Phylogeny and bioinformatics for evolution

Bayesian method

September, 2007

Lecture outline

Bayesian setting
Definition
Simple example

MCMC
Normalizing constant
MCMC in practice
Proposal distribution
Prior distribution
Posterior probabilities
Application of Bayesian methods

Demo: McRobot (P. Lewis)

Lecture outline

Lecture outline

©N. Salamin
Sept 2007

Lecture outline
Bayesian setting
Definition
Simple example
MCMC
Normalizing constant
MCMC in practice
Proposal distribution
Prior distribution
Posterior probabilities
Application of Bayesian methods
Demo: McRobot (P. Lewis)

Bayesian methods

The likelihood calculation is used as well by Bayesian methods. However, another component is added to the method: the prior distributions.

Before observing any data, each parameter will be assigned a prior distribution

- topologies
- branch lengths
- each parameter of the model of evolution

The prior distributions are then combined with the likelihood of the data to give the posterior distribution.

This is a highly attractive quantity because it computes what we most need: the probabilities of different hypotheses in the light of the data.

Bayes theorem

To combine all this together, we use the Bayes theorem

$$Prob(T|D) = \frac{Prob(T \cup D)}{Prob(D)}$$

where $Prob(T \cup D) = Prob(T)Prob(D|T)$

so that

$$Prob(T|D) = \frac{Prob(T)Prob(D|T)}{Prob(D)}$$

©N. Salamin
Sept 2007

Lecture outline
Bayesian setting
Definition
Simple example
MCMC
Normalizing constant
MCMC in practice
Proposal distribution
Prior distribution
Posterior probabilities
Application of Bayesian methods
Demo: McRobot (P. Lewis)

Normalizing constant

The denominator $Prob(D)$ is the sum of the numerator $Prob(T)Prob(D|T)$ over all possible trees T.

This quantity is needed to normalize the probabilities of all T so that they add up to 1.

This leads to

$$Prob(T|D) = \frac{Prob(T)Prob(D|T)}{\sum_T Prob(T)Prob(D|T)}$$

In words:

$$\text{posterior probability} = \frac{\text{prior probability} \times \text{likelihood}}{\text{normalizing constant}}$$

Odds ratio

Bayes theorem can be put in the form of odds-ratio, which is the odds favoring one hypothesis over another

- odds a person has initially (the prior odds)
- multiplied by likelihood ratio under the data

- suppose we favor, in advance, T_1 over T_2 with odds $3:2$
- some data gives $Prob(D|T_1)/Prob(D|T_2) = 1/2$
- data say that T_1 is half as probable than T_2
- posterior odds ratio $(3/2) \times (1/2) = 3/4$

After looking at the data, we favor T_2 over T_1 by a factor of $4:3$

Coin toss

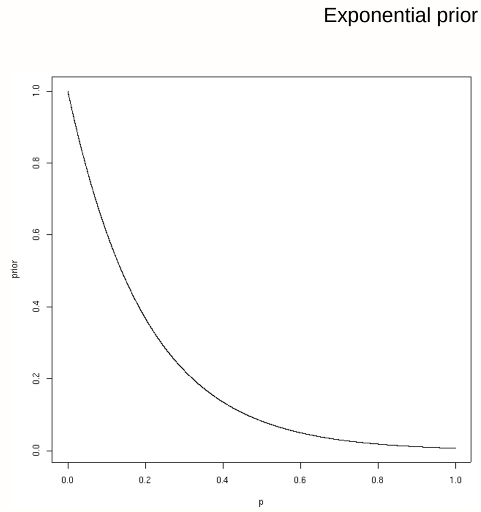
We want to estimate p, the probability of obtaining head, by tossing a coin n times, which results in n_h heads and n_t tails

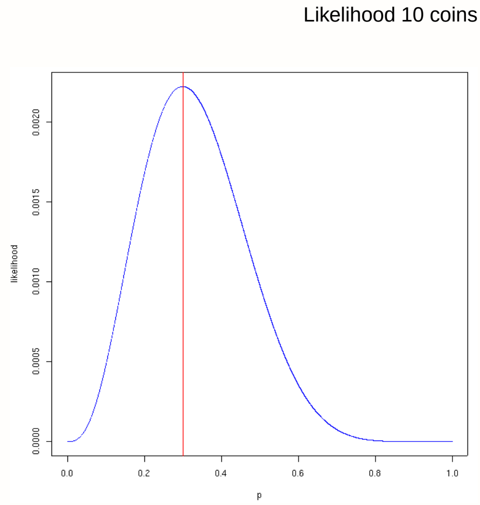
- binomial distribution to calculate the likelihood of p

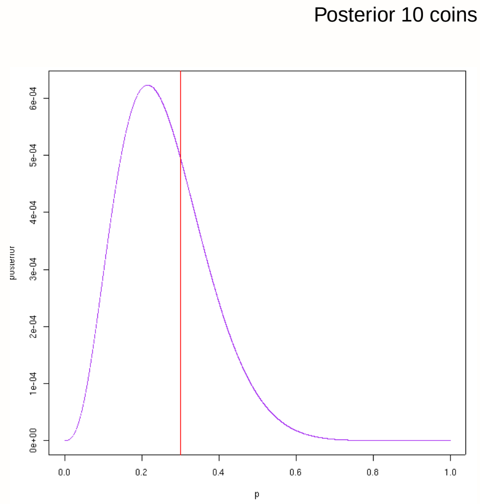
$$\mathcal{B}(n, n_h, p) = \binom{n}{n_h} p^{n_h} (1-p)^{n-n_h}$$

- we make two trials of 10 and 1000 draws resulting in
  - 3 heads and 7 tails
  - 300 heads and 700 tails

Exponential prior 10 tosses

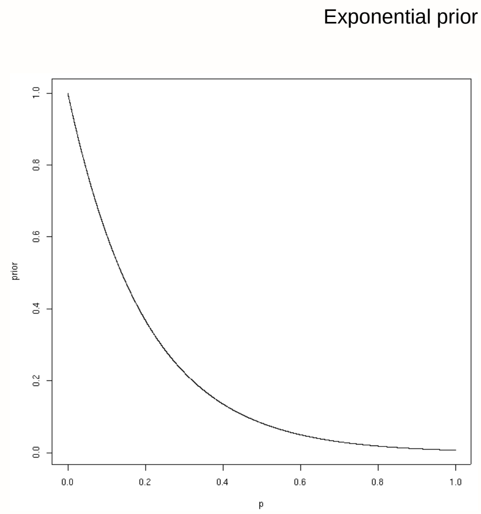


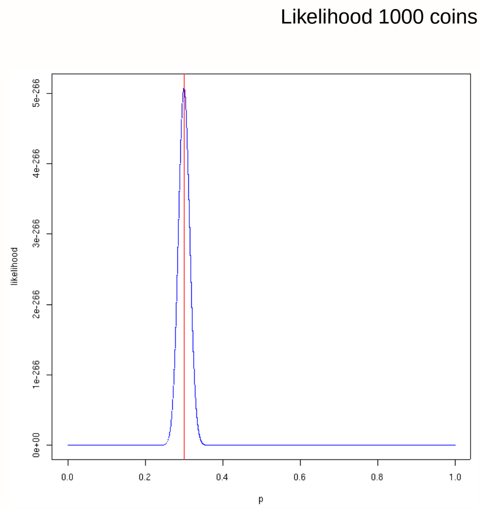


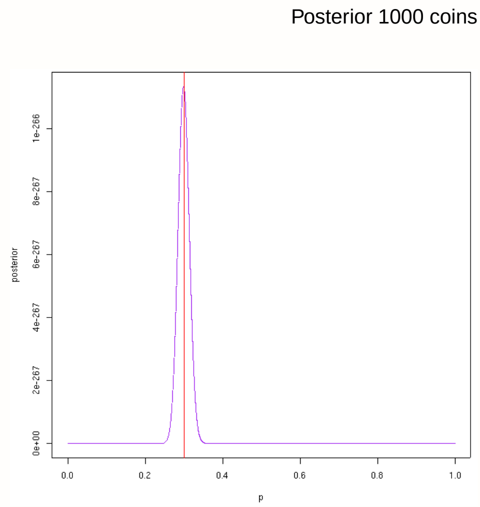


Lecture outline
Bayesian setting
Definition
Simple example
MCMC
Normalizing constant
MCMC in practice
Proposal distribution
Prior distribution
Posterior probabilities
Application of Bayesian methods
Demo: McRobot (P. Lewis)

Exponential prior 1000 tosses

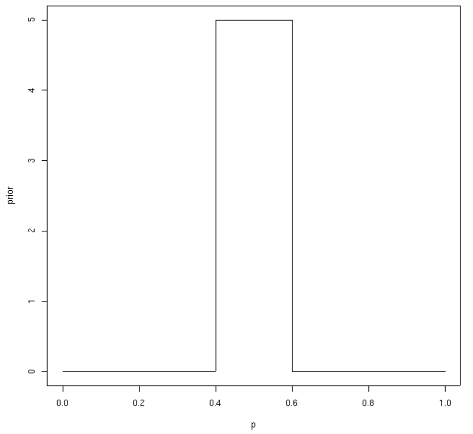




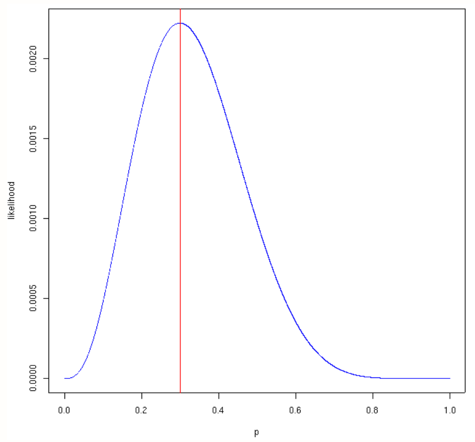


Flat prior 10 tosses

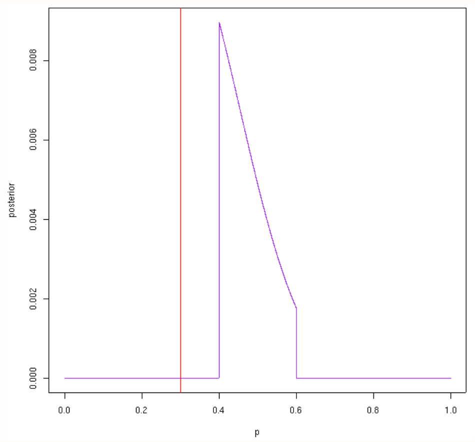
Exponential prior



Likelihood 10 coins



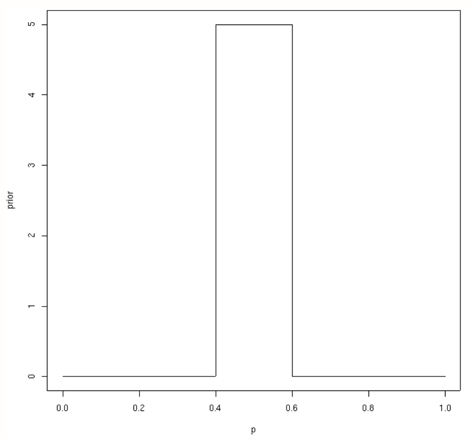
Posterior 10 coins



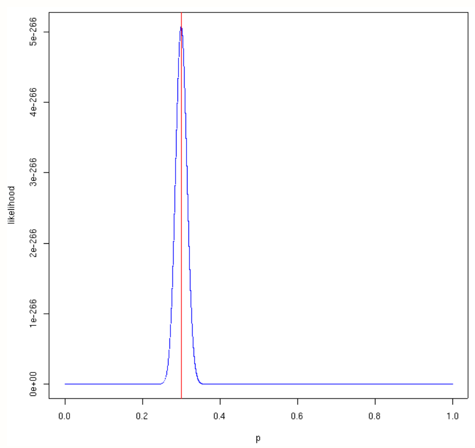
Lecture outline
Bayesian setting
Definition
Simple example
MCMC
Normalizing constant
MCMC in practice
Proposal distribution
Prior distribution
Posterior probabilities
Application of Bayesian methods
Demo: McRobot (P. Lewis)

Flat prior 1000 tosses

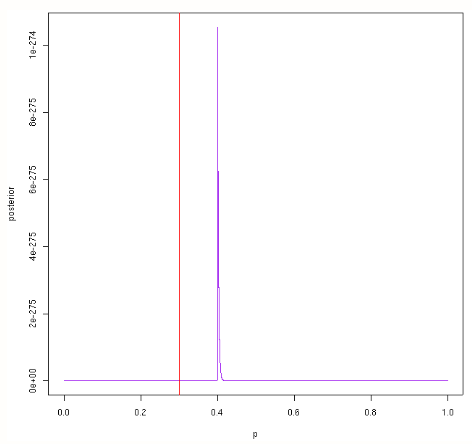
Exponential prior



Likelihood 1000 coins



Posterior 1000 coins



Lecture outline

©N. Salamin
Sept 2007

Lecture outline
Bayesian setting
Definition
Simple example
MCMC
Normalizing constant
MCMC in practice
Proposal distribution
Prior distribution
Posterior probabilities
Application of Bayesian methods
Demo: McRobot (P. Lewis)

Estimating normalizing constant

Posterior distribution expression has a denominator, i.e. $\sum_T Prob(T)Prob(D|T)$, that is often impossible to compute.

Fortunately, samples from the posterior distribution can be drawn using a Markov chain that does not need to know the denominator

- draw a random sample from posterior distribution of trees
- becomes possible to make probability statements about true tree
- e.g. if 96% of the samples from posterior distribution have (human,chimp) as monophyletic group, probability of this group is 96%

Makov chain Monte Carlo

Idea: to wander randomly through tree space by sampling trees until we settle down into an equilibrium distribution of trees that has the desired distribution, i.e. posterior distribution.

- Markov chain: the new proposed tree will depend only on the previous one
- to reach equilibrium distribution, the Markov chain must be
  - aperiodic – no cycles should be present in the Markov chain
  - irreducible – every trees must be accessible from any other tree
  - probability of proposing T_j when we are at T_i is the same as probability of proposing T_i when we are at T_j
- the Markov chain has no end

Lecture outline
Bayesian setting
Definition
Simple example
MCMC
Normalizing constant
MCMC in practice
Proposal distribution
Prior distribution
Posterior probabilities
Application of Bayesian methods
Demo: McRobot (P. Lewis)

MCMC in practice

Metropolis algorithm

- start with a random tree T_i
- select a new tree T_j by modifying T_i in some way
- compute

$$R = \frac{Prob(T_j|D)}{Prob(T_i|D)}$$

the normalizing constant being the same, this is

$$R = \frac{Prob(T_j)Prob(D|T_j)}{Prob(T_i)Prob(D|T_i)}$$

- if $R \geq 1$, accept T_j
- if $R < 1$, draw a random number n between $[0, 1]$ and accept T_j if $R > n$, otherwise keep T_i

How to propose a new tree

We could invent any type of proposal distribution to wander through the tree space

- e.g. NNI by selecting a node at random
- erase part of the tree and propose new branch lengths
- should be able to reach all trees from any starting tree
- at least after "sufficient" running, but impossible to know how much running is enough

Should be careful because

- if trees proposed are too different ⇒ these trees will be rejected too often
- if trees proposed are too similar ⇒ tree space won't be sampled well enough

Type of prior distributions

Prior distributions for topologies have been proposed

- stochastic process of random speciation and extinction
- uniform distribution of all possible rooted trees

Prior distributions on branch lengths

- exponential distribution
- uniform distribution

Prior on model parameters

- Dirichlet distribution on nucleotide frequency
- uniform or exponential distribution on shape of Γ distribution
- uniform distribution on proportion of invariant sites

Problems with priors

As shown before, we have to be careful with prior because they can exclude possible estimated values.

Other problematic aspects:

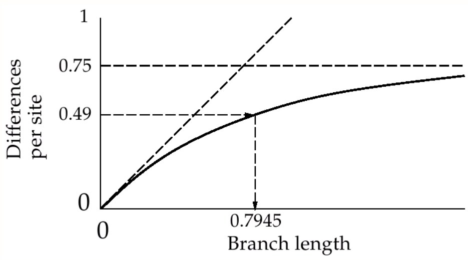
- universality of priors
- use of “uninformative” flat priors
  - issues of scale
  - unbounded quantities

©N. Salamin
Sept 2007

Lecture outline
Bayesian setting
Definition
Simple example
MCMC
Normalizing constant
MCMC in practice
Proposal distribution
Prior distribution
Posterior probabilities
Application of Bayesian methods
Demo: McRobot (P. Lewis)

Jukes-Cantor model

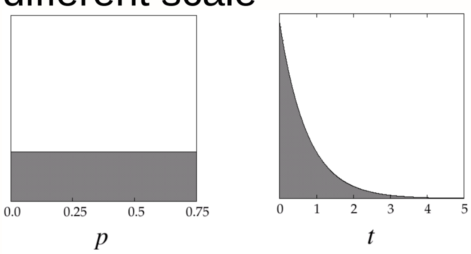
figures from Felsenstein 2004



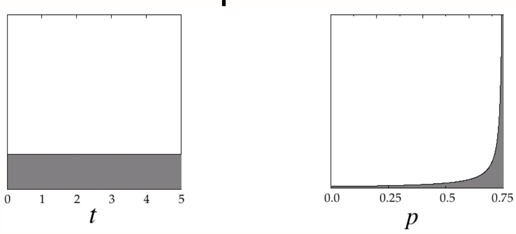
Sequence disimilarity and branch length under this model:

$$p = \frac{3}{4}\left(1 - e^{-\frac{4}{3}t}\right)$$

different scale



unbounded priors



Summarizing posterior

What is the posterior probability of each tree?

- do we take branch length into account?
- if noise in data, no single tree will have high probability
- what if only part of the tree is supported?

We therefore have to take clade probabilities

- for clade of interest, sum the posterior probabilities of all trees containing that clade
- but clade prior distribution is not clearly defined (yet!)
- might be the reason why posterior probabilities are always larger than bootstrappercentages

To have good estimation of posterior probabilities, we have to sample the Markov chain for long enough to reach equilibrium.

What can we do with Bayesian

Beside their use in building phylogenetic trees, Bayesian methods are useful to deal with complex biological problems

- testing hypotheses about rates of host switching and cospeciation in host/parasite systems
- dating phylogenetic trees using autocorrelated prior distribution on rates of evolution
- infer rate of change of states of a character and the bias in the rate of gain of the character
- infer accuracy of inference of ancestral states
- infer position of the root of the tree
- testing rates of speciation and assess key innovations associated with changes in rates

Lecture outline

1. Bayesian setting
   - Definition
   - Simple example
2. Markov chain Monte Carlo
   - Normalizing constant
   - MCMC in practice
   - Proposal distribution
   - Prior distribution
   - Posterior probabilities
   - Application of Bayesian methods
3. Demo: McRobot (P. Lewis)