Introduction to Bayesian Inference & MCMC in RevBayes

Tracy A. Heath

Ecology, Evolution, & Organismal Biology
Iowa State University

Bayesian Phylogenetics & Macroevolution in RevBayes Workshop
Canberra, Australia
March 26-28, 2019

Bayesian or Maximum Likelihood?

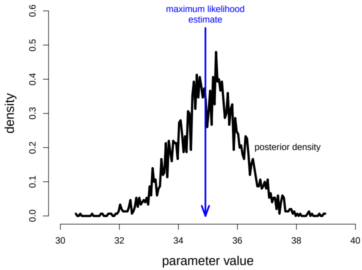
What's the difference?

Bayesian:

- estimates $f(\theta \mid \mathcal{D})$
- estimates a distribution
- parameters are random variables
- average over nuisance parameters

Maximum Likelihood:

- maximizes $f(\mathcal{D} \mid \theta)$
- point estimate
- parameters are fixed/unknown
- optimize nuisance parameters



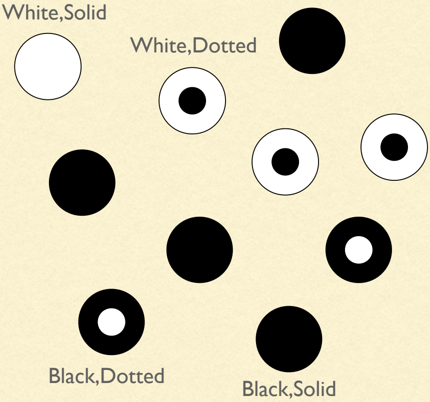
Joint Probabilities

Let's start with joint probability and the simple example that Paul Lewis gives in his lecture on Bayesian phylogenetics

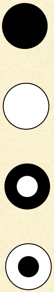
Slides source: https://molevol.mbl.edu/index.php/Paul_Lewis

(photo of Paul Lewis by David Hillis)

Joint probabilities



10 marbles in a bag
Sampling with replacement



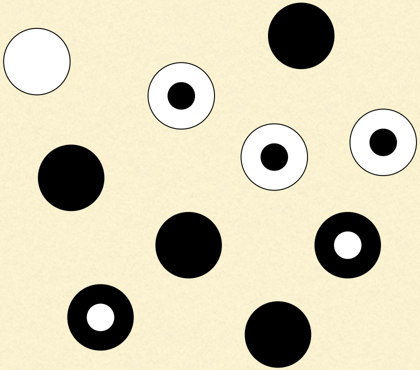
$\Pr(B,S) = 0.4$

$\Pr(W,S) = 0.1$

$\Pr(B,D) = 0.2$

$\Pr(W,D) = 0.3$

Conditional probabilities



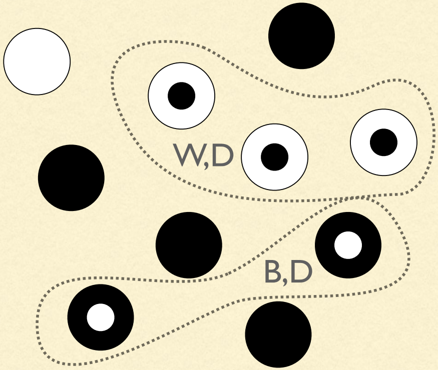
What's the probability that a marble is black given that it is dotted?

5 marbles satisfy the condition (D)

$$\Pr(B|D) = \frac{2}{5}$$

2 remaining marbles are black (B)

Marginal probabilities



Marginalizing over color yields the total probability that a marble is dotted (D)

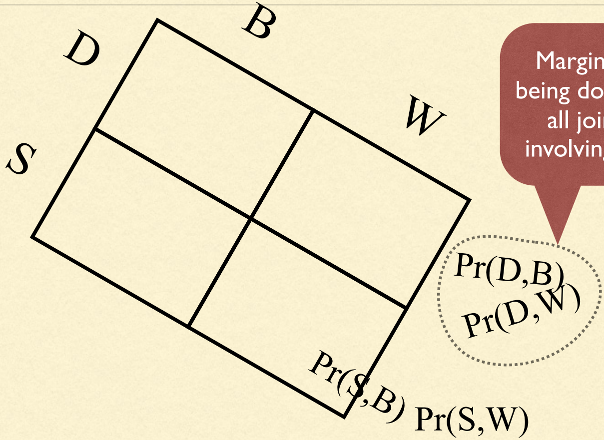
$$\Pr(\mathbf{D}) = \Pr(B,\mathbf{D}) + \Pr(W,\mathbf{D})$$
$$= 0.2 + 0.3$$
$$= 0.5$$

Marginalization involves summing all joint probabilities containing D

Marginalization

	B	W
D	Pr(D,B)	Pr(D,W)
S	Pr(S,B)	Pr(S,W)

Marginalizing over colors

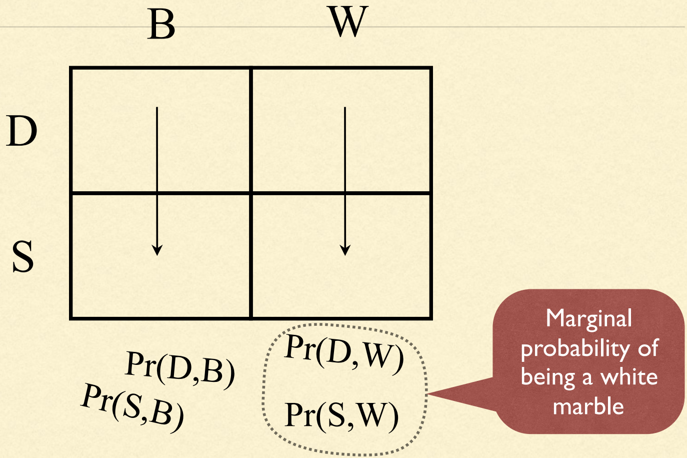


Marginal probability of being dotted is the sum of all joint probabilities involving dotted marbles

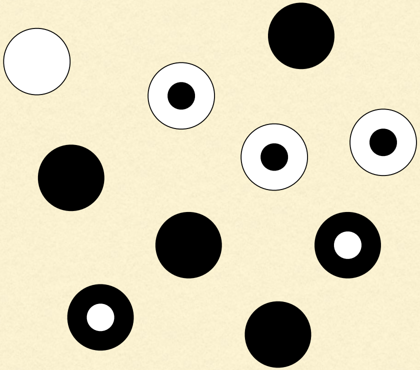
Joint probabilities

	B	W
D	Pr(D,B)	Pr(D,W)
S	Pr(S,B)	Pr(S,W)

Marginalizing over "dottedness"



Bayes' rule



The joint probability Pr(B,D) can be written as the product of a *conditional probability* and the *probability of that condition*

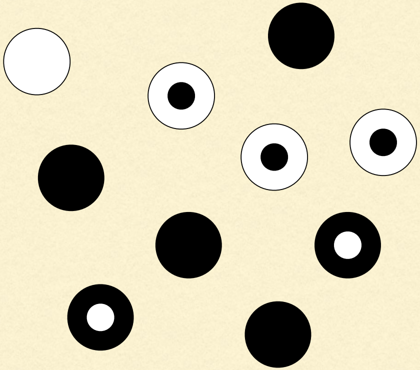
Pr(B|D) Pr(D)

Pr(B,D)

Either B or D can be the condition

Pr(D|B) Pr(B)

Bayes' rule



Equate the two ways of writing Pr(B,D)

$$\Pr(B|D)\,\Pr(D) = \Pr(D|B)\,\Pr(B)$$

Divide both sides by Pr(D)

$$\frac{\Pr(B|D)\,\cancel{\Pr(D)}}{\cancel{\Pr(D)}} = \frac{\Pr(D|B)\,\Pr(B)}{\Pr(D)}$$

Bayes' rule

$$\Pr(B|D) = \frac{\Pr(D|B)\,\Pr(B)}{\Pr(D)}$$

Bayes' rule (variations)

$$\Pr(B|D) = \frac{\Pr(D|B)\Pr(B)}{\Pr(D)}$$

$$= \frac{\Pr(D|B)\Pr(B)}{\Pr(B,D) + \Pr(W,D)}$$

Pr(*D*) is the **marginal probability** of being dotted
To compute it, we **marginalize over colors**

Bayes' rule (variations)

$$\Pr(B|D) = \frac{\Pr(D|B)\Pr(B)}{\Pr(B,D) + \Pr(W,D)}$$

$$= \frac{\Pr(D|B)\Pr(B)}{\Pr(D|B)\Pr(B) + \Pr(D|W)\Pr(W)}$$

$$= \frac{\Pr(D|B)\Pr(B)}{\sum_{\theta\in\{B,W\}}\Pr(D|\theta)\Pr(\theta)}$$

Bayes' rule in statistics

Likelihood of hypothesis θ

Prior probability of hypothesis θ

$$\Pr(\theta|D) = \frac{\Pr(D|\theta)\,\Pr(\theta)}{\sum_\theta \Pr(D|\theta)\,\Pr(\theta)}$$

Posterior probability of hypothesis θ

Marginal probability of the data (marginalizing over hypotheses)

Bayesian Inference

Estimate the **probability** of a hypothesis (model) conditional on observed data.

The probability represents the researcher's **degree of belief**.

Bayes' Rule (also called Bayes Theorem) specifies the conditional probability of the hypothesis given the data.

Bayes' Rule

The posterior probability of a discrete parameter δ conditional on the data D is

$$\Pr(\delta \mid D) = \frac{\Pr(D \mid \delta)\Pr(\delta)}{\sum_{\delta}\Pr(D \mid \delta)\Pr(\delta)}$$

$\sum_{\delta}\Pr(D \mid \delta)\Pr(\delta)$ is the likelihood **marginalized** over all possible values of δ.

Bayes' Rule

The posterior probability **density** a continuous parameter θ conditional on the data D is

$$f(\theta \mid D) = \frac{f(D \mid \theta) f(\theta)}{\int_\theta f(D \mid \theta) f(\theta) d\theta}$$

$\int_\theta f(D \mid \theta) f(\theta) d\theta$ is the likelihood **marginalized** over all possible values of θ.

Priors

Priors distributions are an important part of Bayesian statistics

The the distribution of θ before any data are collected is the prior

$$f(\theta)$$

The prior describes your uncertainty in the parameters of your model

Priors

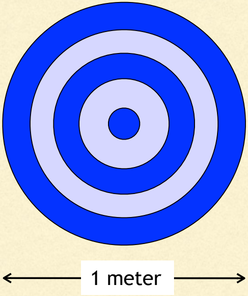
Paul Lewis gives a clear example of a prior in action...

Slides source: https://molevol.mbl.edu/index.php/Paul_Lewis

(photo of Paul Lewis by David Hillis)

If you had to guess...

Photo by Tracy Heath

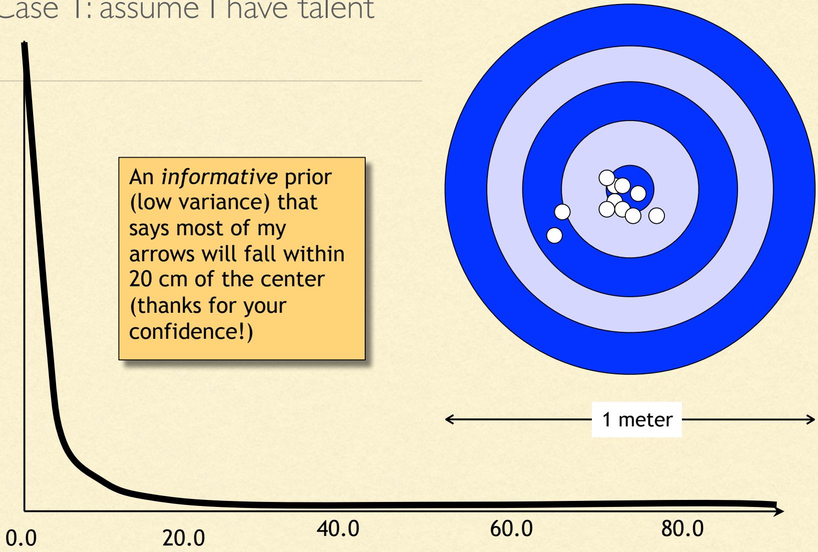


Not knowing anything about my archery abilities, draw a curve representing your view of the chances of my arrow landing a distance *d* centimeters from the center of the target.

0.0

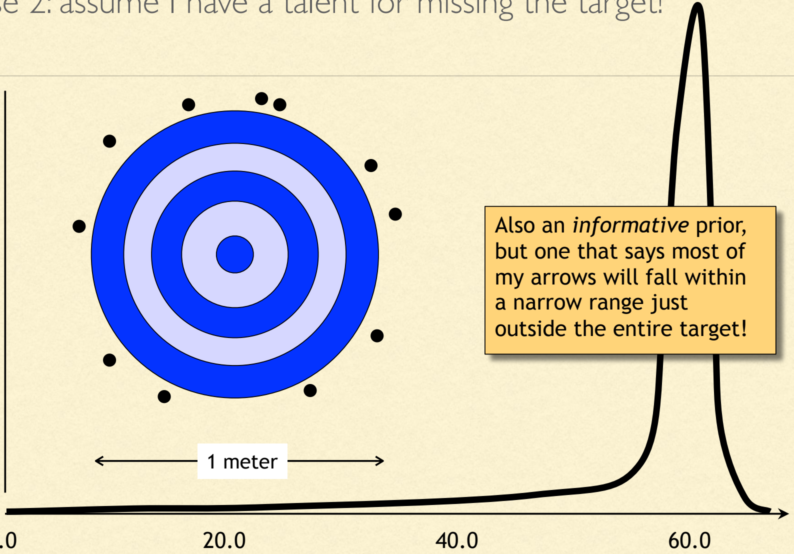
d (centimeters from target center)

Case 1: assume I have talent



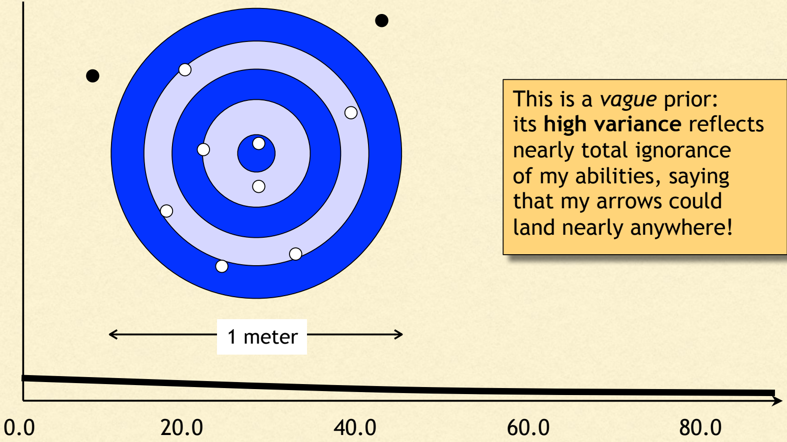
An *informative* prior (low variance) that says most of my arrows will fall within 20 cm of the center (thanks for your confidence!)

Case 2: assume I have a talent for missing the target!



Also an *informative* prior, but one that says most of my arrows will fall within a narrow range just outside the entire target!

Case 3: assume I have no talent



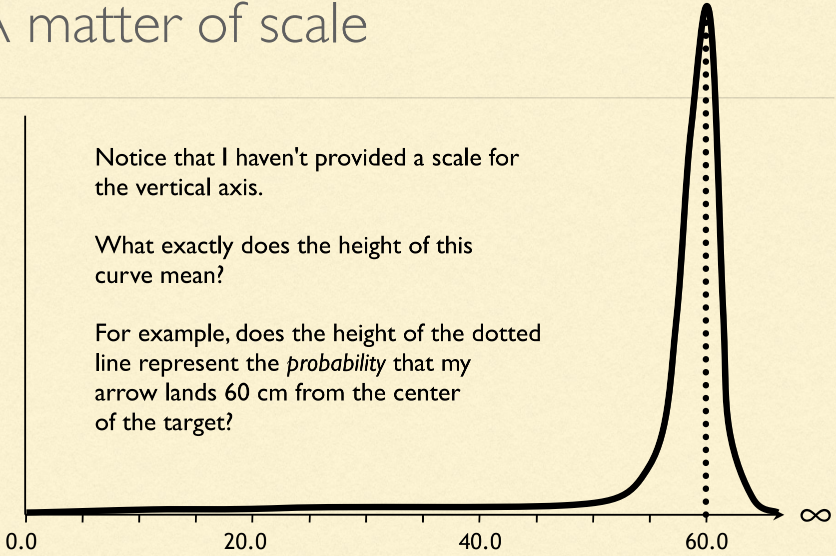
This is a *vague* prior:
its **high variance** reflects
nearly total ignorance
of my abilities, saying
that my arrows could
land nearly anywhere!

A matter of scale

Notice that I haven't provided a scale for the vertical axis.

What exactly does the height of this curve mean?

For example, does the height of the dotted line represent the *probability* that my arrow lands 60 cm from the center of the target?

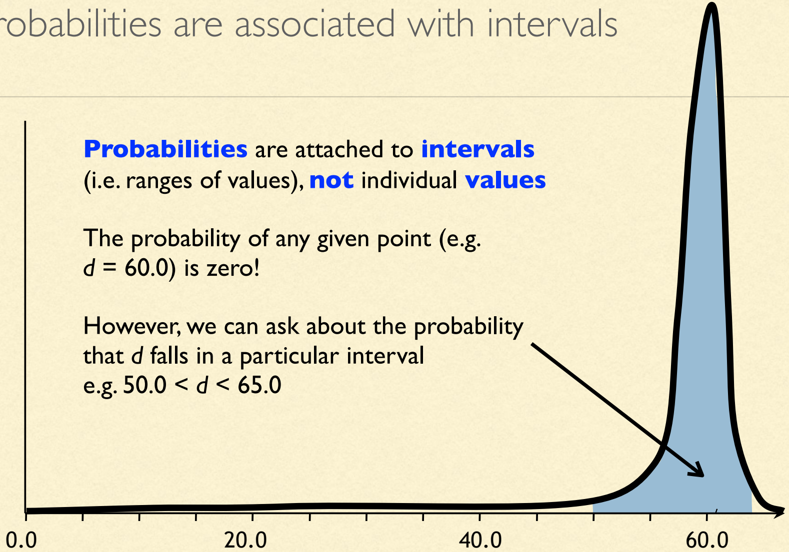


Probabilities are associated with intervals

Probabilities are attached to **intervals** (i.e. ranges of values), **not** individual **values**

The probability of any given point (e.g. d = 60.0) is zero!

However, we can ask about the probability that d falls in a particular interval e.g. $50.0 < d < 65.0$

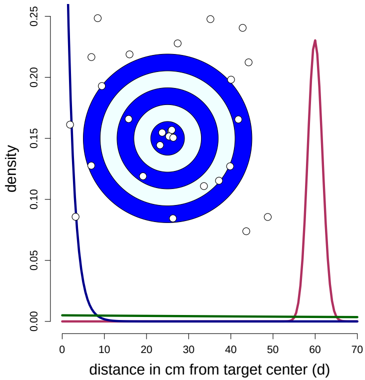


Priors: Archery Example

Let's continue with the archery example: we may assume a gamma-prior distribution on my archery skill (distance from bullseye $= d$) with a shape parameter α and a rate parameter β.

$$d \sim \text{Gamma}(\alpha, \beta)$$

$$f(d \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} d^{\alpha-1} e^{-\frac{d}{\beta}}$$



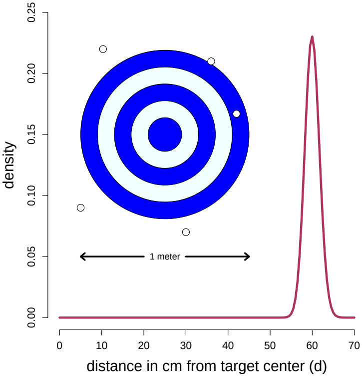
This requires us to assign values for α and β based on our **prior belief**

Priors: Archery Example

Let's assume that I will consistently miss the target, this corresponds to a gamma distribution with a mean (m) of 60 and a variance (v) of 3.

mean = accuracy

variance = precision



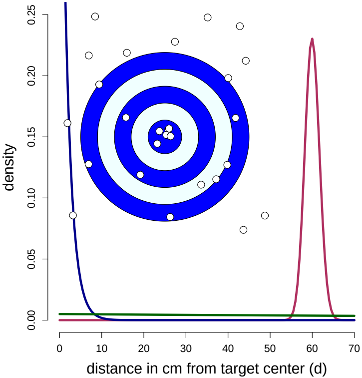
Priors: Archery Example

If we have some prior knowledge of the mean (m) and variance (v) of the gamma distribution, we can compute α and β.

$$m = \frac{\alpha}{\beta}, \quad \alpha = \frac{m^2}{v}$$

$$v = \frac{\alpha}{\beta^2}, \quad \beta = \frac{m}{v}$$

$$d \sim \text{Gamma}(\alpha, \beta)$$



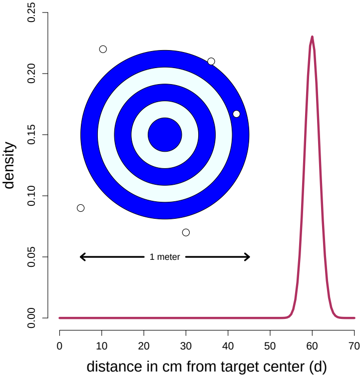
Priors: Archery Example

Let's assume that I will consistently miss the target, this corresponds to a gamma distribution with a mean (m) of 60 and a variance (v) of 3.

$$\alpha = \frac{60^2}{3} = 1200$$

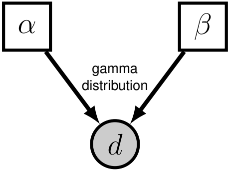
$$\beta = \frac{60}{3} = 20$$

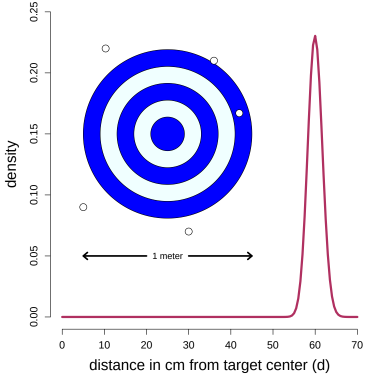
$$d \sim \text{Gamma}(\alpha, \beta)$$



Priors: Archery Example

Another way of expressing $d \sim \text{Gamma}(\alpha, \beta)$ is with a **probabilistic graphical model**

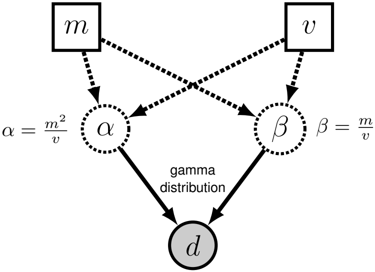


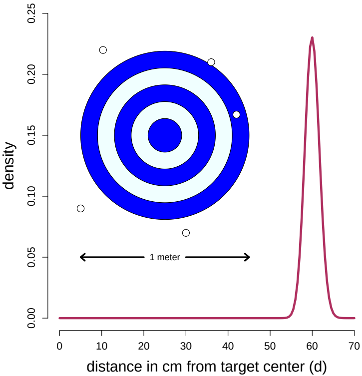


This shows that our observed datum (d = a single observed shot) is conditionally dependent on the shape (α) and rate (β) of the gamma distribution.

Priors: Archery Example

We can parameterize the model using the mean (m) and variance (v), where α and β are computed using m and v.



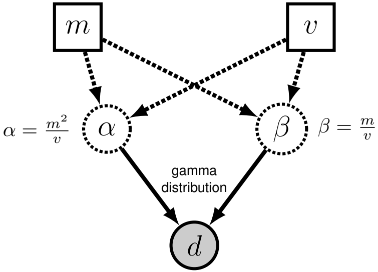


Sometimes it's better to use alternative parameterization. We may have more intuition about mean and variance than we have about shape and rate.

Priors: Archery Example

If somehow we happened to **know** the true mean and variance of my accuracy at the archery range, we can easily calculate the likelihood of any observed shot:

$$f(d \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} d^{\alpha-1} e^{-\frac{d}{\beta}}$$

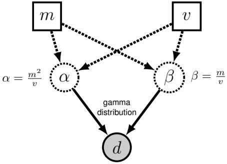


$$f(d = 39.76 \mid \alpha = 1200, \beta = 20) = 7.89916e - 40$$

RevBayes Demo: Archery Accuracy

RevBayes

Fully integrative Bayesian inference of phylogenetic parameters using probabilistic graphical models and an interpreted language



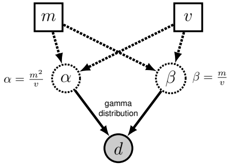
http://RevBayes.com

Höhna, Landis, Heath, Boussau, Lartillot, Moore, Huelsenbeck, Ronquist. 2016. RevBayes: **Bayesian phylogenetic inference using graphical models and an interactive model-specification language.** *Systematic Biology*. (doi: 10.1093/sysbio/syu021)

Graphical Models in RevBayes

Graphical models provide tools for visually & computationally representing complex, parameter-rich probabilistic models

We can depict the conditional dependence structure of various parameters and other random variables



Höhna, Heath, Boussau, Landis, Ronquist, Huelsenbeck. 2014. **Probabilistic Graphical Model Representation in Phylogenetics**. *Systematic Biology*. (doi: 10.1093/sysbio/syu039)

RevBayes Demo: Model on Archery Skill

The Rev language calculating the probability of 1 data observation (`observed_shot`) given a mean and variance.

```
mean <- 60
var <- 3

alpha := (mean * mean) / var
beta := mean / var

observed_shot = 39.76

d ~ dnGamma(alpha,beta)
d.clamp(observed_shot)

d.lnProbability()

-90.0366
```

Example: Model on Archery Skill

What if we **do not know** m and v?

We can use maximum likelihood or Bayesian methods to estimate their values.

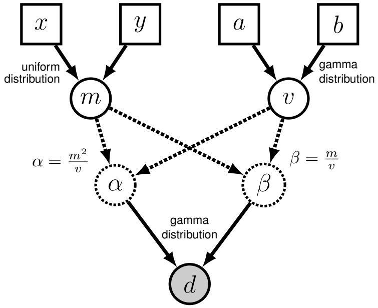
Maximum likelihood methods require us to find the values of m and v that **maximize** $f(d \mid m, v)$.

Bayesian methods use **prior distributions** to describe our uncertainty in m and v and estimate $f(m, v \mid d)$.

Example: Hierarchical Archery Model

We must define prior distributions for m and v to account for uncertainty and estimate the posterior densities of those parameters

Now x and y are the parameters of the uniform prior distribution on m and a and b are the shape and rate parameters of the gamma prior distribution on v.

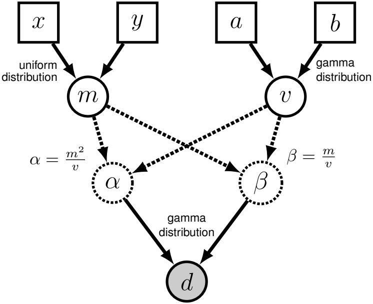


Example: Hierarchical Archery Model

The values we choose for the parameters of these hyperprior distributions should reflect our prior knowledge. The previous observed shot was **39.76 cm**, thus we may use this to parameterize our hyperpriors for analysis of future observations.

$$m \sim \text{Uniform}(x, y)$$
$$x = 10.0$$
$$y = 50.0$$
$$\mathbb{E}(m) = 30.0$$

$$v \sim \text{Gamma}(a, b)$$
$$a = 20.0$$
$$b = 2.0$$
$$\mathbb{E}(v) = 10.0$$



RevBayes Demo: Hierarchical Archery Model

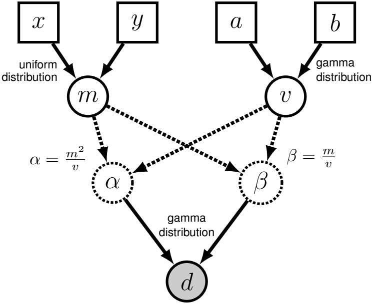
The Rev language specifying a hierarchical model on shot accuracy based on 1 new observation.

```
mean ~ dnUnif(10,50)
var ~ dnGamma(20,2)

alpha := (mean * mean) / var
beta := mean / var

observed_shot = 35.21

d ~ dnGamma(alpha,beta)
d.clamp(observed_shot)

d.lnProbability()

depends on initial value of mean & var
```

Example: Hierarchical Archery Model

Now that we have a defined model, how do we estimate the posterior probability density?

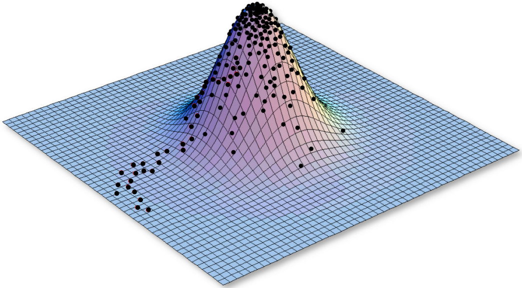
$$m \sim \text{Uniform}(x, y)$$
$$v \sim \text{Gamma}(a, b)$$
$$d \sim \text{Gamma}(\alpha, \beta)$$



$$f(m, v \mid d, a, b, x, y) \propto f(d \mid \alpha = \frac{m^2}{v}, \beta = \frac{m}{v}) f(m \mid x, y) f(v \mid a, b)$$

Markov Chain Monte Carlo (MCMC)

An algorithm for approximating the posterior distribution



Metropolis, Rosenblusth, Rosenbluth, Teller, Teller. 1953. Equations of state calculations by fast computing machines. *J. Chem. Phys.*

Hastings. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika.*

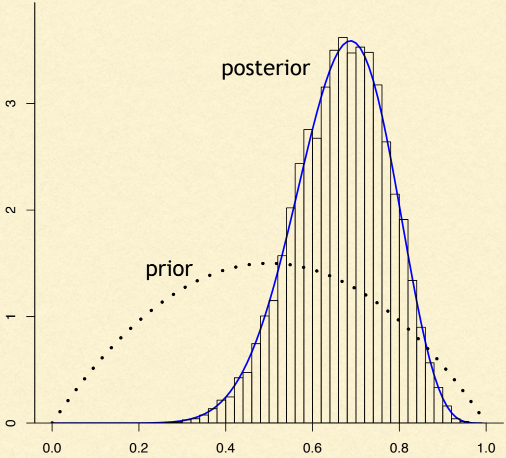
Markov Chain Monte Carlo (MCMC)

More on MCMC from Paul Lewis and his lecture on Bayesian phylogenetics

Slides source: https://molevol.mbl.edu/index.php/Paul_Lewis

(photo of Paul Lewis by David Hillis)

Markov chain Monte Carlo (MCMC)

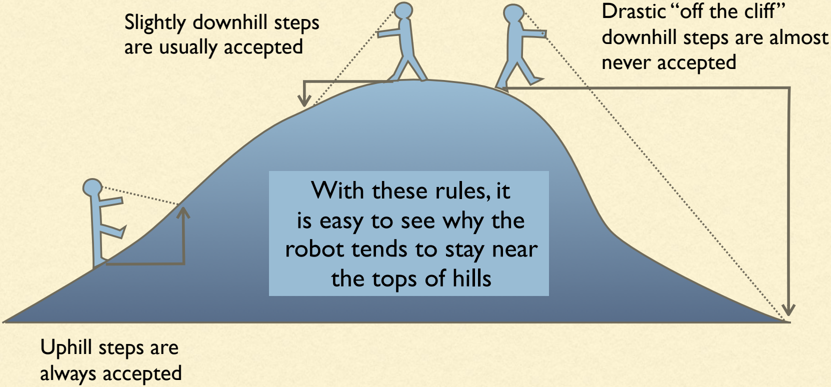


For more complex problems, we might settle for a

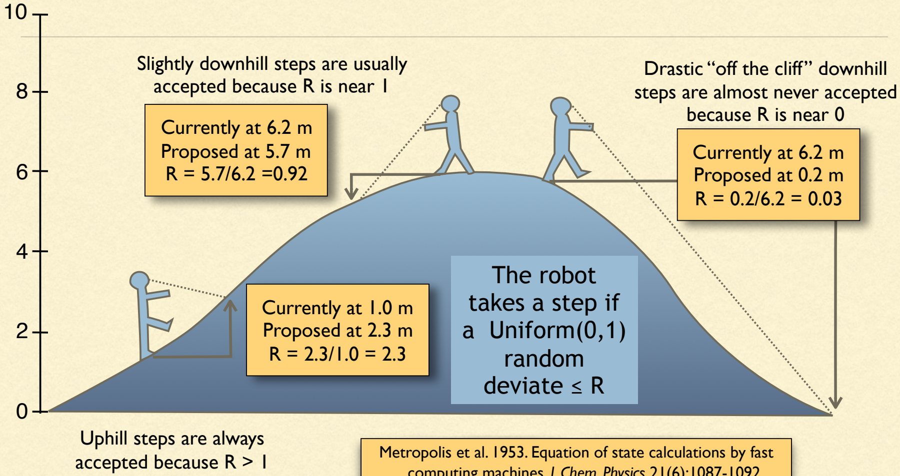
good approximation

to the posterior distribution

MCMC robot's rules



Actual rules (Metropolis algorithm)



Metropolis et al. 1953. Equation of state calculations by fast computing machines. *J. Chem. Physics* 21(6):1087-1092.

Cancellation of marginal likelihood

When calculating the ratio (R) of posterior densities, the marginal probability of the data cancels.

$$\frac{p(\theta^* \mid D)}{p(\theta \mid D)} = \frac{\dfrac{p(D \mid \theta^*)\, p(\theta^*)}{\cancel{p(D)}}}{\dfrac{p(D \mid \theta)\, p(\theta)}{\cancel{p(D)}}} = \frac{p(D \mid \theta^*)\, p(\theta^*)}{p(D \mid \theta)\, p(\theta)}$$

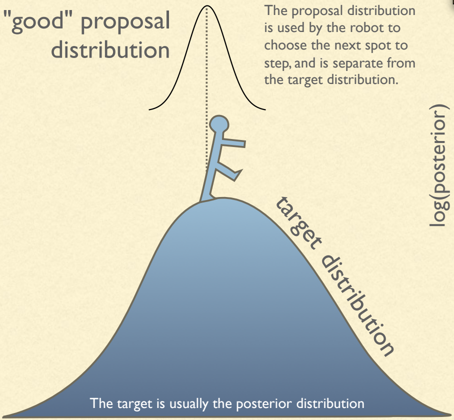
Posterior odds

Apply Bayes' rule to both top and bottom

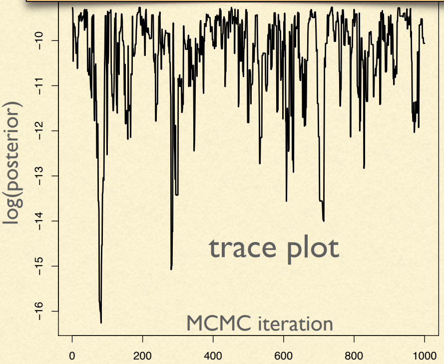
Likelihood ratio

Prior odds

Target vs. Proposal Distributions

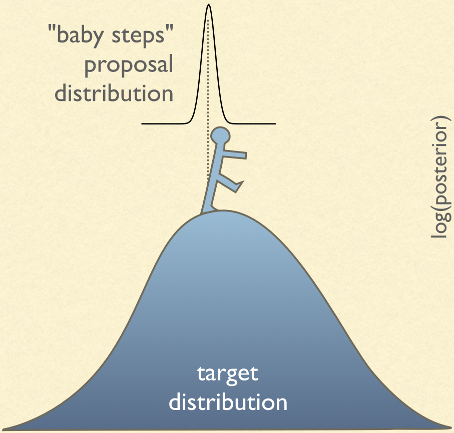


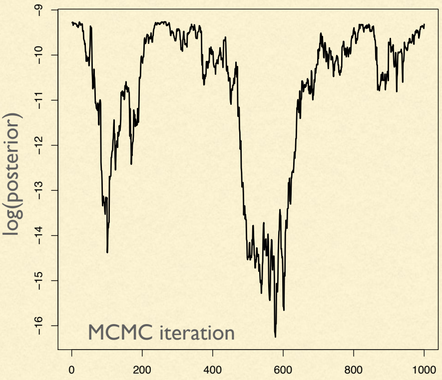
Tracer (app for generating trace plots from MCMC output):
https://github.com/beast-dev/tracer/releases/tag/v1.7.1



White noise appearance is a sign of good mixing

Target vs. Proposal Distributions

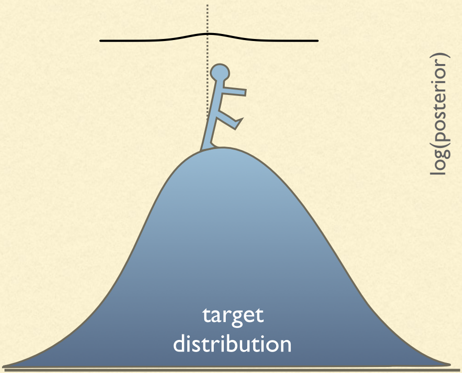


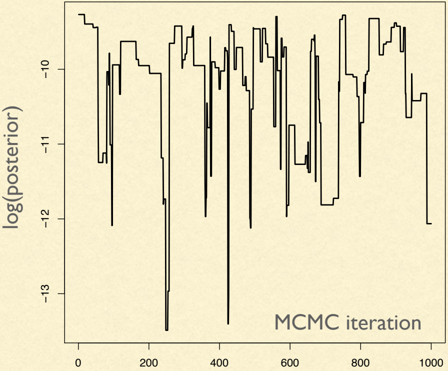


Big waves in trace plot indicate robot is crawling around

Target vs. Proposal Distributions

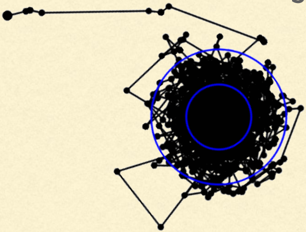
"overly bold" proposal distribution



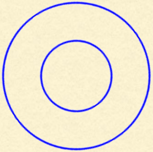


Plateaus in trace plot indicate robot is often stuck in one place

Metropolis-coupled Markov chain Monte Carlo (MCMCMC)



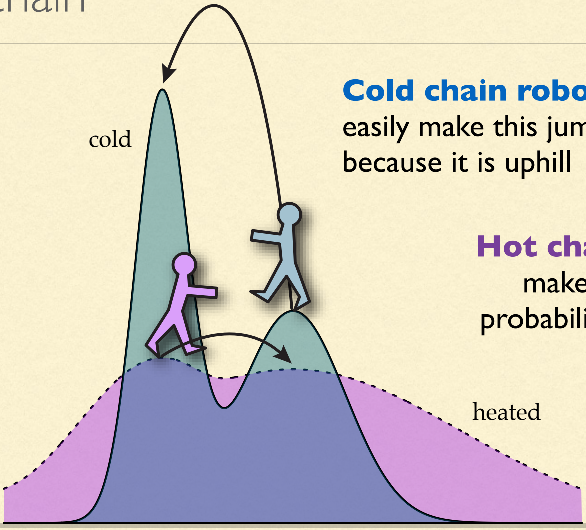
Sometimes the robot needs some help,



MCMCMC introduces helpers in the form of "heated chain" robots that can act as scouts.

Geyer, C. J. 1991. Markov chain Monte Carlo maximum likelihood for dependent data. Pages 156-163 *in* Computing Science and Statistics (E. Keramidas, ed.).

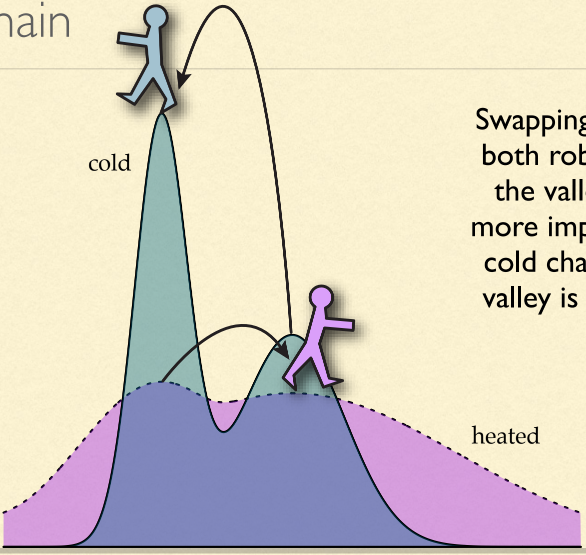
Heated chains act as scouts for the cold chain



Cold chain robot can easily make this jump because it is uphill

Hot chain robot can also make this jump with high probability because it is only slightly downhill

Heated chains act as scouts for the cold chain



Swapping places means both robots can cross the valley, but this is more important for the cold chain because its valley is much deeper.

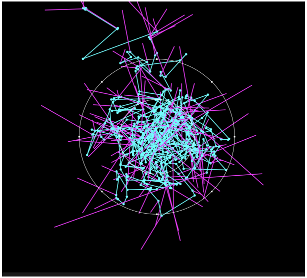
Markov Chain Monte Carlo (MCMC)

Thanks, Paul!

Slides source: https://molevol.mbl.edu/index.php/Paul_Lewis

See MCMCRobot, a helpful software program for learning MCMC by Paul Lewis

https://phylogeny.uconn.edu/mcmc-robot



RevBayes Demo: Hierarchical Archery Model

We can use MCMC to estimate m and v.

First, let's generate our observed data using simulation.

```
true_accuracy = 35.0
true_variance = 4.0
true_alpha = (true_accuracy^2) / true_variance
true_beta = true_accuracy / true_variance

n = 6
observed_shots = rgamma(n, true_alpha, true_beta)
```

The values in `observed_shots` are data generated from the true underlying distribution.

RevBayes Demo: Hierarchical Archery Model

Now we can specify the model for our new observations.

```
mean ~ dnUnif(10,50)
var ~ dnGamma(20,2)

alpha := (mean^2) / var
beta := mean / var

for(i in 1:n){
    d[i] ~ dnGamma(alpha,beta)
    d[i].clamp(observed_shots)
}
```

RevBayes Demo: Hierarchical Archery Model

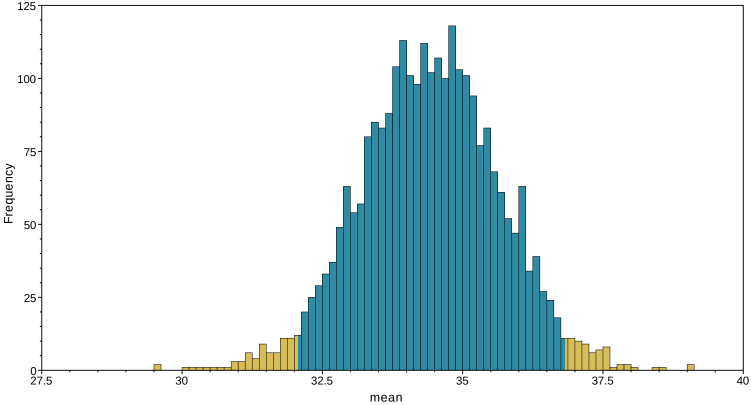
The Rev language specifying the MCMC sampler for the hierarchical model on archery accuracy.

```
mymodel = model(beta)

moves[1] = mvSlide(mean, weight=3.0)
moves[2] = mvScale(var, weight=3.0)

monitors[1] = mnModel(file="archery_mcmc_1.log", printgen=10)
monitors[2] = mnScreen(printgen=1000, mean, var)

mymcmc = mcmc(mymodel, monitors, moves)

mymcmc.burnin(generations=10000,tuningInterval=1000)

mymcmc.run(generations=40000)
```

MCMC screen output

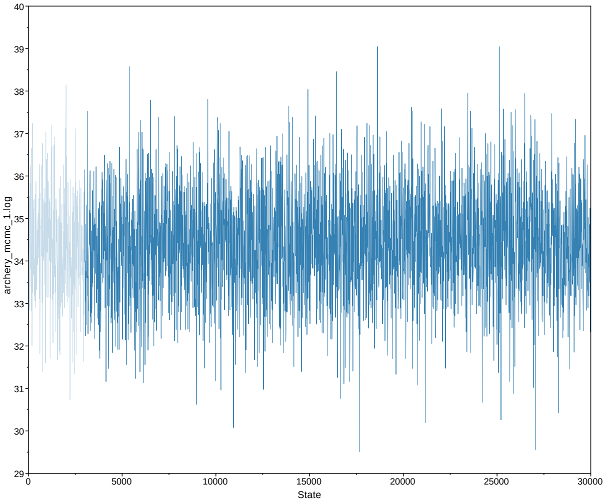
RevBayes Demo: Hierarchical Archery Model

Summary of the MCMC sample for the mean distance from target center.



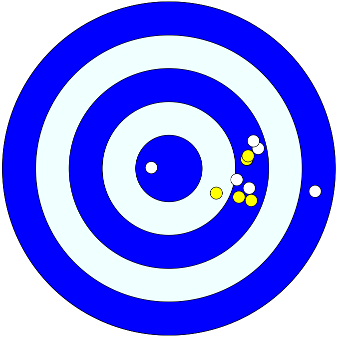
RevBayes Demo: Hierarchical Archery Model

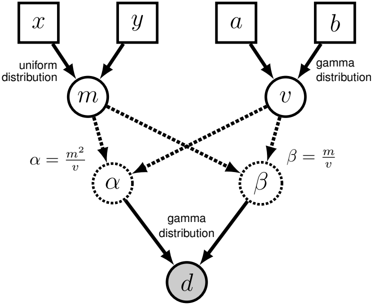
The trace-plot of the MCMC samples for the mean distance from target center



Example: Hierarchical Archery Model

Under this model, we do a good job of estimating the mean, but when judging archery skill, precision (variance) is as (if not more) important than accuracy

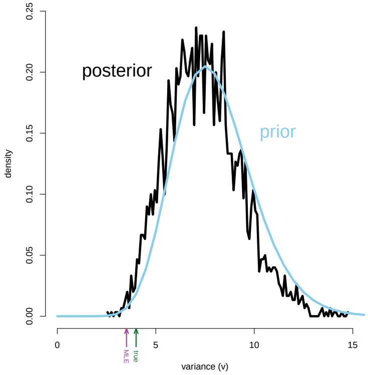




Thus, it is also worth evaluating the estimated posterior distribution for the variance component of our model

Example: Variance

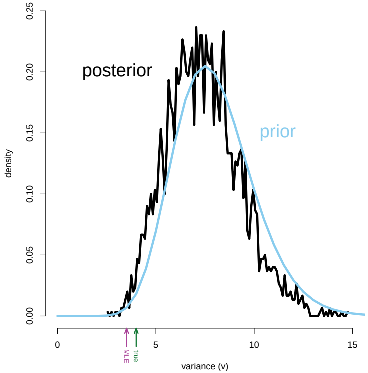
The posterior estimate of the variance (v) is quite different from the true value (4.0) and from the highest likelihood value found by our MCMC (MLE = 3.51374).



This indicates that the prior is having a strong influence on the posterior. Why do you think that is?

Example: Variance

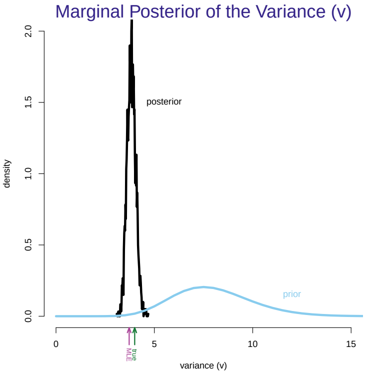
When the prior closely matches the posterior, it can indicate that the data are not very informative for this parameter.

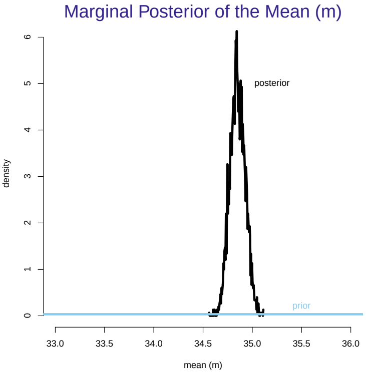


Remember that our data were only 6 observed shots. What would happen if I had 600 arrows?

Example: With a lot More Data

With 100X more observations, we can estimate the mean and variance with greater precision.

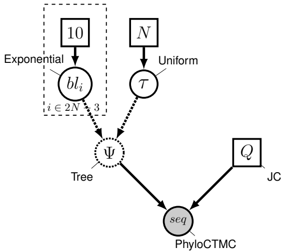




Bayesian Phylogenetics

How is this applied to phylogenetic inference?

Jukes-Cantor (1969) on an unrooted tree




```
for (i in 1:n_branches) {
        bl[i] ~ dnExponential(10.0)
}
topology ~ dnUniformTopology(taxa)
psi := treeAssembly(topology, bl)

Q <- fnJC(4)

seq ~ dnPhyloCTMC( tree=psi, Q=Q, type="DNA" )

seq.clamp( data )
```

(image source RevBayes Substitution Models Tutorial)

We can assemble a phylogenetic model in the same way, using previously described models and probability distributions as priors.

Bayesian Phylogenetics

With a defined model we simply then have to:

- draw starting values for every random variable in the model
- define moves on each random variable that propose new values
- then for each step in our MCMC, choose a parameter and update it according to the correct proposal.
  - propose a new tree topology and accept or reject
  - propose a new model parameter value and accept or reject
- save the current state of every random variable (tree, branch lengths, base frequencies, etc.) after every k number of states
- after n MCMC steps, evaluate the run for signs of non-convergence
- summarize the tree and other parameters