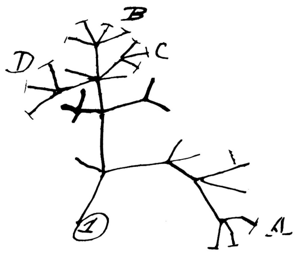


IPS-164 INTRODUCTION TO PHYLOGENETICS 2022

Lecture 10
Bayesian Inference & Model Selection

Sergei Tarasov

Beetle curator & Docent

Finnish Museum of Natural History, University of Helsinki

- @tarasov_sergio
- sergei.tarasov@helsinki.fi
- https://www.tarasovlab.com

PLAN OF THE TODAY'S LECTURE

1. Bayesian Inference
2. Model Selection
3. Dating divergence time (Part I)

Bayesian interpretation of probability

- Bayesian interpretation expresses a degree of belief in an event
- This degree of belief is based on prior knowledge about the event

Thomas Bayes (1701 –1761)

Bayes' theorem:

Likelihood: Probability of *B* given *A*

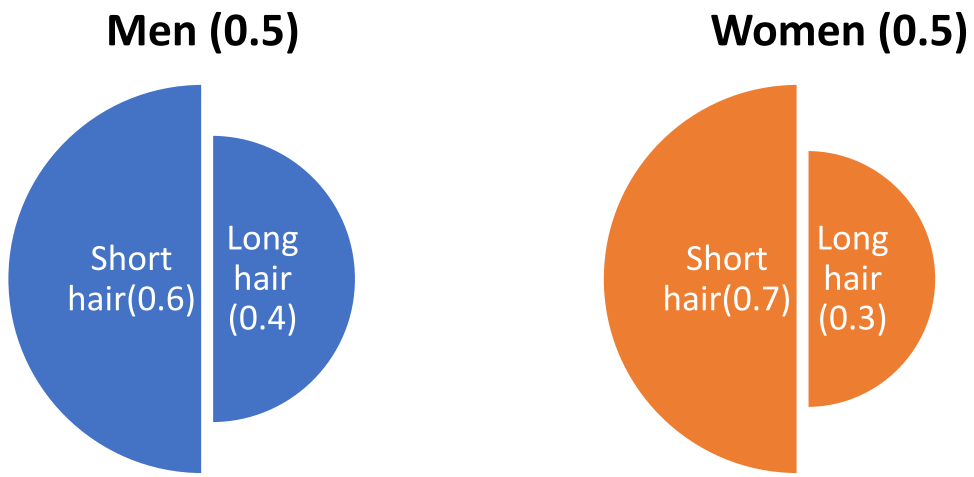
Prior: Probability of *A* before gathering the data

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Posterior: Probability that *A* is true given *B* is observed

Probability of *B* (=**probability of data**, =**marginal probability**)

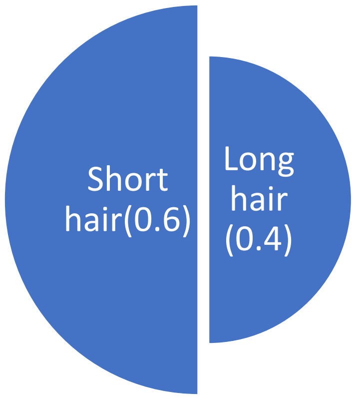
Using the Bayes theorem: $P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$



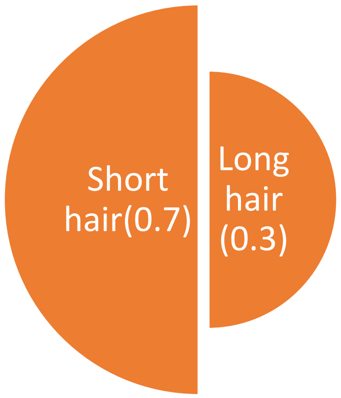
If we see someone has long hair, what is the probability that this person is a man (or a woman)?

Using the Bayes theorem: $P(A|B) = \dfrac{P(B|A)\,P(A)}{P(B)}$

Men (0.5)



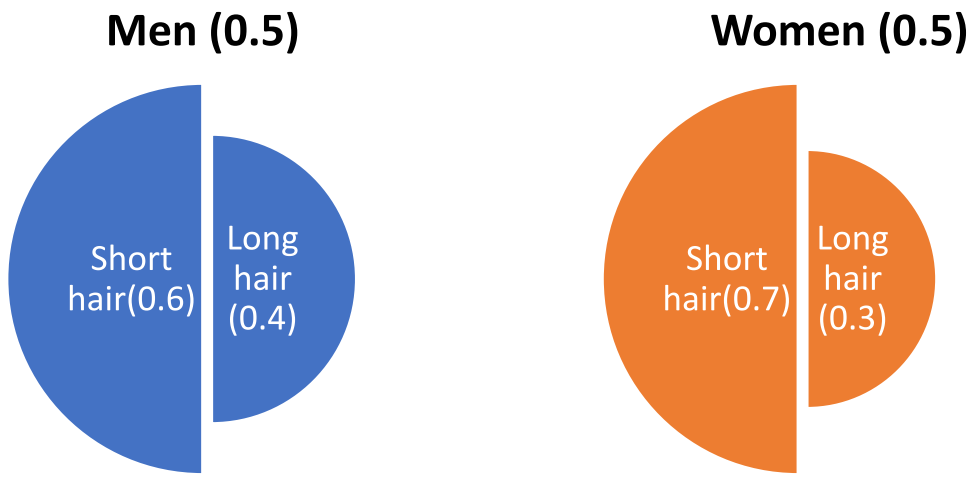
Women (0.5)



If we see someone has long hair, what is the probability that this person is a man (or a woman)?

M & W	Total Prob.	Hair Length Prob.
Short	0.6*0.5 = 0.3	
Long	0.4*0.5 = 0.2	
Short	0.7*0.5 = 0.35	
Long	0.3*0.5 = 0.15	
Total	1	

Using the Bayes theorem: $P(A|B) = \dfrac{P(B|A)\,P(A)}{P(B)}$

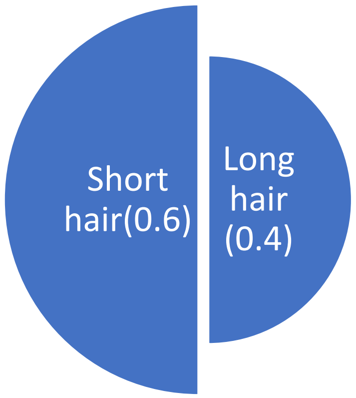


If we see someone has long hair, what is the probability that this person is a man (or a woman)?

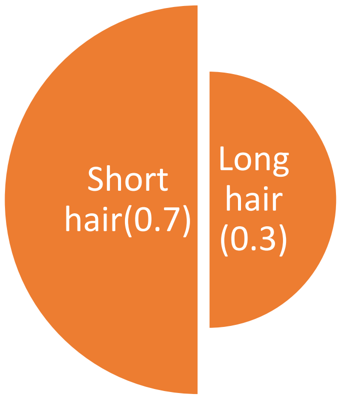
M & W	Total Prob.	Hair Length Prob.
Short	0.6*0.5 = 0.3	
Long	0.4*0.5 = 0.2	0.2/(0.2+0.15) = **0.57**
Short	0.7*0.5 = 0.35	
Long	0.3*0.5 = 0.15	0.15/(0.2+0.15) = **0.43**
Total	1	

Using the Bayes theorem: $\boldsymbol{P(A|B) = \dfrac{P(B|A)\,P(A)}{P(B)}}$

Men (0.5)



Women (0.5)

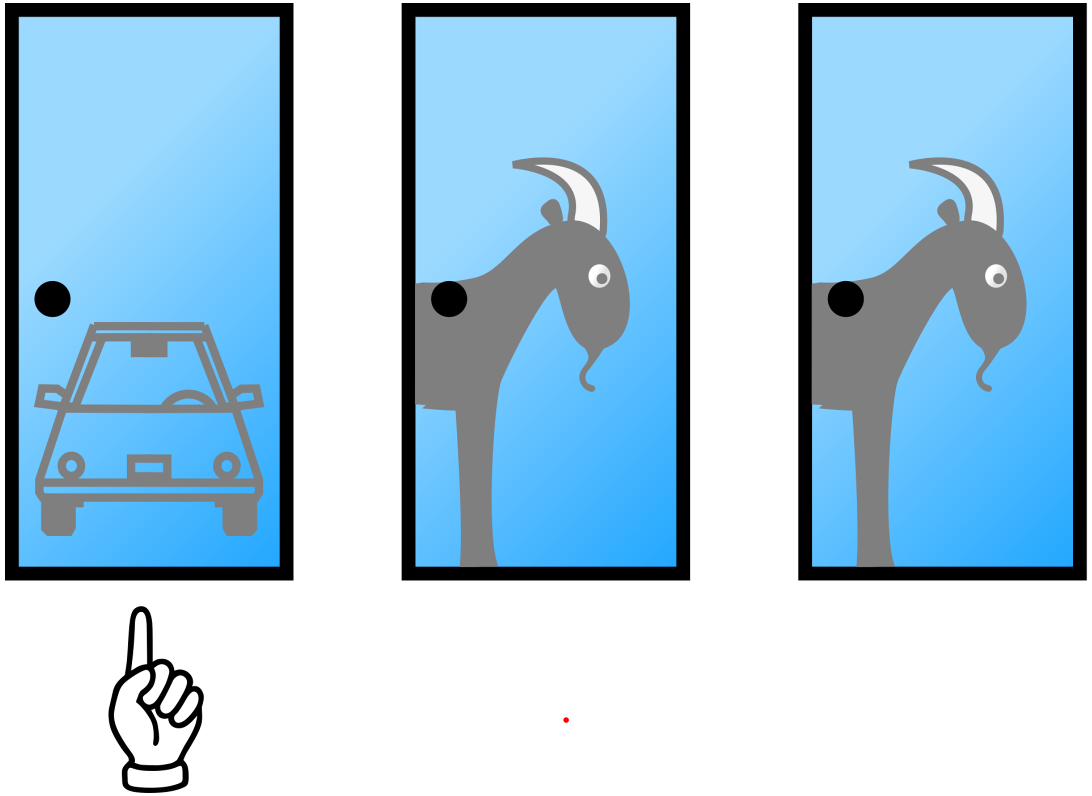


$$P(man \mid long\ hair) = \frac{\overset{\mathbf{0.4}}{P(long\ hair|man)}\ \overset{\mathbf{0.5}}{P(man)}}{\underset{\mathbf{0.35}}{P(long\ hair)}} = \mathbf{0.57}$$

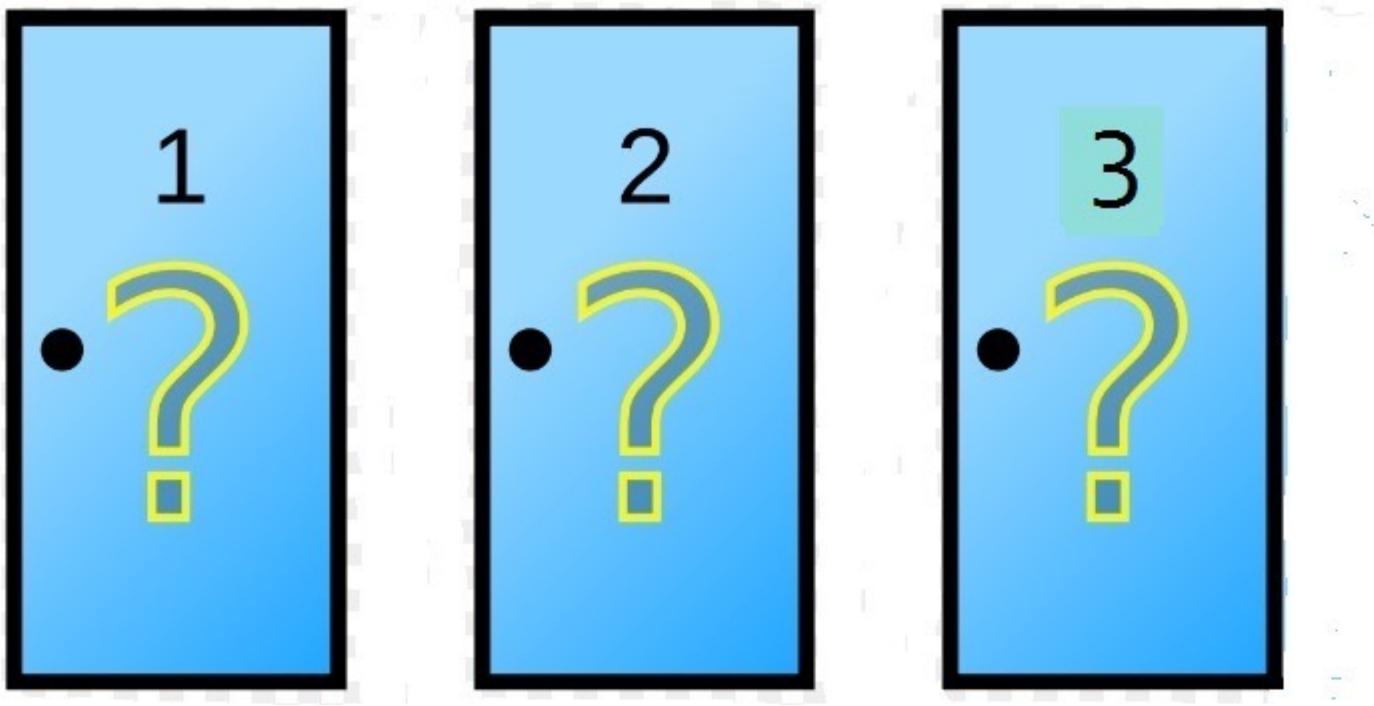
$$P(long\ hair) = \overset{\mathbf{0.4}}{P(Long\ hair)}\overset{\mathbf{0.5}}{P(man)} + \overset{\mathbf{0.3}}{P(Long\ hair)}\overset{\mathbf{0.5}}{P(woman)} = \boldsymbol{0.35}$$

If we see someone has long hair, what is the probability that this person is a man (or a woman)?

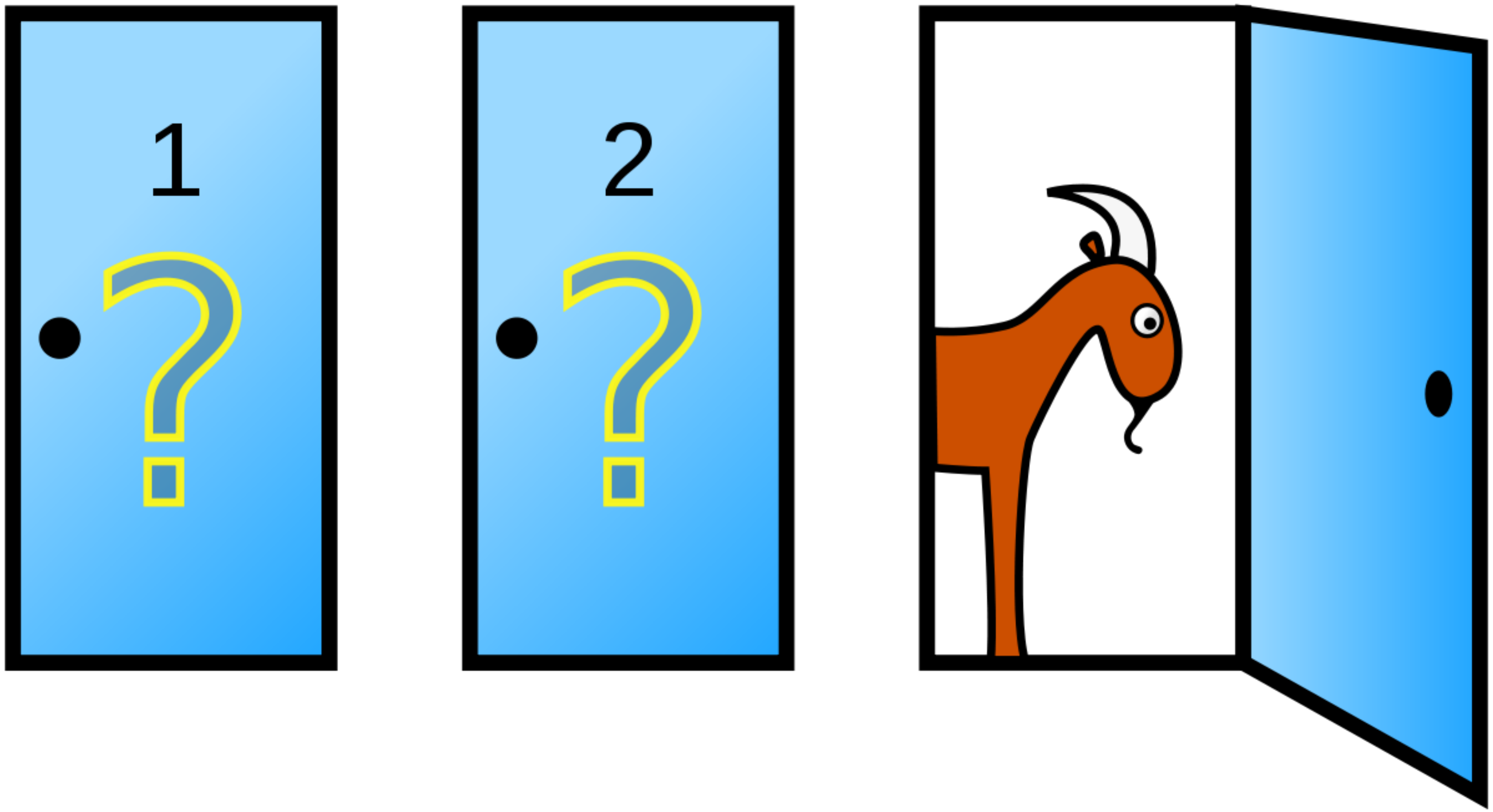
Monty Hall Paradox



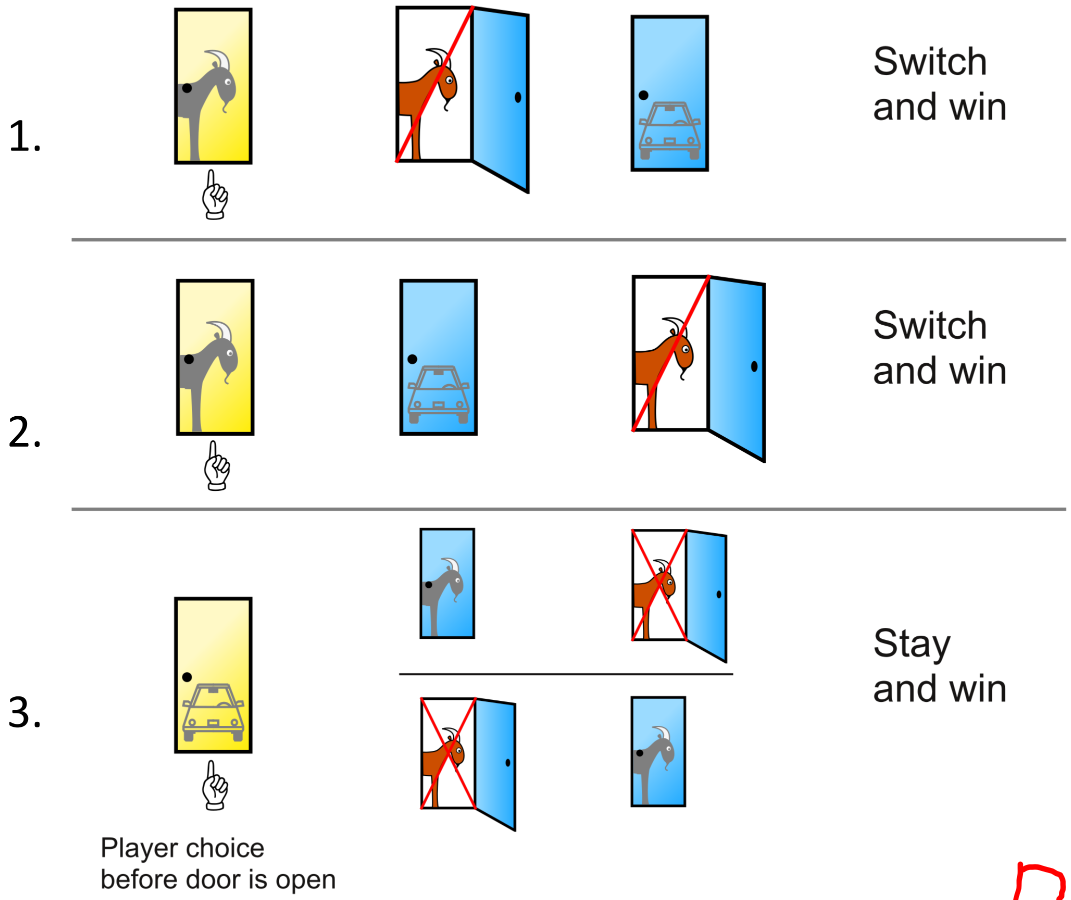
Monty Hall Paradox



Monty Hall Paradox



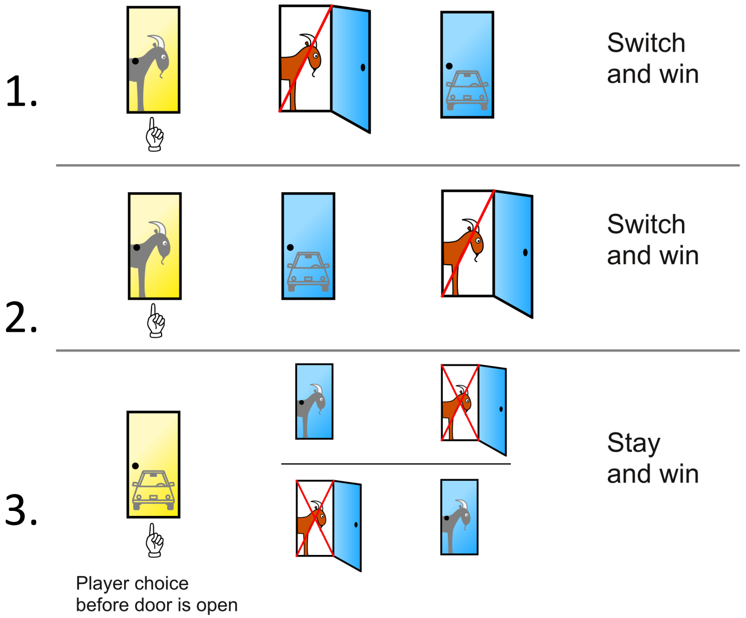
Monty Hall Paradox



Stay	Switch
goat	car
g	CAR
c	g

$$Pr(win) = \frac{1}{3}; \quad \frac{2}{3}$$

Monty Hall Paradox



(3) $\mathbf{p}(\textit{the car behind door 1} \mid \text{Monty Hall opens door 3})=$

$$\frac{\mathbf{p}(\text{Monty Hall opens door 3} \mid \text{the car behind door 1}) * \mathbf{p}(\text{the car behind door 1})}{\mathbf{p}(\text{Monty Hall opens door 3})}$$

$= (1/2 * 1/3) / (1/2) = \mathbf{1/3}$

(1&2) $\mathbf{p}(\textit{the car behind door 2} \mid \text{Monty Hall opens door 3})=$

$$\frac{\mathbf{p}(\text{Monty Hall opens door 3} \mid \text{the car behind door 2}) * \mathbf{p}(\text{the car behind door 2})}{\mathbf{p}(\text{Monty Hall opens door 3})}$$

$= (1 * 1/3) / (1/2) = \mathbf{2/3}$

Bayesian interpretation of probability

- Bayesian interpretation expresses a degree of belief in an event
- This degree of belief is based on prior knowledge about the event

Bayes' theorem:

Likelihood: Probability of *B* given *A*

Prior: Probability of *A* before gathering the data

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Posterior: Probability that *A* is true given *B* is observed

Probability of *B* (=**probability of data**, =**marginal probability**)

Without loss of generality posterior can be written as:

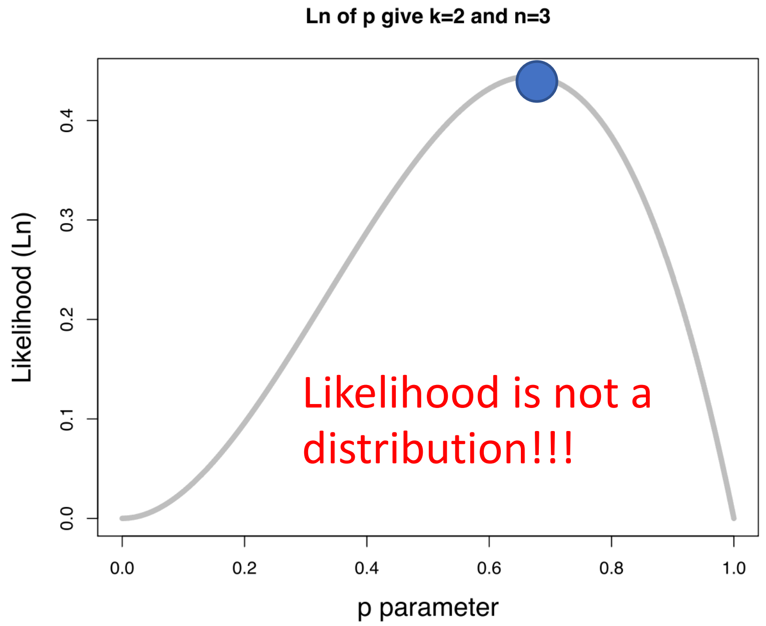
*Posterior ∝ Likelihood * Prior*

Likelihood function of Binomial distribution

Given *n* and *k*, infer probability for every *p*

Ln(*p*|*n*,*k*)

$$Likelihood(p \mid n=3, k=2) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{3}{2} p^2 (1-p)^{3-2}$$

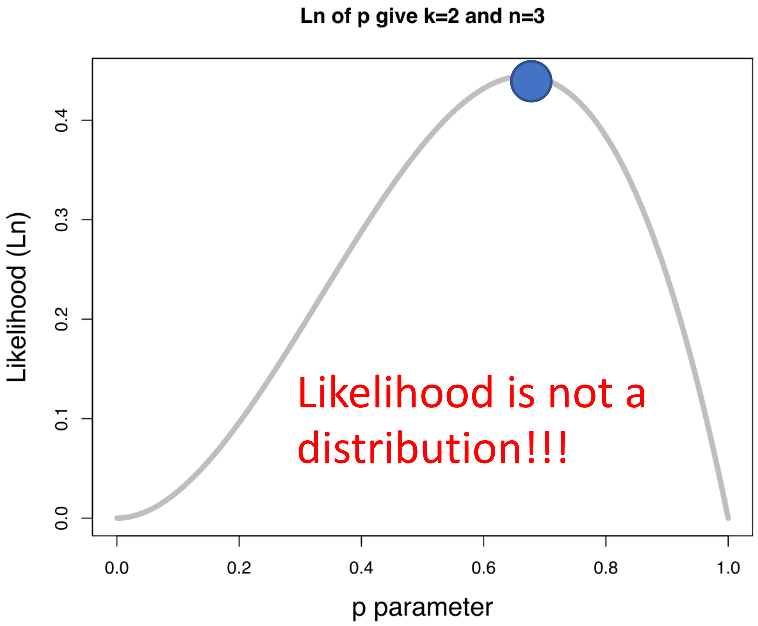


Likelihood function of Binomial distribution

Given *n* and *k*, infer probability for every *p*

Ln(*p*|*n*,*k*)

$$Likelihood(p \mid n = 3, k = 2) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{3}{2} p^2 (1-p)^{3-2}$$



Informal Axiom of Statistics:

Any measured quantity of any set of objects in the Universe has some probability distribution

What if the parameter *p* is not just a maximum point but has some distribution?

Use the Bayes theorem!

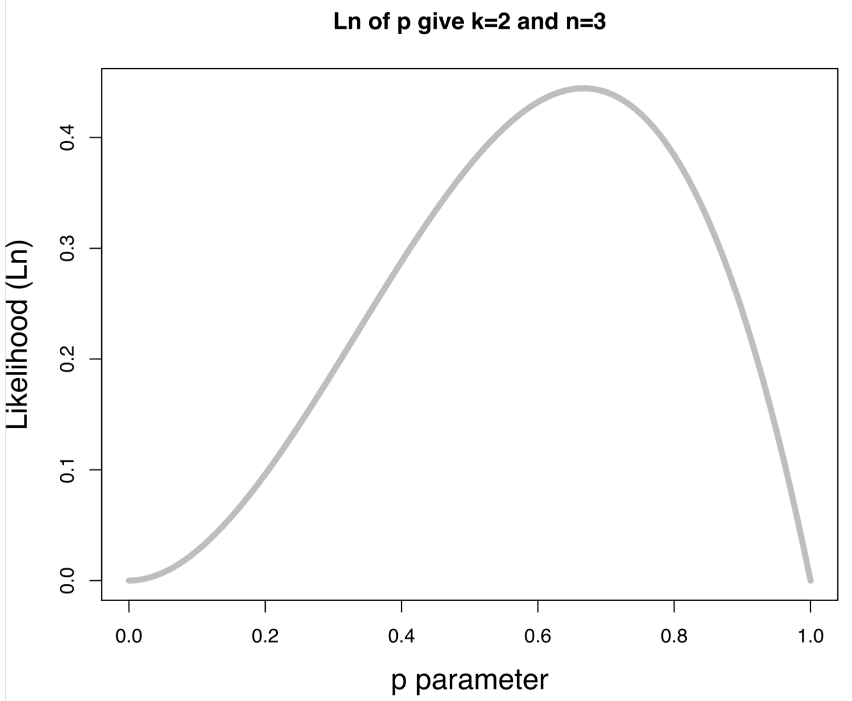
*Posterior ∝ Likelihood * Prior*

Bayesian theorem applied to probability distribution

- We can find the distribution of p using Bayes theorem:

  Posterior ∝ Likelihood * Prior

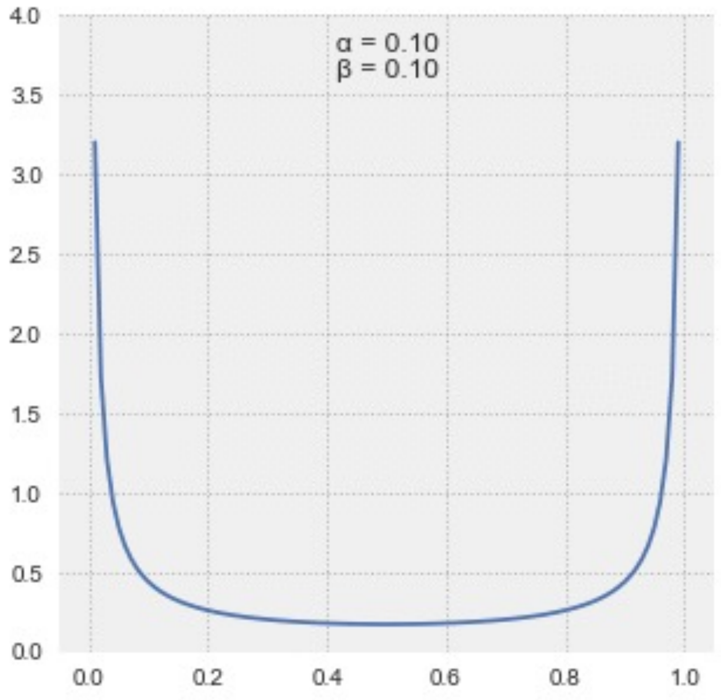
- Bayes theorem requires prior choice



Likelihood

$\binom{n}{k}p^k(1-p)^{n-k} = \binom{3}{2}p^2(1-p)^{3-2}$

Beta distribution is the natural choice as the prior for the Binomial Likelihood

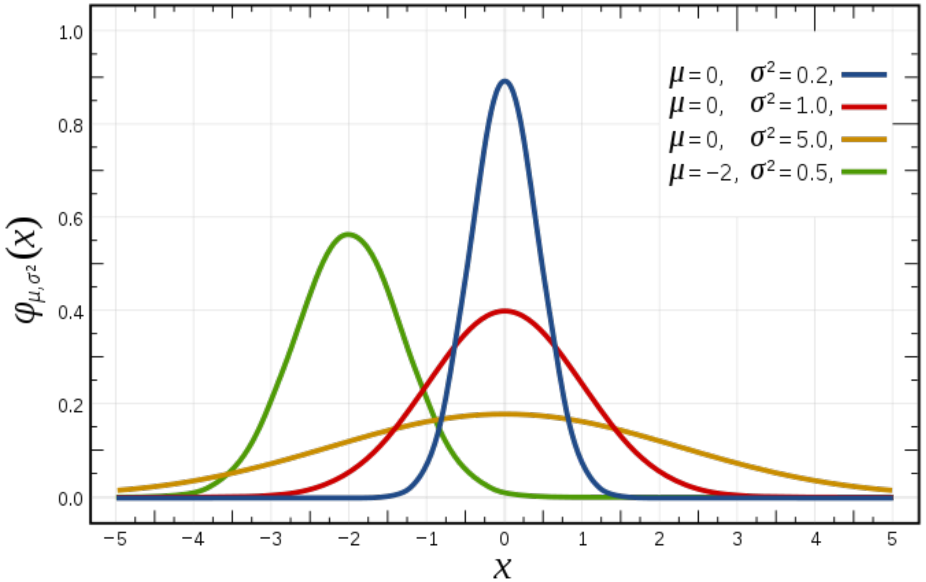


$$f(x|a,b) = \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

https://en.wikipedia.org/wiki/Beta_distribution

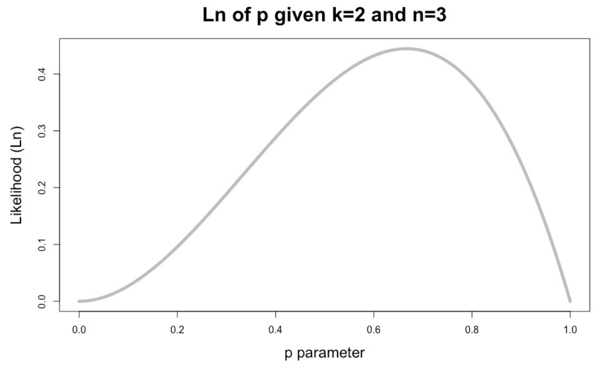
Why not Normal?

Normal Distribution. It has two parameters:
Mean (mu) and Variance (sigma)

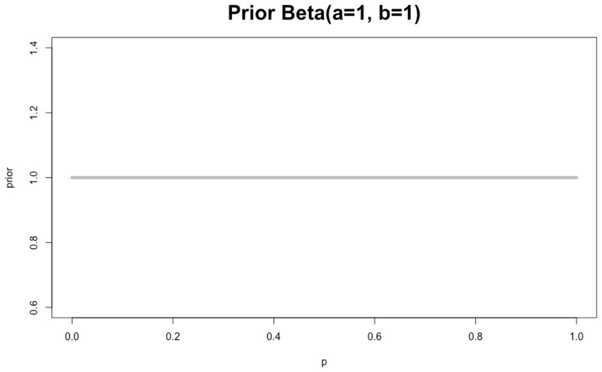


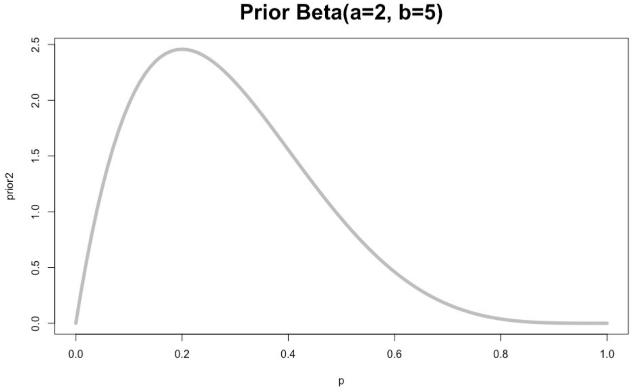
$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

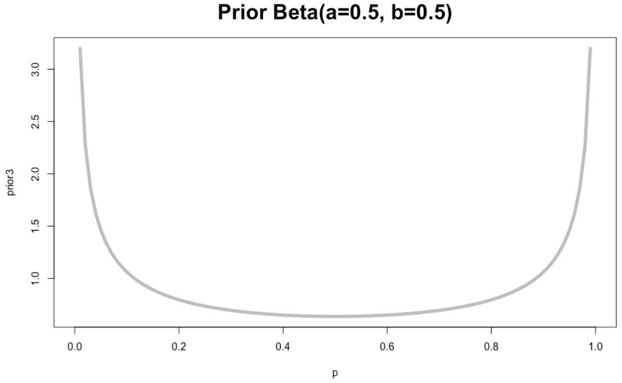
Posterior for coin toss



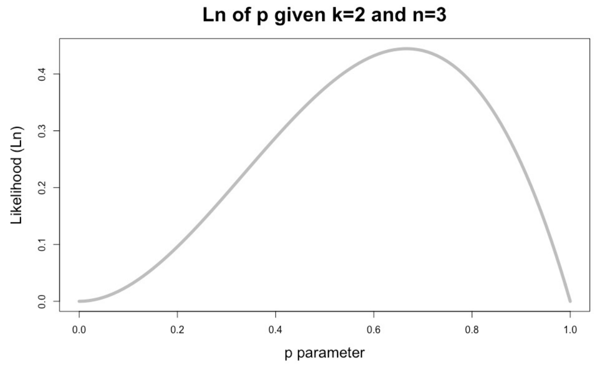
Priors:



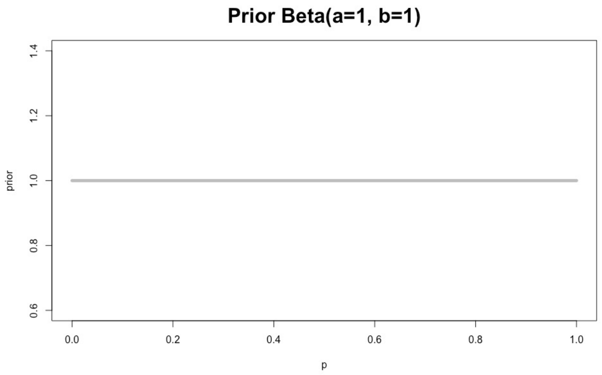


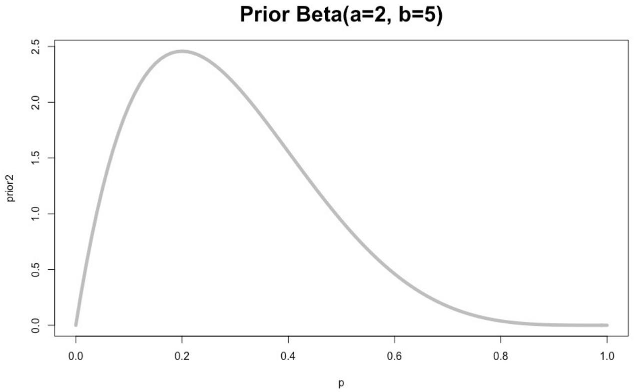


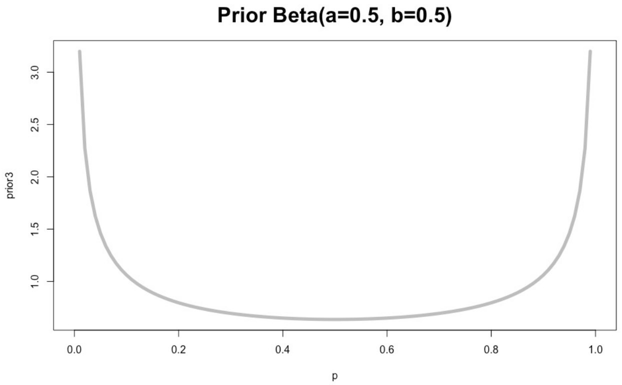
Posterior for coin toss



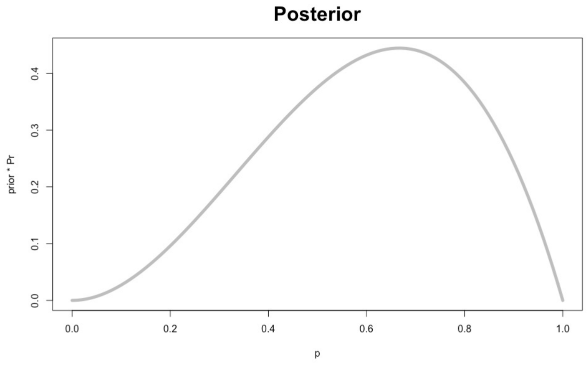
Priors:

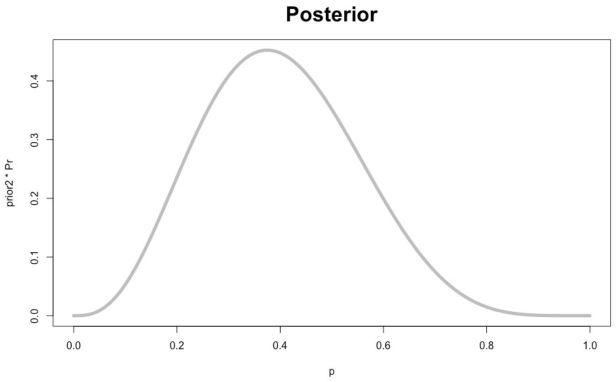


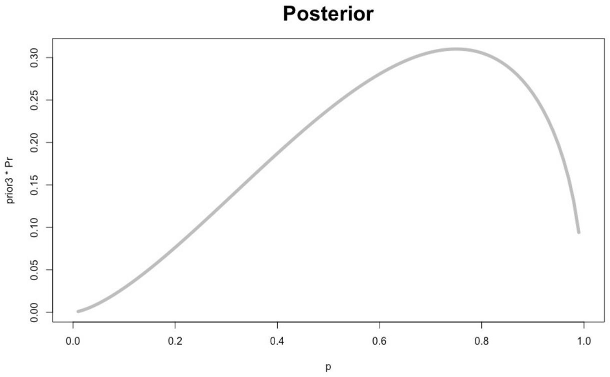




Posteriors:

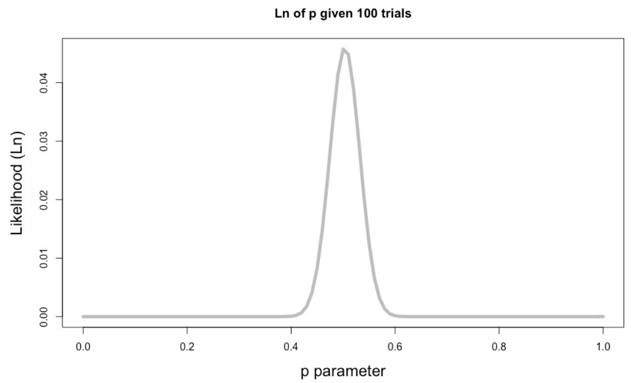




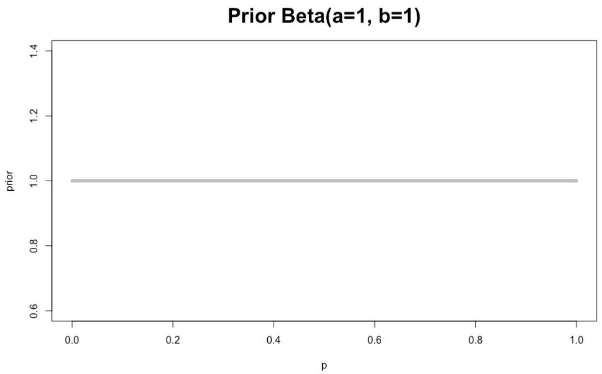


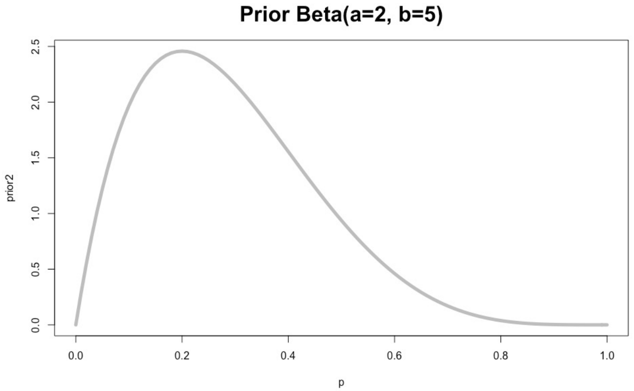
Posterior for coin toss

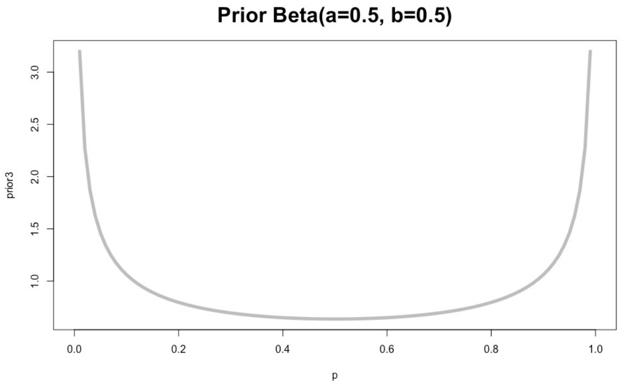
100 trials



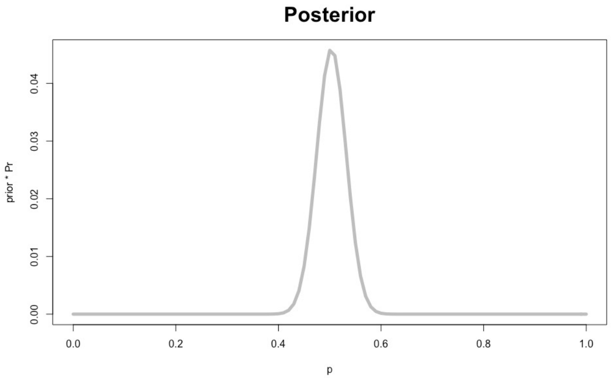
Priors:

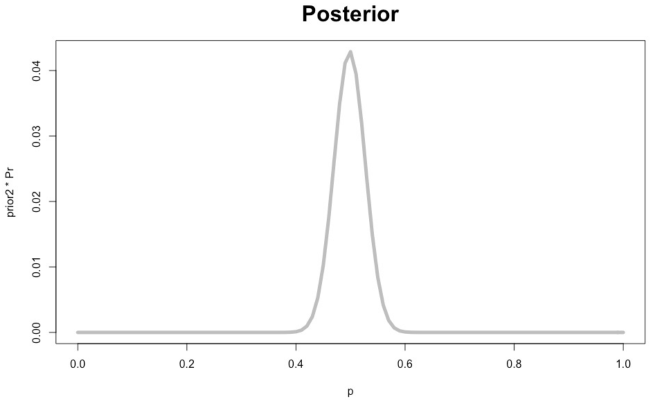


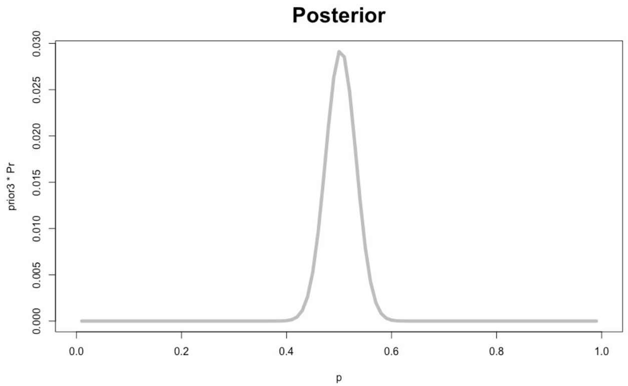




Posteriors:







Hyperpriors are the priors for the priors

Informal Axiom of Statistics:

Any measured quantity of any set of objects in the Universe has some probability distribution

Model with priors

Likelihood: $(p \mid n, k) = \binom{n}{k} p^k (1-p)^{n-k}$

Prior: $p \sim Beta(a, b)$

Model with hyperpriors

Likelihood: $(p \mid n, k) = \binom{n}{k} p^k (1-p)^{n-k}$

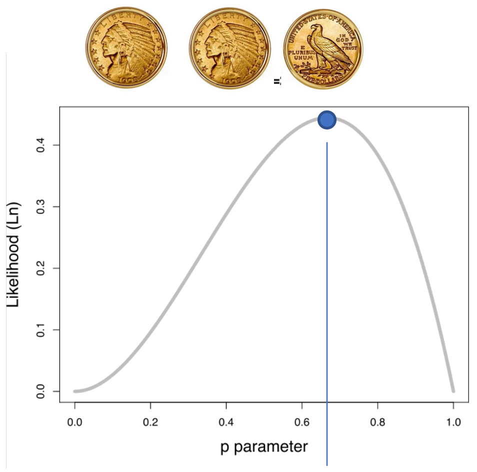
Prior: $p \sim Beta(a, b)$

Hyperprior: $a \sim Gamma(k_1, \theta_1)$

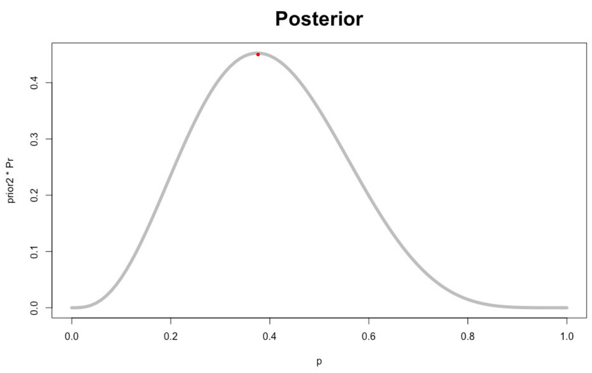
Hyperprior: $b \sim Gamma(k_2, \theta_2)$

Bayesian inference

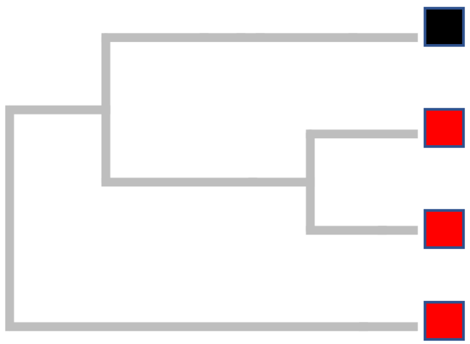
- Sample parameters from their joint posterior distribution
- Your parameter sample is a distribution
- It's not a point estimate as in the Likelihood method



ML



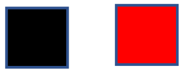
BI



Topology and branch lengths

$$\boldsymbol{Q}=\begin{bmatrix} -\alpha & \alpha \\ \beta & -\beta \end{bmatrix}$$

Rates of the rate matrix



$$\pi = (\pi_1, \pi_2)$$

Initial vector at the root of tree

Felsenstein's pruning algorithm is the same for the Bayesian Inference but add Priors

Given values:

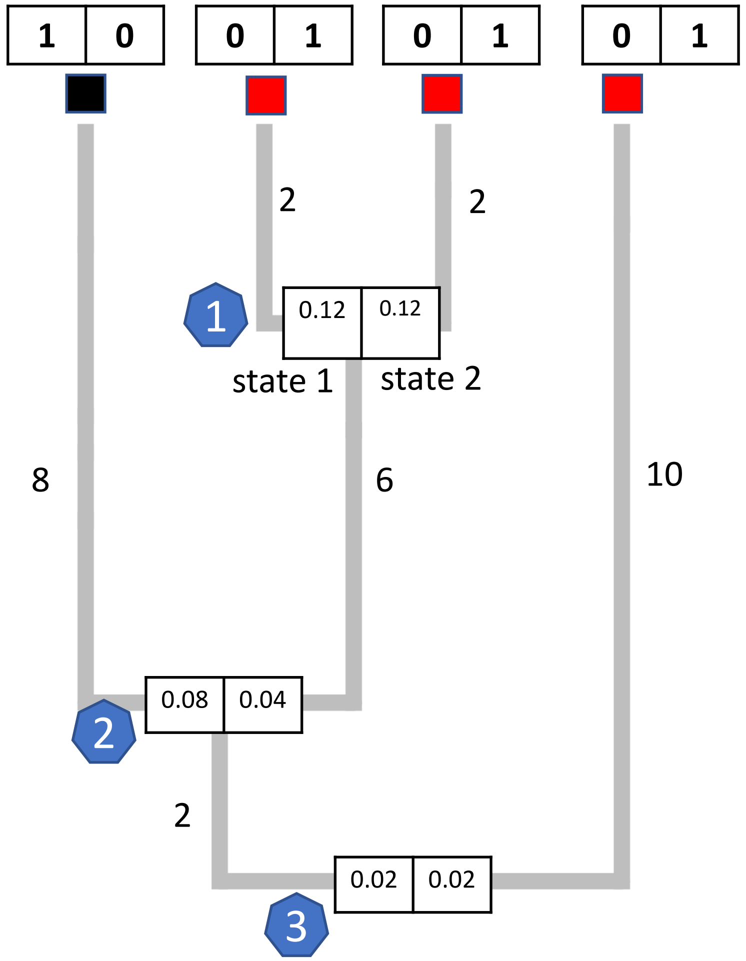
state 1 state 2

1	0

0	1

$$\boldsymbol{Q}=\begin{bmatrix}-1 & 1 \\ 2 & -2\end{bmatrix}$$

$\pi = (1/2, 1/2)$



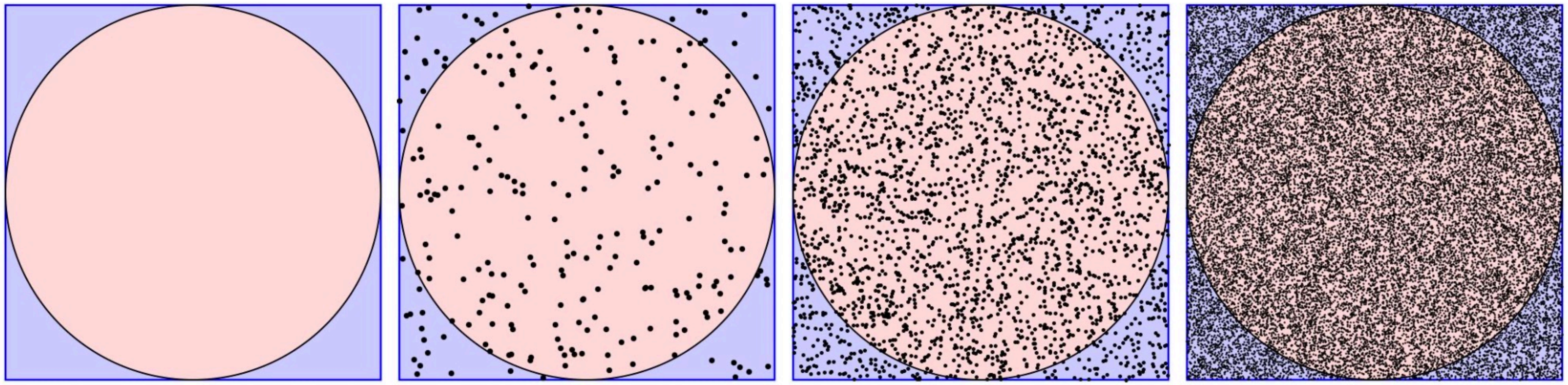
Likelihood (at the root):

$L(tree)$ = Pr(**black**)* $\boldsymbol{\pi_1}$+Pr(**red**)* π_2 = 0.02*1/2+ 0.02*1/2 = **0.02**

Posterior ∝ L(tree) * Prior

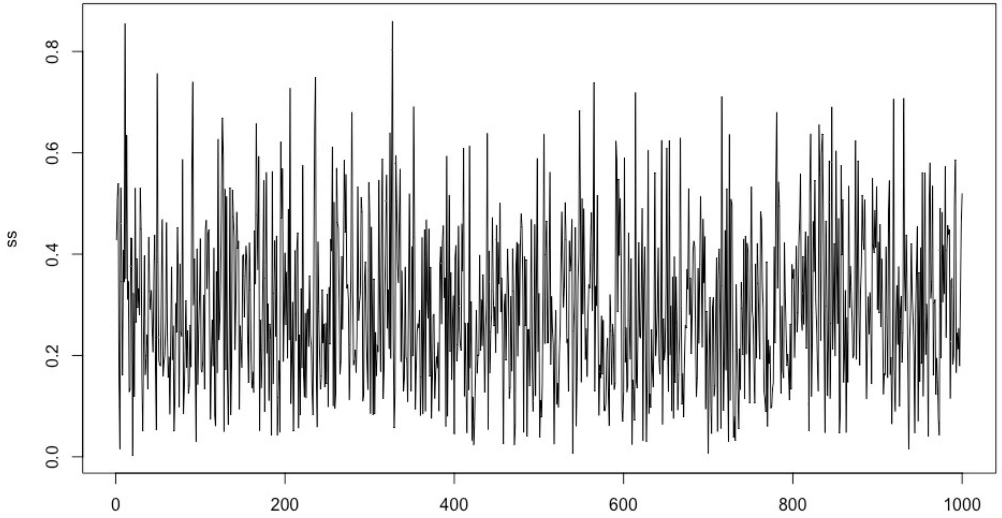
Approximating the posterior distribution with Markov Chain Monte Carlo (MCMC) method using Metropolis-Hasting algorithm

Estimating area of the circle using Monte Carlo method

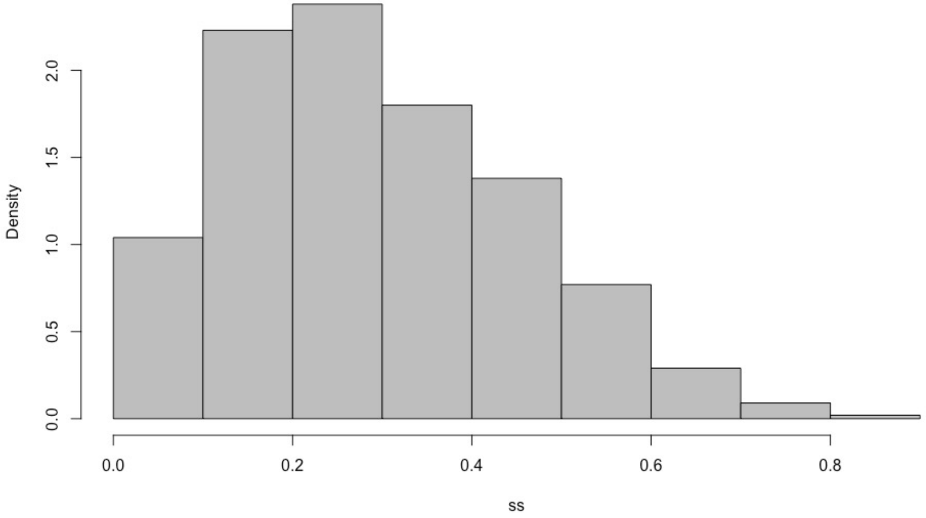


Approximating the posterior distribution with Markov Chain Monte Carlo (MCMC) method using Metropolis-Hasting algorithm

MCMC is a Markov chain that being at stationary randomly samples from the posterior distribution



Posterior

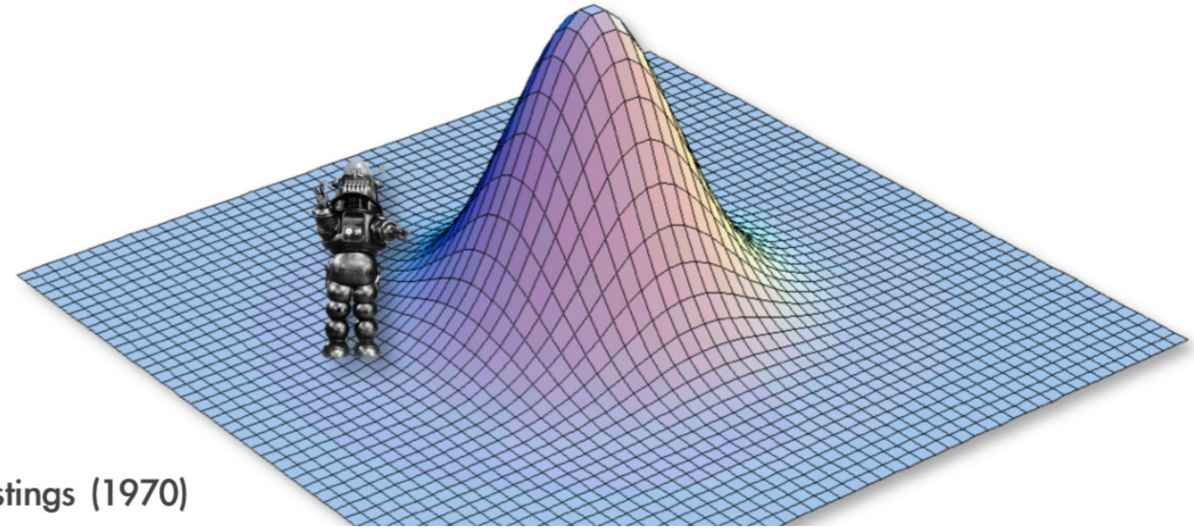


Approximating the Joint Posterior Probability Density with MCMC

Programming our MCMC robot...

Our robot parachutes into a random location in the joint posterior density and will explore parameter space by following these simple rules:

From the presentation of Brian Moore
(Univ. of Davis)



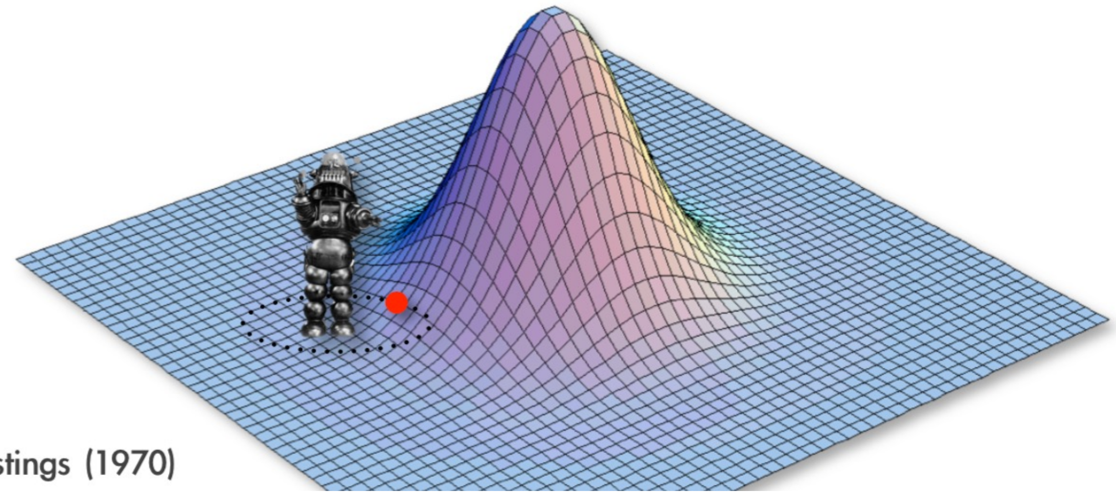
Metropolis et al. (1953); Hastings (1970)

Approximating the Joint Posterior Probability Density with MCMC

Programming our MCMC robot...

Our robot parachutes into a random location in the joint posterior density and will explore parameter space by following these simple rules:

From the presentation of Brian Moore (Univ. of Davis)



Metropolis et al. (1953); Hastings (1970)

Approximating the Joint Posterior Probability Density with MCMC

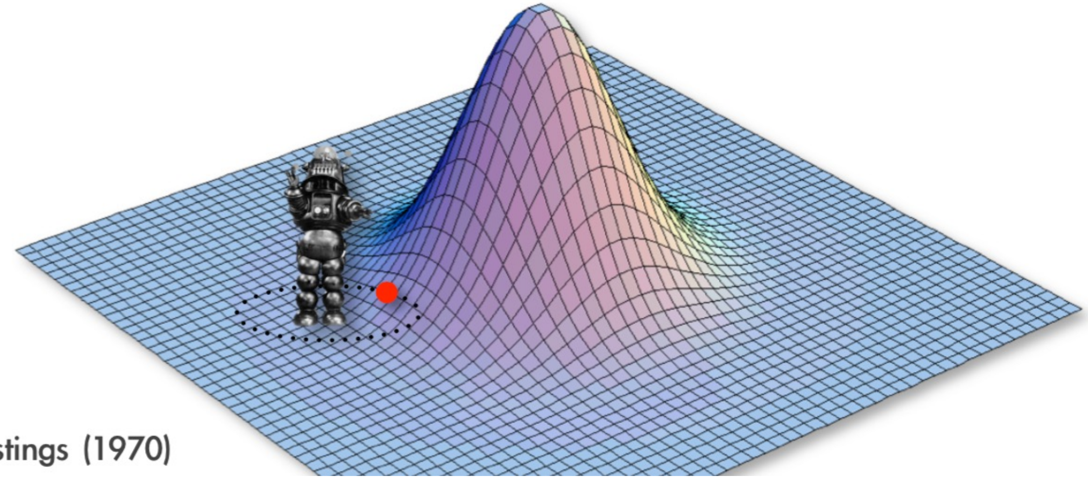
Programming our MCMC robot...

Our robot parachutes into a random location in the joint posterior density and will explore parameter space by following these simple rules:

1. If the proposed step will take the robot uphill, it automatically takes the step

From the presentation of Brian Moore (Univ. of Davis)

Pr[Accept] = 1



Metropolis et al. (1953); Hastings (1970)

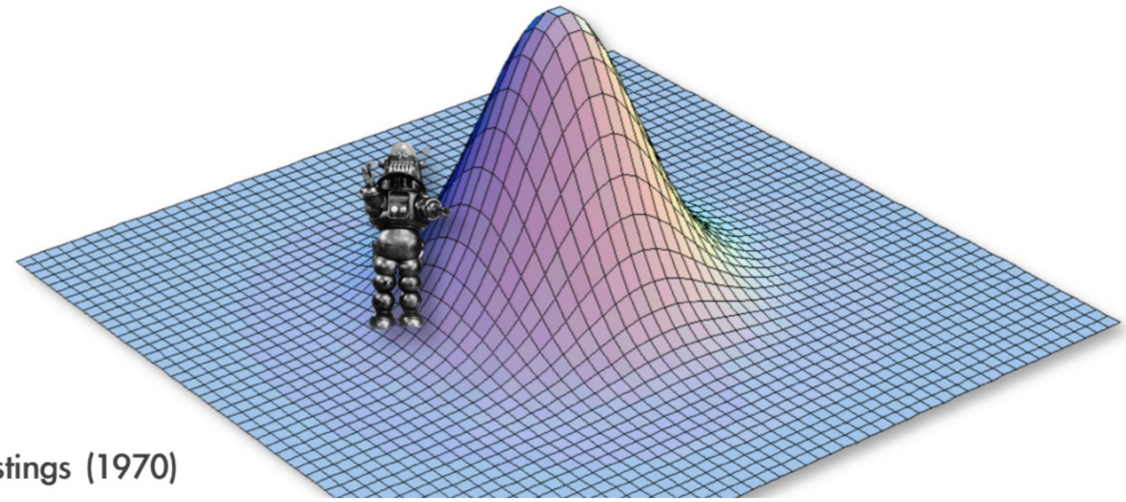
Approximating the Joint Posterior Probability Density with MCMC

Programming our MCMC robot...

Our robot parachutes into a random location in the joint posterior density and will explore parameter space by following these simple rules:

1. If the proposed step will take the robot uphill, it automatically takes the step

From the presentation of Brian Moore
(Univ. of Davis)



Metropolis et al. (1953); Hastings (1970)

Approximating the Joint Posterior Probability Density with MCMC

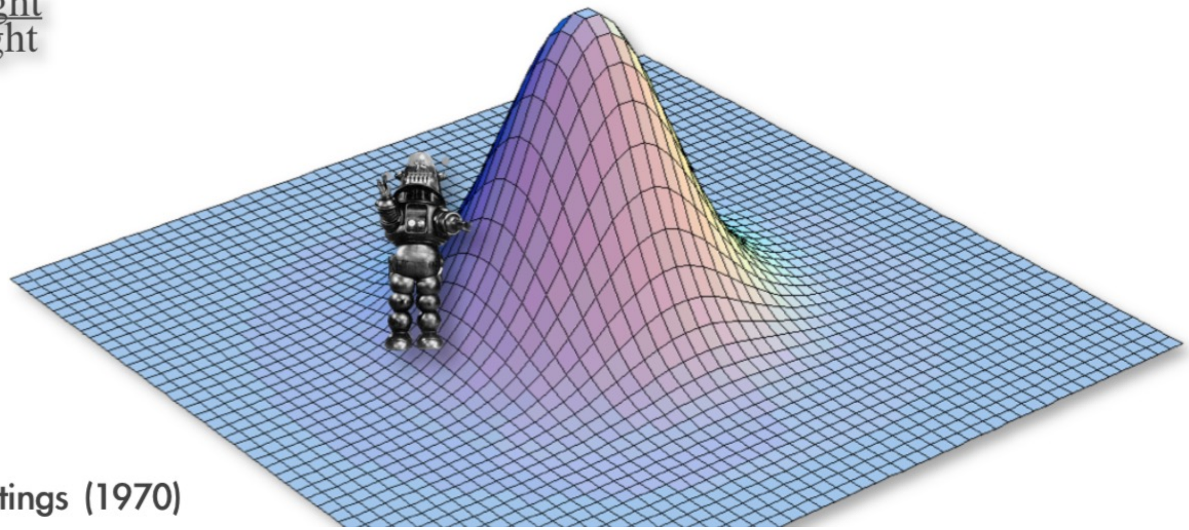
Programming our MCMC robot...

Our robot parachutes into a random location in the joint posterior density and will explore parameter space by following these simple rules:

1. If the proposed step will take the robot uphill, it automatically takes the step
2. If the proposed step will take the robot downhill, it divides the elevation of the proposed location by the current location, and it only takes the step if the quotient is less than a uniform random variable, U[0,1]

From the presentation of Brian Moore (Univ. of Davis)

$$\Pr[\text{Accept}] = \frac{\text{new height}}{\text{old height}}$$



Metropolis et al. (1953); Hastings (1970)

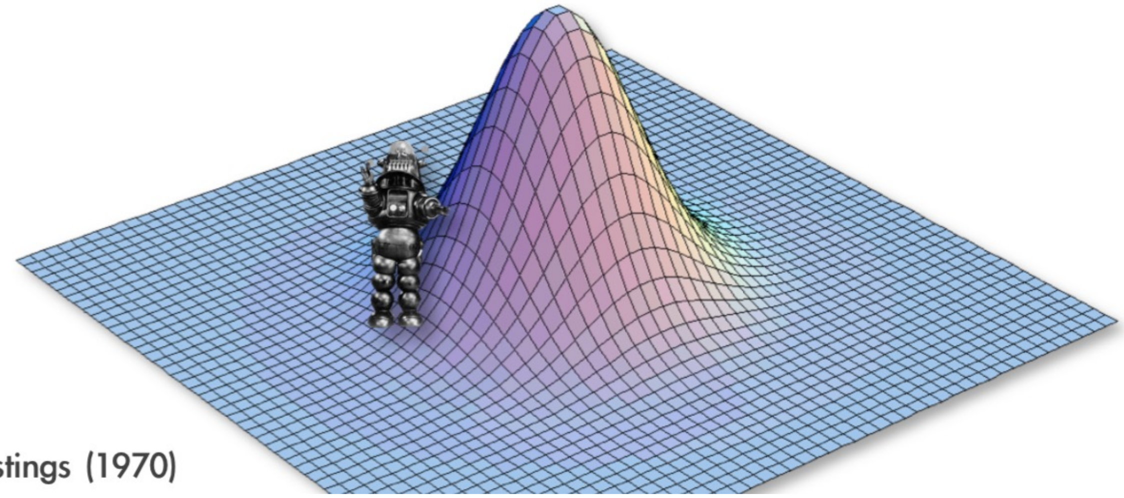
Approximating the Joint Posterior Probability Density with MCMC

Programming our MCMC robot...

Our robot parachutes into a random location in the joint posterior density and will explore parameter space by following these simple rules:

1. If the proposed step will take the robot uphill, it automatically takes the step
2. If the proposed step will take the robot downhill, it divides the elevation of the proposed location by the current location, and it only takes the step if the quotient is less than a uniform random variable, $U[0,1]$
3. The proposal distribution is symmetrical, so $\Pr[A \rightarrow B] = \Pr[B \rightarrow A]$

From the presentation of Brian Moore (Univ. of Davis)



Metropolis et al. (1953); Hastings (1970)

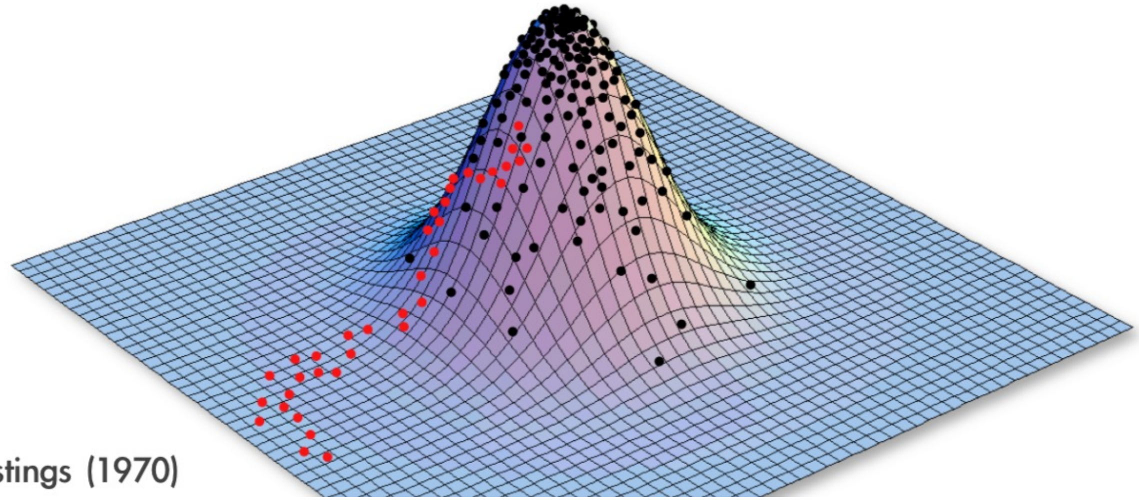
Approximating the Joint Posterior Probability Density with MCMC

Programming our MCMC robot...

Our robot parachutes into a random location in the joint posterior density and will explore parameter space by following these simple rules:

1. If the proposed step will take the robot uphill, it automatically takes the step
2. If the proposed step will take the robot downhill, it divides the elevation of the proposed location by the current location, and it only takes the step if the quotient is less than a uniform random variable, $U[0,1]$
3. The proposal distribution is symmetrical, so $\Pr[A \rightarrow B] = \Pr[B \rightarrow A]$

From the presentation of Brian Moore (Univ. of Davis)



Metropolis et al. (1953); Hastings (1970)

Assessing MCMC Performance: Three Main Issues

1. Convergence

   Has the chain (robot) successfully targeted the stationary distribution?

2. Mixing

   Is the chain (robot) successfully integrating over the joint posterior probability?

3. Sampling intensity

   Has the robot collected enough samples to adequately describe the posterior probability distribution?

From the presentation of Brian Moore (Univ. of Davis)

Software for accessing diagnostics:

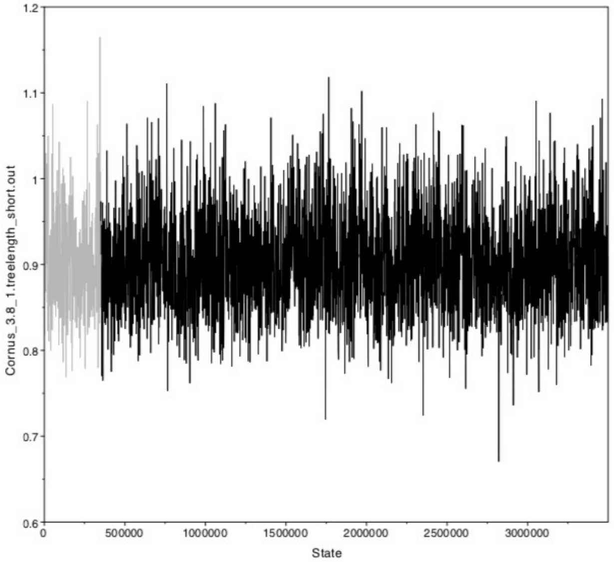
- Tracer https://github.com/beast-dev/tracer/releases/tag/v1.7.1
- Bonsai (R package)
- AWTY

Assessing MCMC Performance: Diagnostics Based on Single Runs

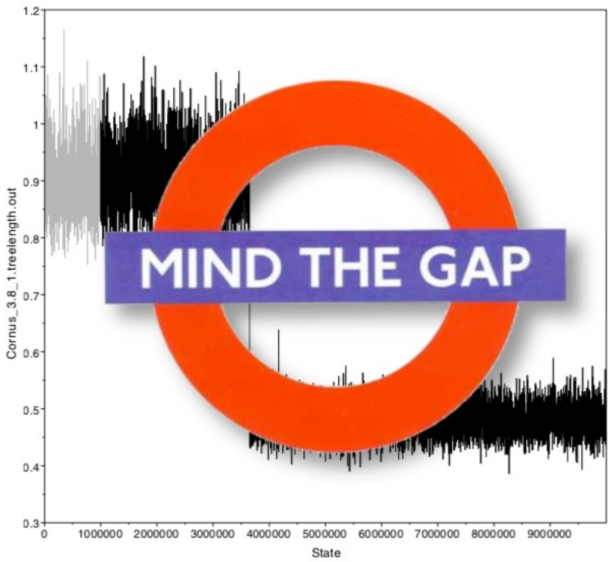
Example: Tracer plots of tree-length at two stages of a single MrBayes run

From the presentation of Brian Moore (Univ. of Davis)

bad convergence



better convergence



fast* slow*

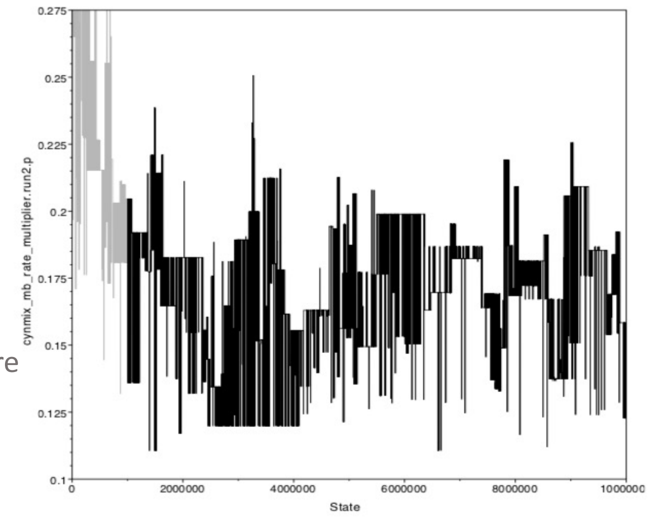
lnL base freq. sub. rates ASRV TL topology

*somewhat data-set dependent

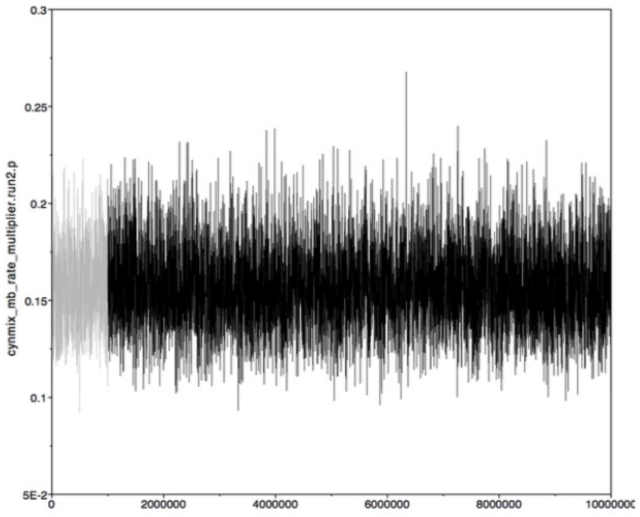
Assessing MCMC Performance: Diagnostics Based on Single Runs

Example: Tracer plots of relative-rate multipliers from two MrBayes runs

bad mixing



better mixing



From the presentation of Brian Moore (Univ. of Davis)

Assessing MCMC Performance: Diagnostics Based on Single Runs

Example: Tracer plots of relative-rate multipliers from two MrBayes runs

bad mixing

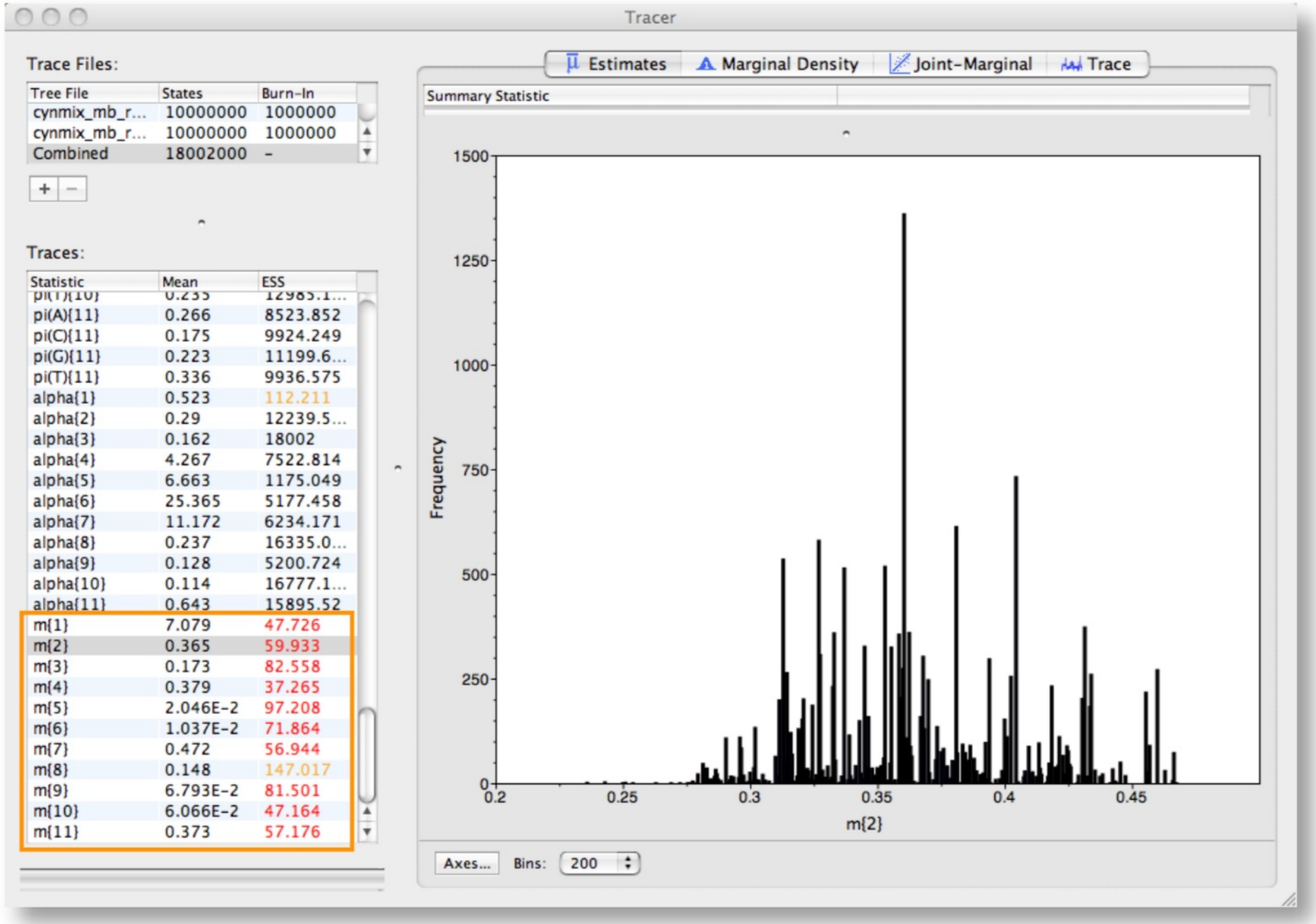
better mixing

From the presentation of Brian Moore (Univ. of Davis)

Assessing MCMC Performance: Diagnostics Based on Single Runs

Example: ESS values for relative-rate multipliers from two MrBayes runs

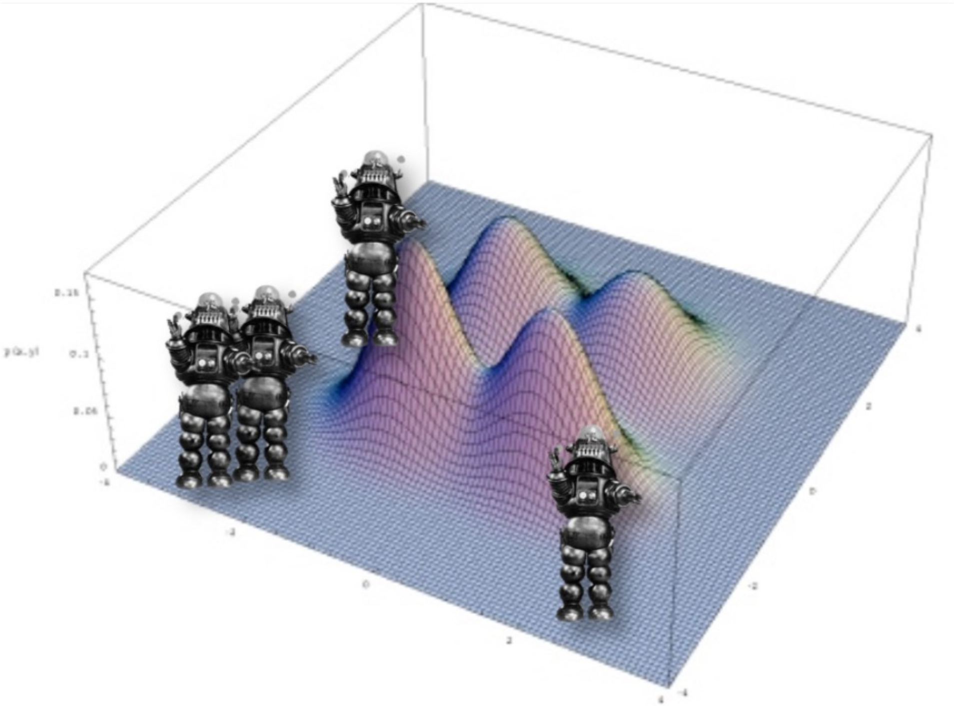
low intensity



From the presentation of Brian Moore (Univ. of Davis)

Approximating the Joint Posterior Probability Density with Metropolis-Couples MCMC

Robot Squadron!!



From the presentation of Brian Moore
(Univ. of Davis)

Summary: Some General Strategies for Assessing MCMC Performance

You can never be absolutely certain that the MCMC is reliable, you can only identify when something has gone wrong. Gelman

1. When do you need to assess MCMC performance?

   ALWAYS

2. When should you assess the performance of individual runs?

   ALWAYS

3. Which diagnostics should you use to assess individual runs?

   ALL that are relevant for the models/parameters you are estimating under

4. When is a single run sufficient to assess MCMC performance?

   NEVER

5. When should you estimate under the prior?

   WHENEVER POSSIBLE (and be wary of programs where it is not possible)

From the presentation of Brian Moore
(Univ. of Davis)

Summary: Some General Strategies for Assessing MCMC Performance

You can never be absolutely certain that the MCMC is reliable, you can only identify when something has gone wrong. Gelman

6. When should you use Metropolis-Coupling?

   Whenever you cannot be certain that standard MCMC is adequate

   i.e., ALWAYS (and be wary of programs where it is not possible)

7. When should you perform multiple independent MCMC runs?

   ALWAYS (and be wary of pseudo-independence)

8. Which diagnostics should you use to assess individual runs?

   ALL that are relevant for the models/parameters you are estimating under

9. How many independent MCMC runs are sufficient?

   AS MANY AS POSSIBLE (i.e., as many as you think your data/problem deserve)

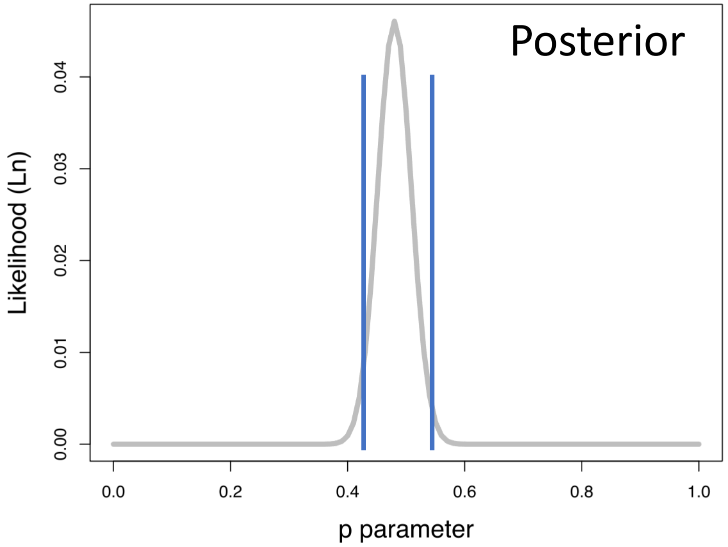
10. How long should you run each MCMC analysis?

    AS LONG AS POSSIBLE (i.e., as long as you think your data/problem deserve)

From the presentation of Brian Moore
(Univ. of Davis)

Credible interval in BI

- Credible interval is an interval within which a parameter value falls with a particular probability
- A measure of the parameter uncertainty



Credible interval 95%

Maximum Likelihood vs. Bayesian Inference

Likelihood:

- Fast
- No priors no subjectivity
- Some types of analyses are challenging due complex likelihood functions

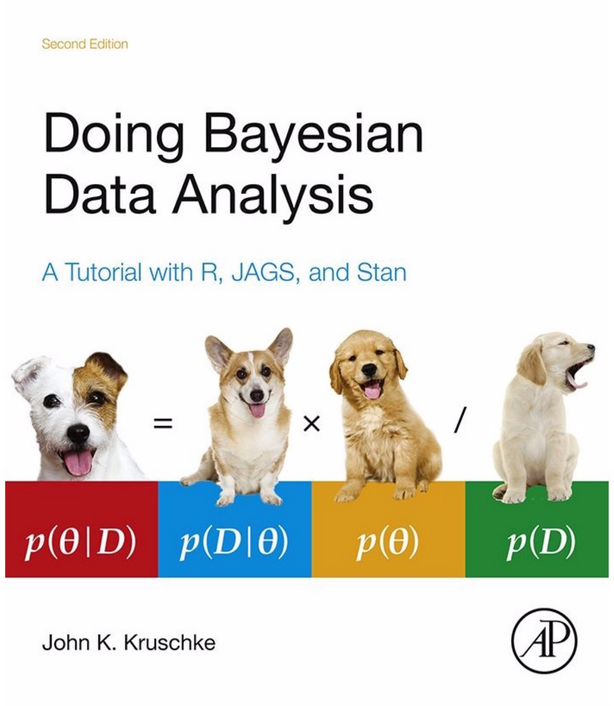
Bayesian:

- Slow
- Priors are logical since everything has a distribution
- Scientists think in a Bayesian way
- Some models can be implemented only in BI. Bayesian non-parametric methods (Dirichlet process prior).

Software for tree reconstruction using BI

- MrBayes: http://nbisweden.github.io/MrBayes/
- RevBayes: https://revbayes.github.io
- Beast: https://www.beast2.org

Suggested literature



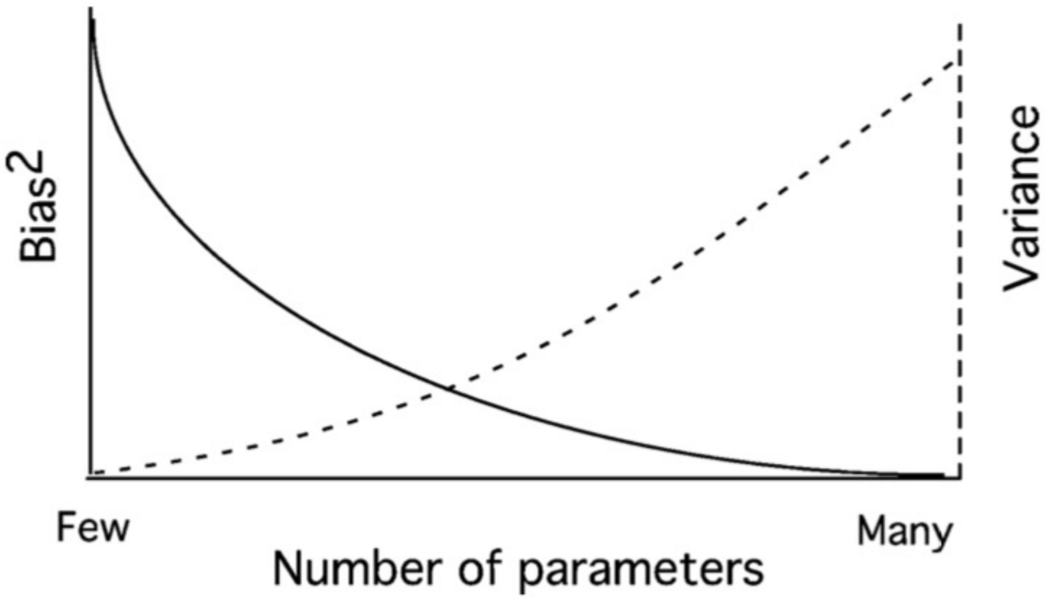
Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan

Quick demo:
Binomial Bayesian Inference

Model selection

Model selection using Maximum Likelihood

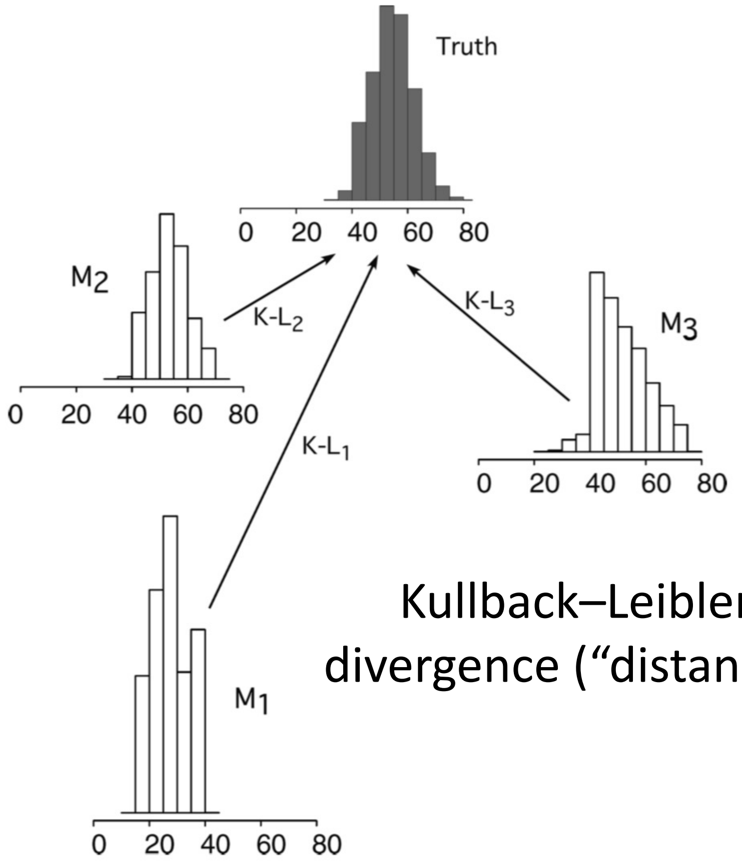
- Different models have different number of parameters



Model selection using Maximum Likelihood

AIC (Akaike information criterion)

- Based on information theory
- AIC estimates the relative amount of information lost by a given model in comparison to the true (unknown) model
- The less information a model loses, the higher the quality of that model



Kullback–Leibler divergence ("distance")

Model selection using Maximum Likelihood

AIC (Akaike information criterion)

- Based on information theory
- AIC estimates the relative amount of information lost by a given model in comparion to the true (unknown) model
- The less information a model loses, the higher the quality of that model
- AIC shows the raltive fit of a model
- The model with a minimum AIC is the best
- Use ΔAIC for comparing multiple models

$$AIC = 2k - 2Ln(L)$$

Where k is the number of the parameters

Delata AIC:

$$\Delta AIC = AIC(M1) - AIC(M2)$$

ΔAIC scale

0 to 2	Not worth more than a bare mention
2 to 6	Positive
6 to 10	Strong
> 10	Very strong

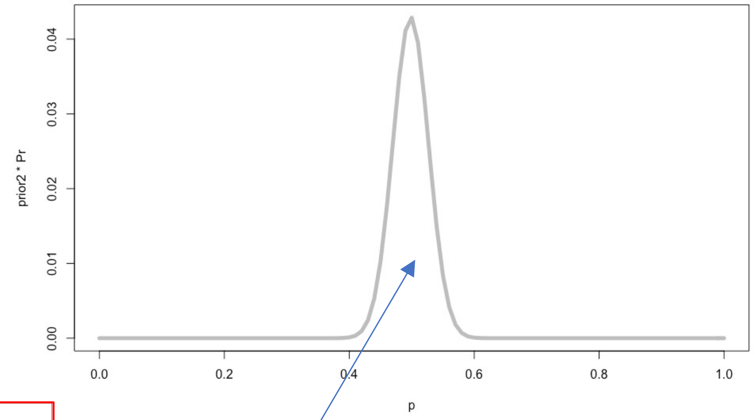
Model selection using Maximum Likelihood

BIC (Bayesian information criterion)

- Motivated by Bayesin thinking but applied to likelihood methods
- BIC approximates the probability of data
- BIC shows the raltive fit of a model
- The model with a minimum BIC in a set of models is the best (= has maximum posterior probability)

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

probability of data (marginal probability)



BIC is the area under *likelihood* function

$$BIC = Log(n)k - 2Ln(L)$$

Where k is the number of the parameters and n is the number of data points

Delata BIC:

$$\Delta BIC = BIC(M1) - BIC(M2)$$

ΔBIC scale

0 to 2	Not worth more than a bare mention
2 to 6	Positive
6 to 10	Strong
> 10	Very strong

Model selection in Bayesian framework

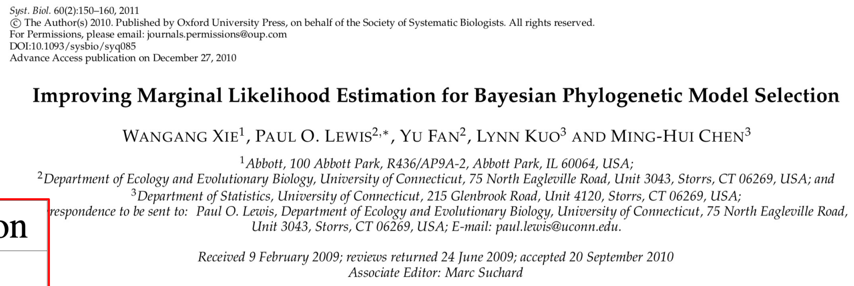
Marginal Likelihood (MLn) and Bayes factor (BF)

- Based on Marginal lekelihood (= probability of data)
- Similar to BIC
- BF is similar to ΔBIC
- BF shows the relative fit of a model
- Marginal likelihood is hard to compute
- Softwares implement special algorithms for computing it (i.e. stepping stone)

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

probability of data
(**marginal probability**)

$$\boxed{BF = MLn(M1) - MLn(M2)}$$


Syst. Biol. 60(2):150–160, 2011

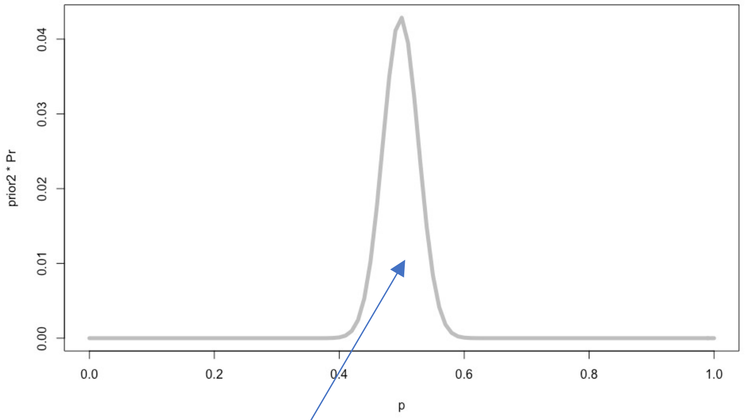
DOI:10.1093/sysbio/syq085
Advance Access publication on December 27, 2010

Improving Marginal Likelihood Estimation for Bayesian Phylogenetic Model Selection

WANGANG XIE[1], PAUL O. LEWIS[2,*], YU FAN[2], LYNN KUO[3] AND MING-HUI CHEN[3]

[1]Abbott, 100 Abbott Park, R436/AP9A-2, Abbott Park, IL 60064, USA;
[2]Department of Ecology and Evolutionary Biology, University of Connecticut, 75 North Eagleville Road, Unit 3043, Storrs, CT 06269, USA; and
[3]Department of Statistics, University of Connecticut, 215 Glenbrook Road, Unit 4120, Storrs, CT 06269, USA;
respondence to be sent to: Paul O. Lewis, Department of Ecology and Evolutionary Biology, University of Connecticut, 75 North Eagleville Road, Unit 3043, Storrs, CT 06269, USA; E-mail: paul.lewis@uconn.edu.

Received 9 February 2009; reviews returned 24 June 2009; accepted 20 September 2010
Associate Editor: Marc Suchard



Marginal likelihood is the area under *posterior distribution* function

$\boxed{BF \text{ scale}}$

0 to 2	Not worth more than a bare mention
2 to 6	Positive
6 to 10	Strong
$>$ 10	Very strong

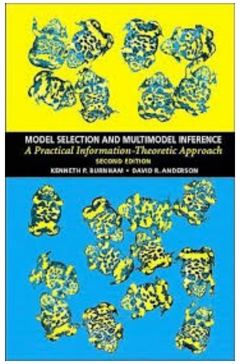
Model selection in practice

- Maximum Likelihood
  - IQ Tree http://www.iqtree.org

- Bayesian
  - MrBayes: http://nbisweden.github.io/MrBayes/
  - RevBayes: https://revbayes.github.io
  - Beast: https://www.beast2.org

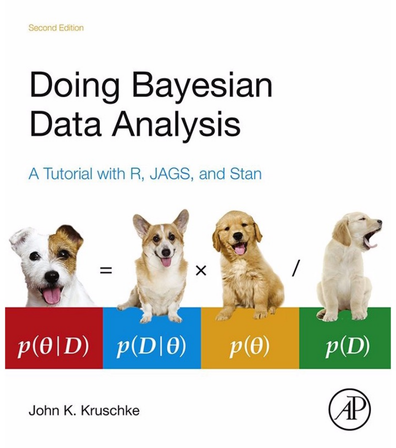
- Old software
  - ParitionFinder http://www.robertlanfear.com/partitionfinder/
  - ModelTest-NG v0.1.5 https://github.com/ddarriba/modeltest/releases

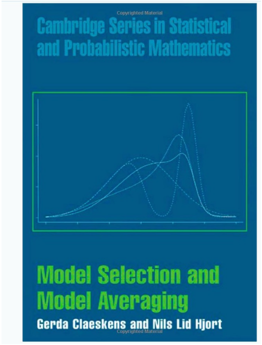
Suggested literature



Model Selection by Burham and Anderson

Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan





Model Selection and Model Averaging

Summary

- Bayesian Inference is a natural extension of Likelihood method for estimating posterior probability of parameters
- Model selection tools allow testing various hypotheses in Maximum Likelihood and Bayesian frameworks