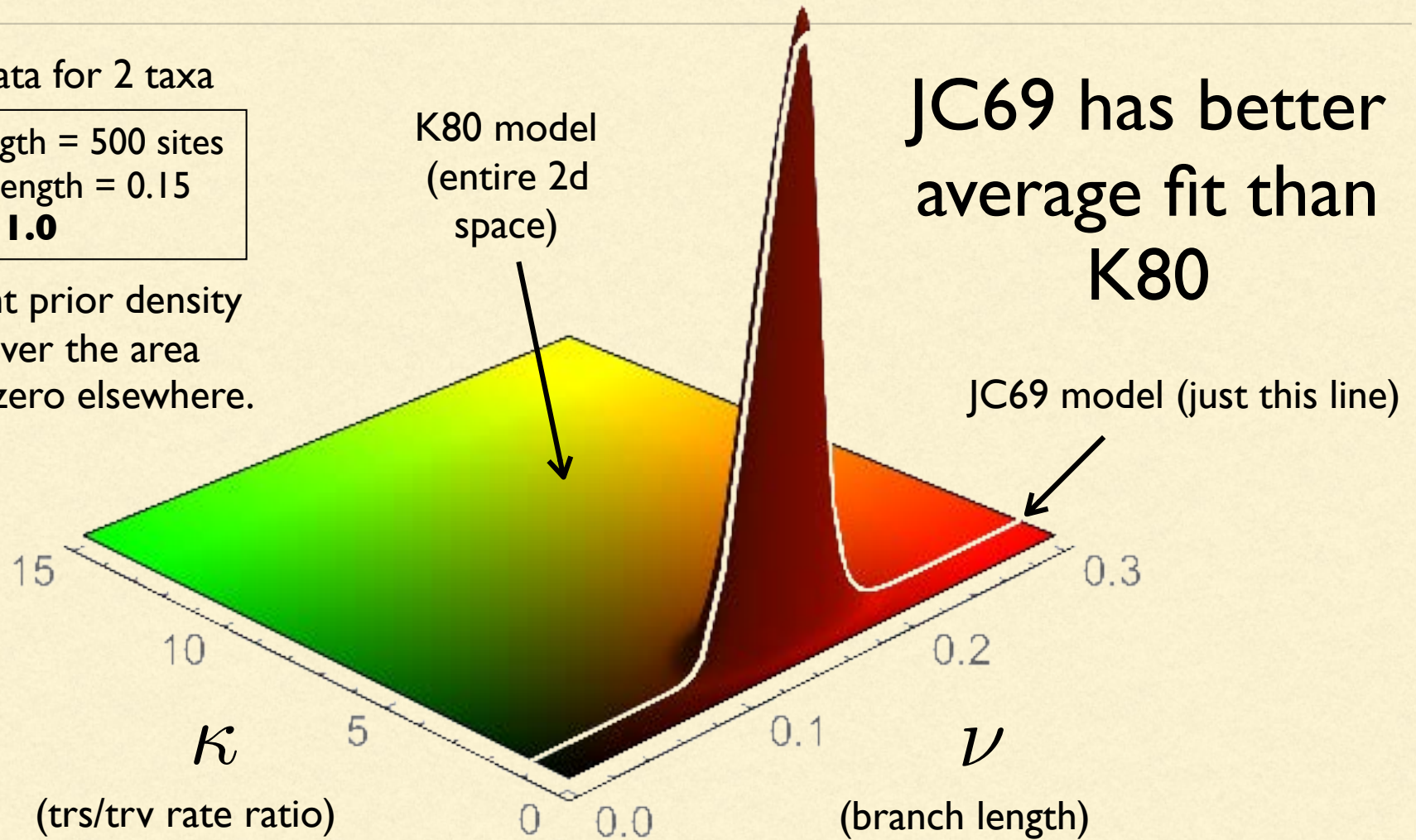0.1

0.2

0.3

ν

(branch length)

Likelihood surface when JC69 true

Simulated data for 2 taxa

sequence length = 500 sites
true branch length = 0.15
true kappa = **1.0**

Assume joint prior density is uniform over the area shown and zero elsewhere.

K80 model (entire 2d space)

JC69 has better average fit than K80

JC69 model (just this line)

κ

(trs/trv rate ratio)

ν

(branch length)

15
10
5
0

0.3
0.2
0.1
0.0

Marginal likelihood and model fit

Bayes' rule: $p(\theta \,|\, D, M) = \dfrac{p(D \,|\, \theta, M)\, p(\theta \,|\, M)}{p(D \,|\, M)}$

MCMC does not require $p(D|M)$.

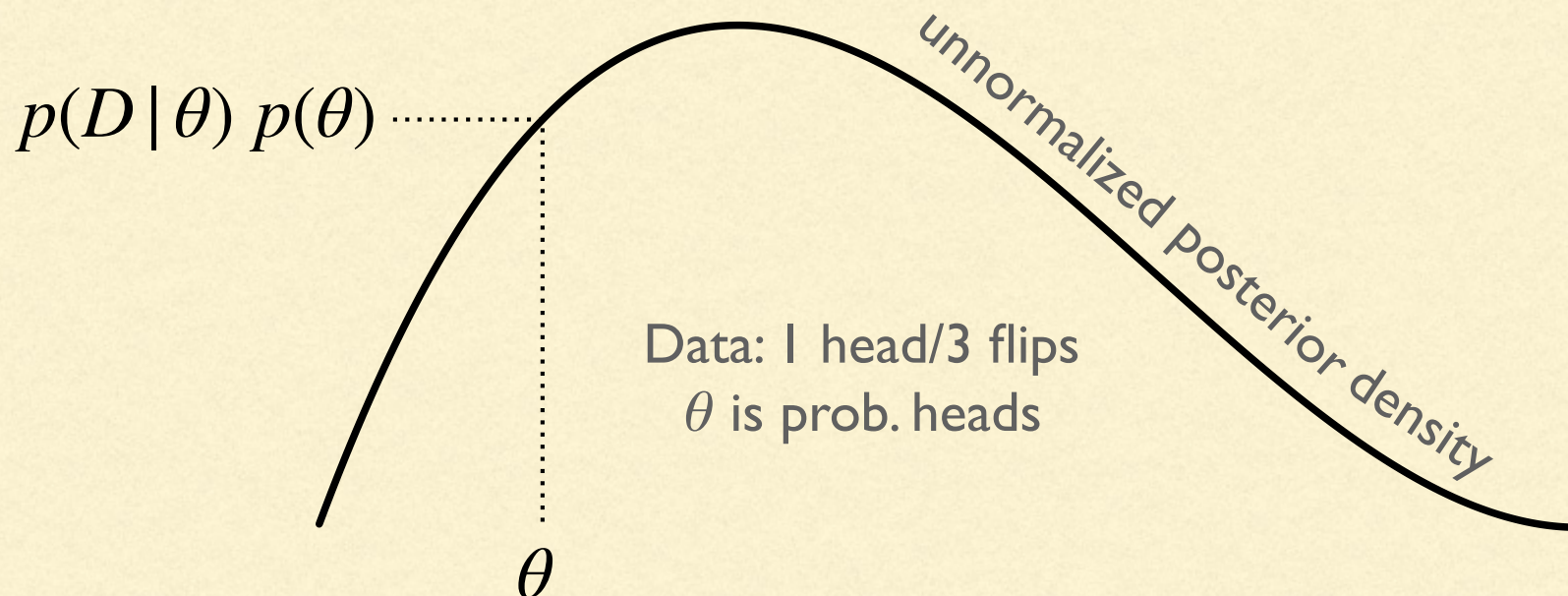$p(D|M)$ measures the average fit of model M to data D and is thus very useful for model choice.

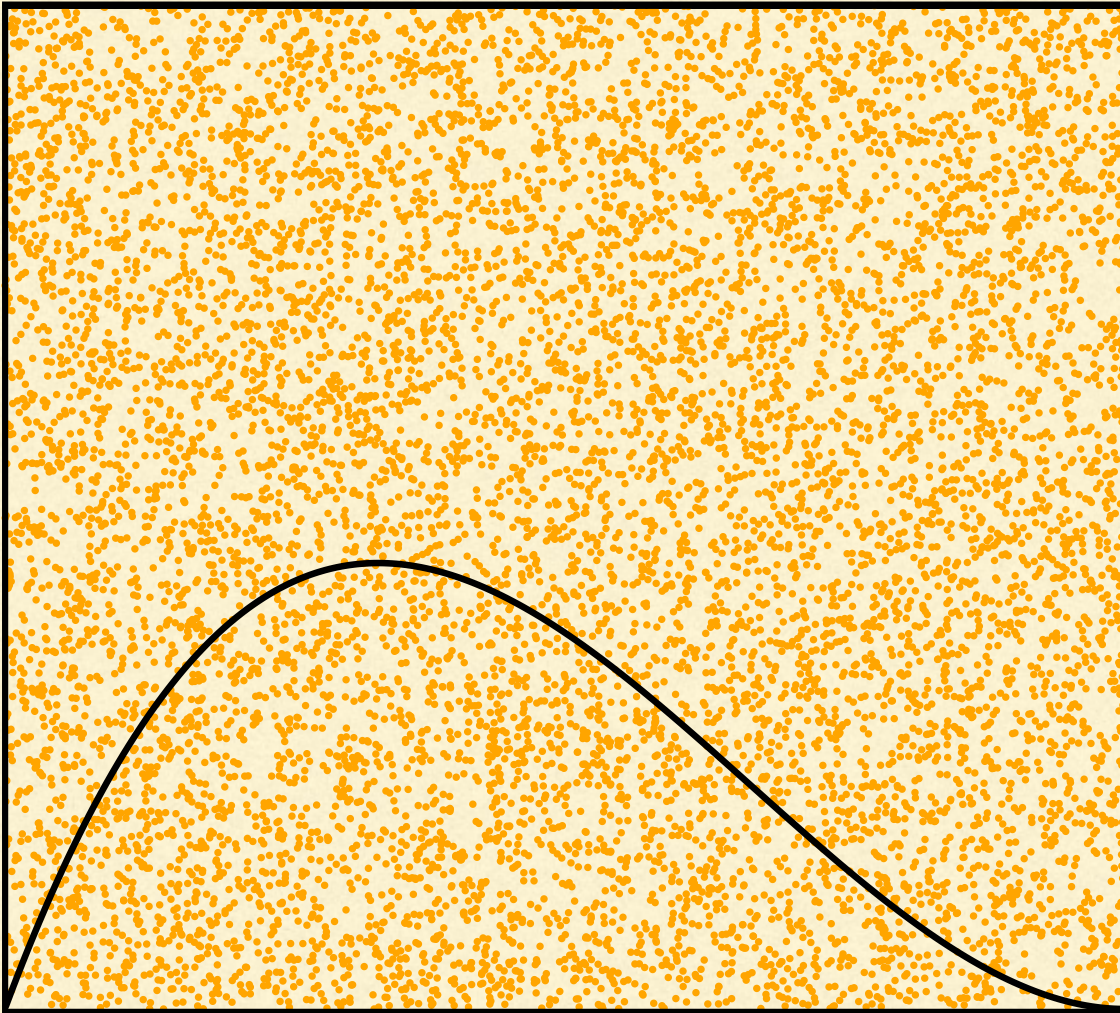So we need to estimate $p(D|M)$ after all in order to choose amongst competing models.

Marginal likelihood estimation

$$p(D) = \int p(D\,|\,\theta)\,p(\theta)\,d\theta$$

Estimating *p(D)* is equivalent to estimating the area under the curve whose height is, for every value of θ, equal to *p(D|θ) p(θ)*

$p(D\,|\,\theta)\,p(\theta)$ ....

unnormalized posterior density

Data: 1 head/3 flips
θ is prob. heads

θ

Simplest method

Sample evenly from a box with known area A that completely encloses the curve.

Area under the curve is just A times the fraction of sampled points that lie under the curve.

2567 = under curve

10000 = inside box
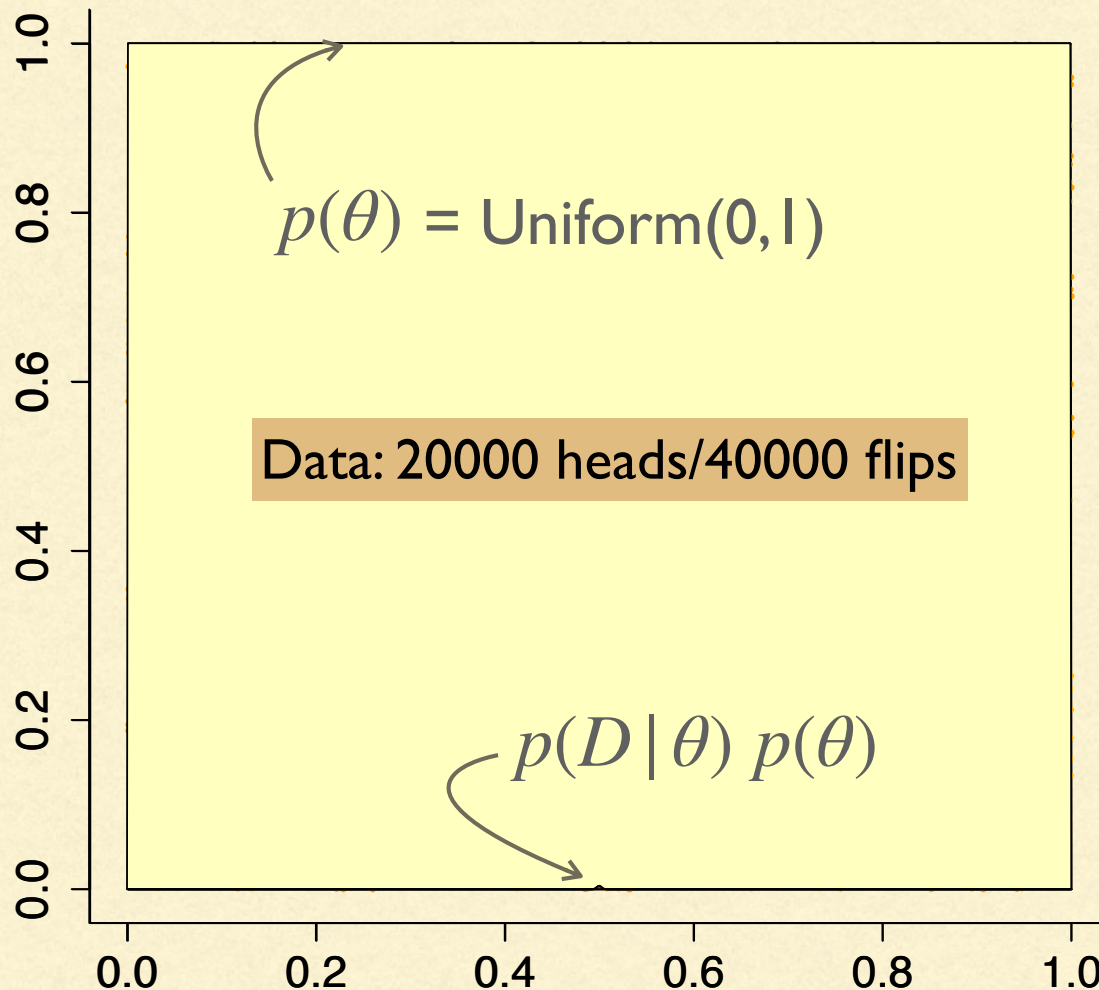
1.0 = area of box

0.2567 = estimated value

0.2500 = true value

With much more data, things get complicated

$p(\theta) = \text{Uniform}(0,1)$

Data: 20000 heads/40000 flips
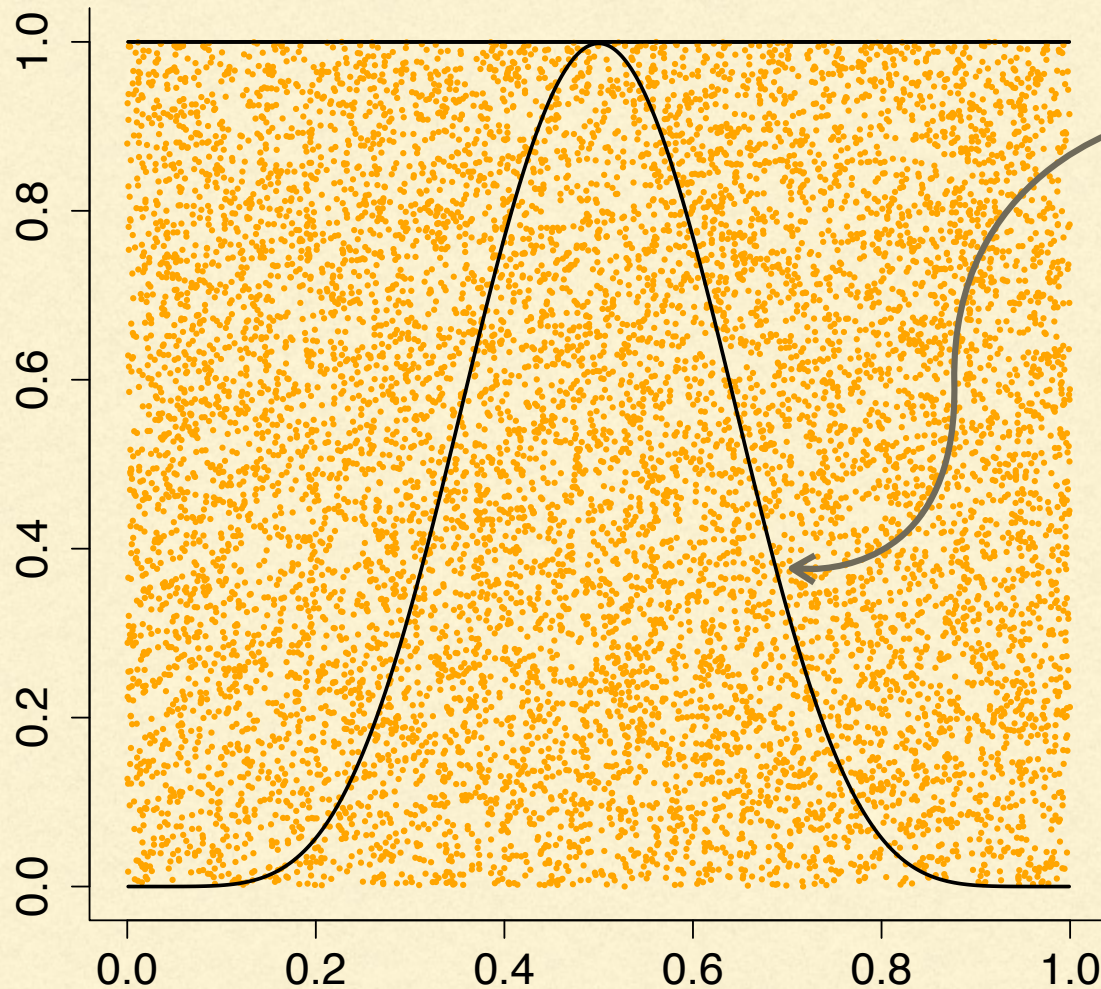
$p(D \mid \theta)\, p(\theta)$

The prior density can serve as our box. It has area 1.0 and is everywhere higher than the posterior *kernel* (likelihood times prior density).

Problem: with lot of data, posterior kernel is minuscule compared to prior.

In fact, of 10000 darts thrown at the prior, 0 fell beneath the posterior kernel (that little bump along the bottom)
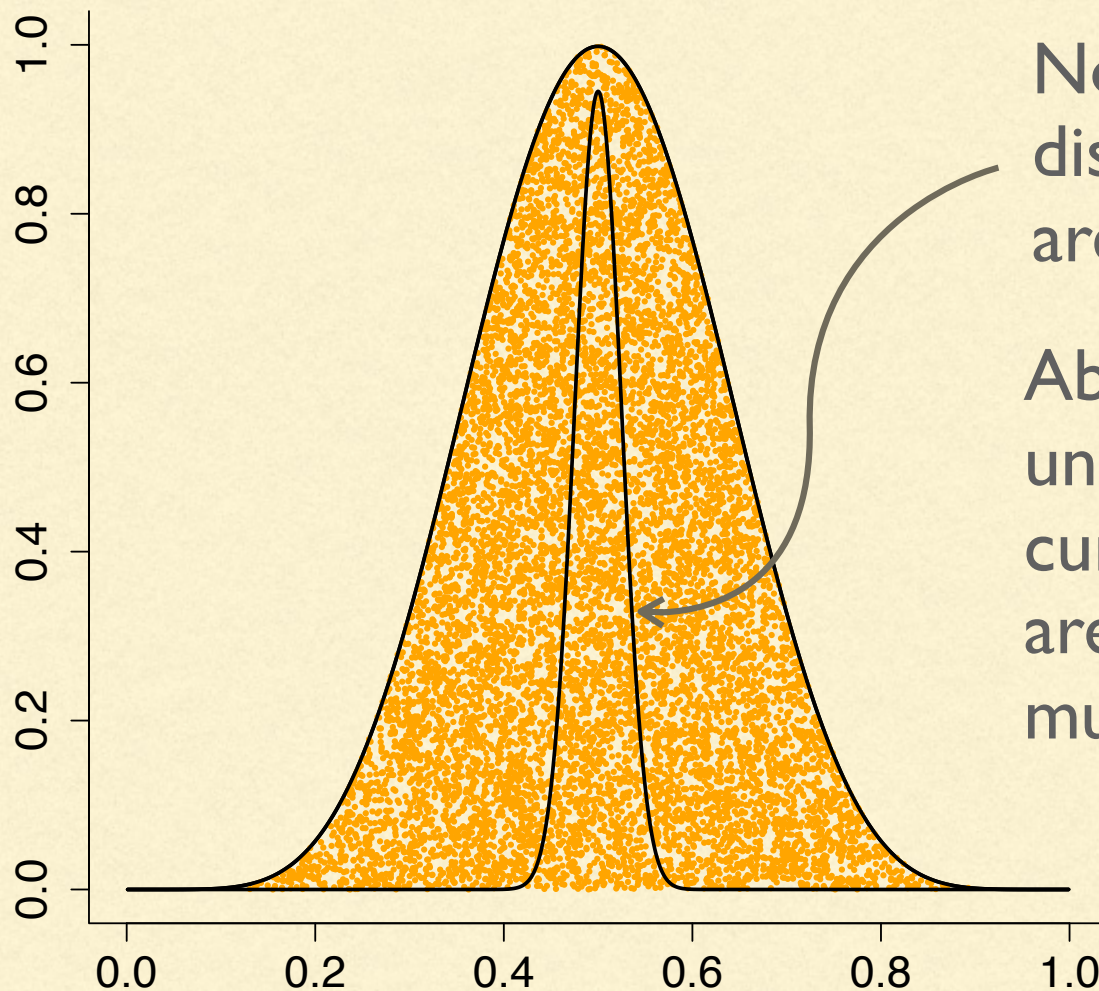
Solution, create stepping stones

Start by estimating the area under this curve (also everywhere higher than the posterior kernel).

About **33%** of the darts thrown at the larger box (which has area **1.0**) fall under the inner curve, so the area under the inner curve must be about **0.33**
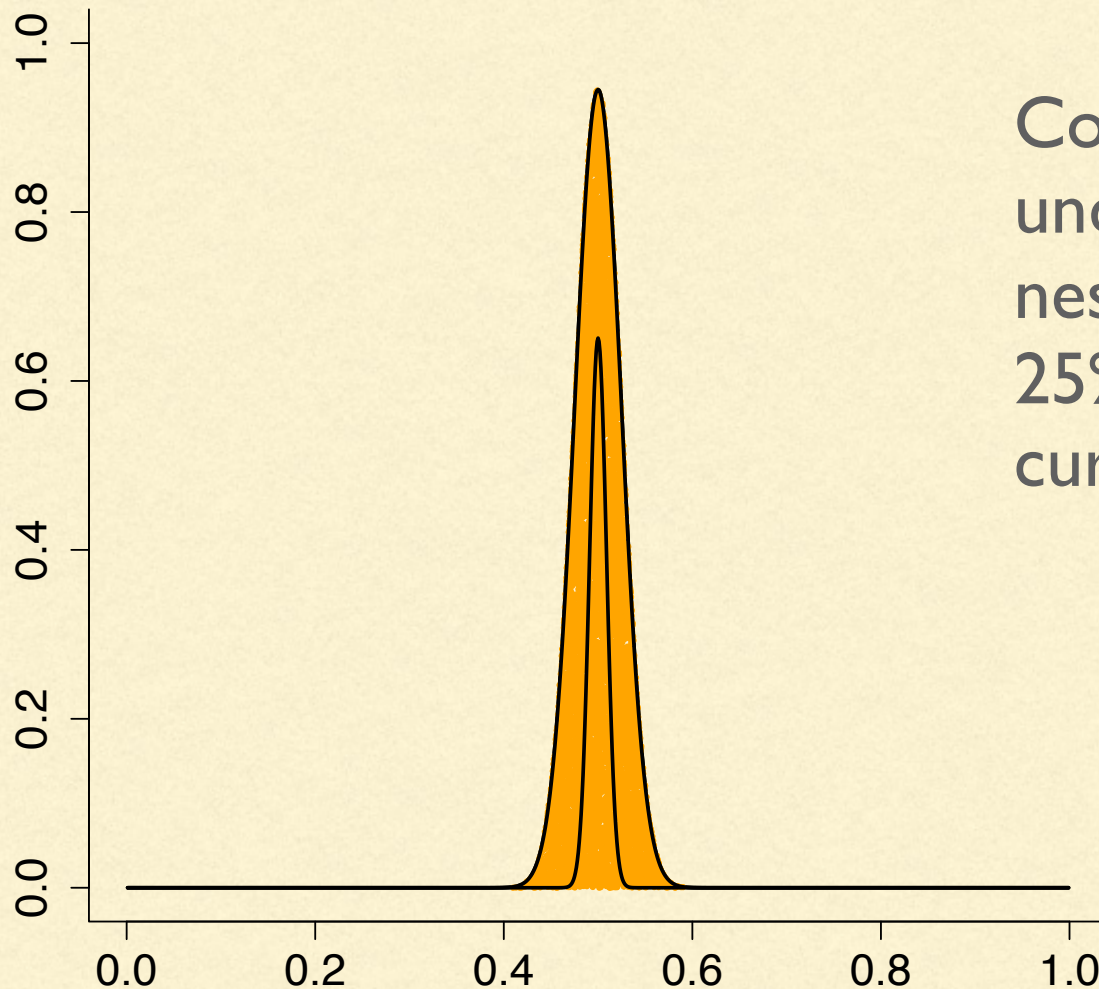
Solution, create stepping stones

Now throw darts at the inner distribution and estimate the area under this curve

About 5.8% of the points fall under the inner, nested curve, so we know that the area under that nested curve must be
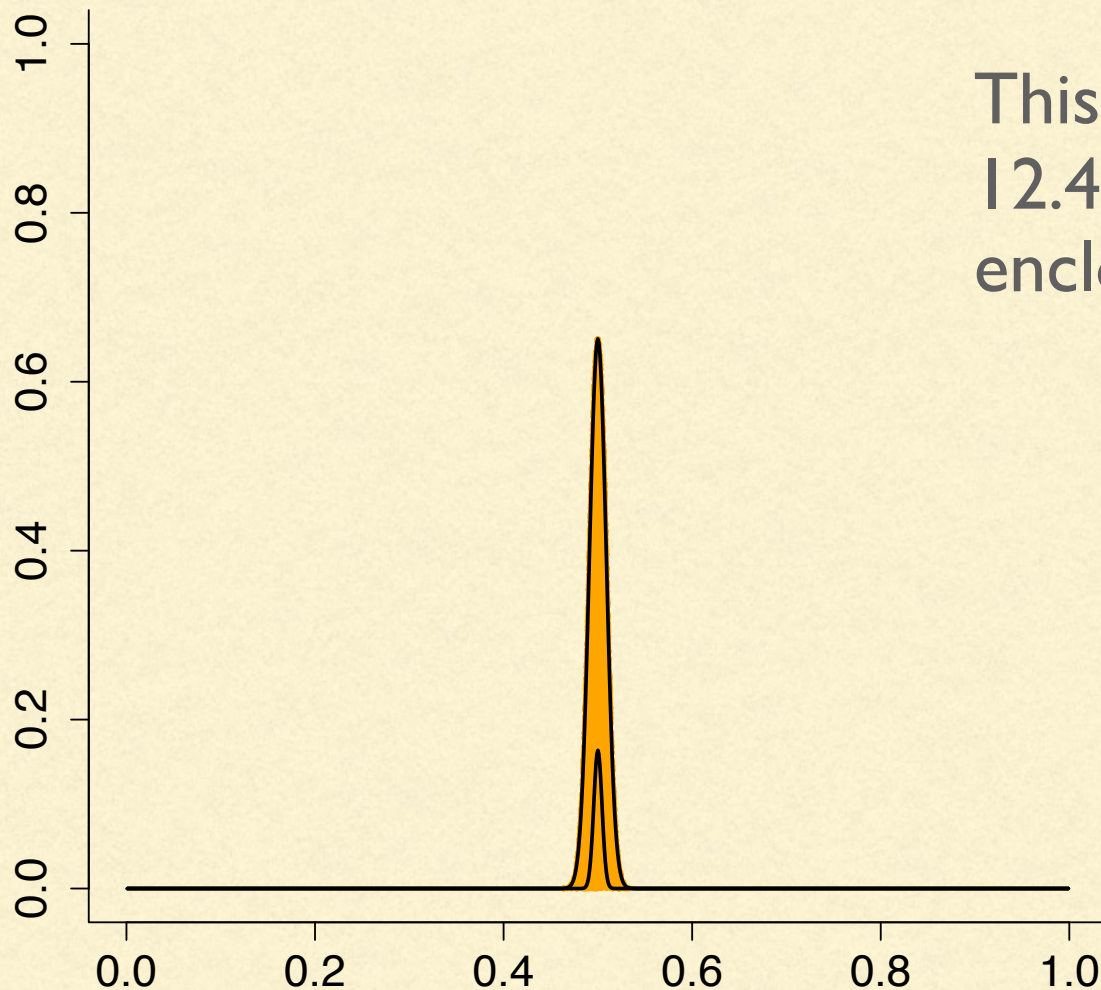
$$(1.0)(0.330)(0.058) = 0.019$$

Solution, create stepping stones

Continue estimating the area under nested curves. The nested one here has about 25% of the are of the outer curve.
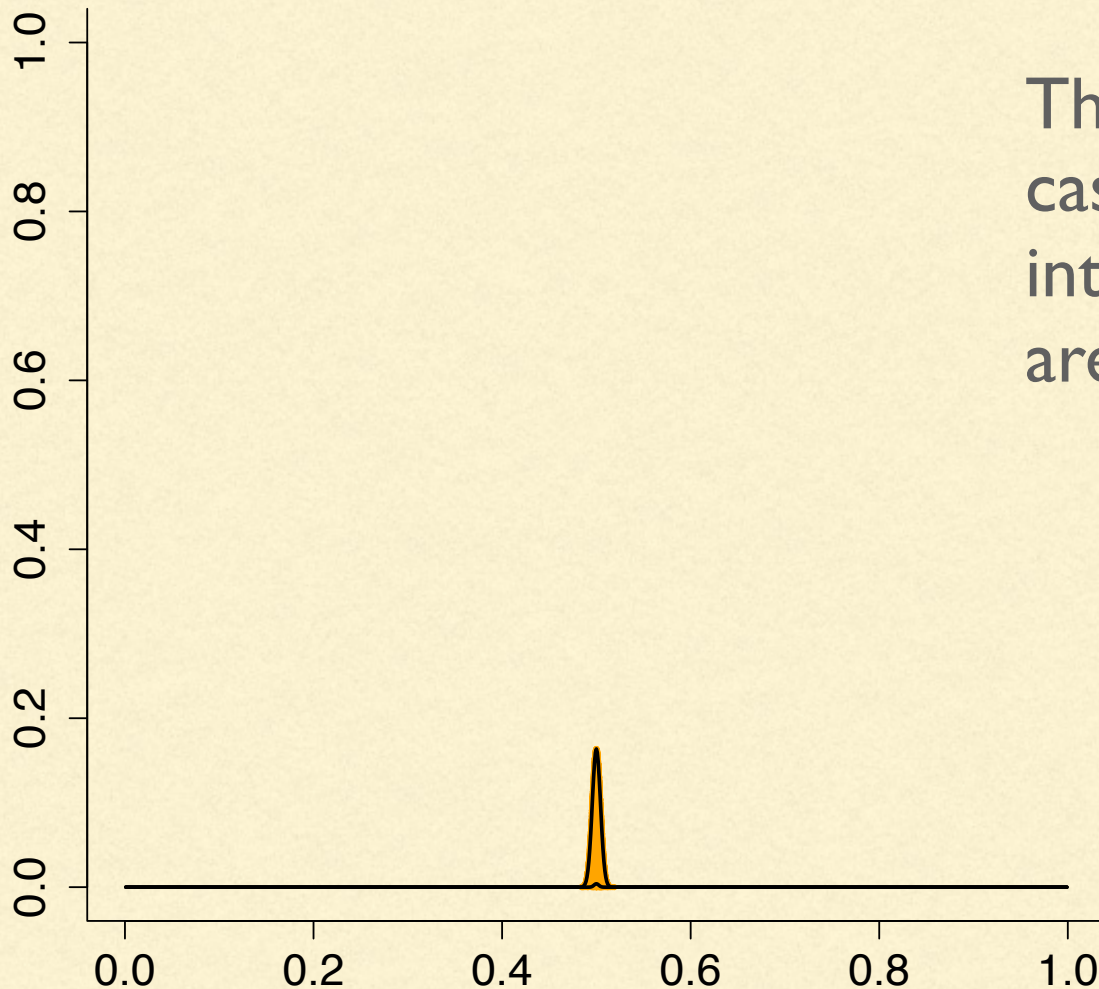
Solution, create stepping stones

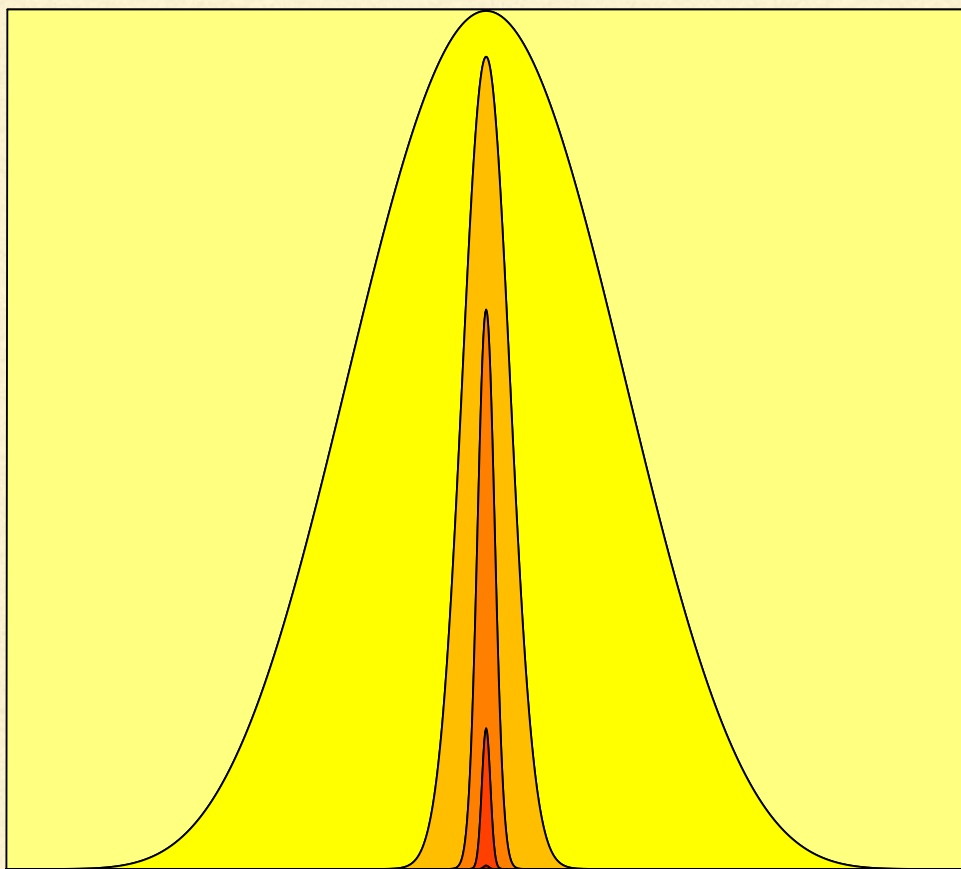This nested curve has about 12.4% of the area of its enclosing curve

Solution, create stepping stones

The nested curve is, in this case, equal to the curve of interest, and has 1.4% of the area of its encosing curve

Solution, create stepping stones

We've broken up the estimation of the area under the posterior kernel into a series of 5 area estimations:

$$\frac{a}{g} = \left(\frac{a}{\cancel{b}}\right)\left(\frac{\cancel{b}}{\cancel{c}}\right)\left(\frac{\cancel{c}}{\cancel{d}}\right)\left(\frac{\cancel{d}}{\cancel{f}}\right)\left(\frac{\cancel{f}}{g}\right)$$

a = area under posterior kernel

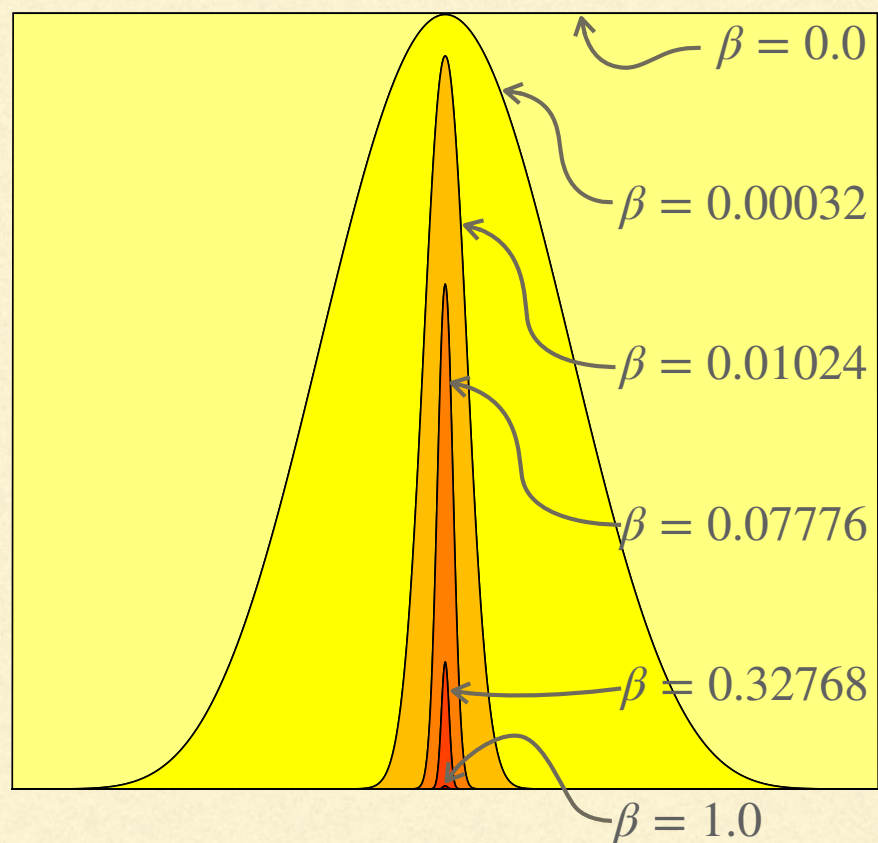g = area under the prior = 1.0

0.0000251 estimated
0.0000250 true

Stepping-stone method

How can we choose the intermediate density functions that serve as stepping stones?

Power posterior kernel:

$$p(D\,|\,\theta)^{\beta}\,p(\theta)$$

$\beta = 1$ posterior kernel

$0 < \beta < 1$ intermediate

$\beta = 0$ prior density

Bayes factor for M_0 against M_1

$$B_{01} = \frac{p(D \mid M_0)}{p(D \mid M_1)} = \frac{\text{marginal likelihood for model } M_0}{\text{marginal likelihood for model } M_1}$$

Normally, both marginal likelihoods are on the log scale, so usually it is the (natural) log Bayes factor that is reported:

$$\log B_{01} = \log p(D \mid M_0) - \log p(D \mid M_1)$$