Review Questions

What is a Markov chain? What is the Markov property?

What are the differences between discrete- and continuous-time Markov chains?

What is the state space for a typical phylogenetic model?

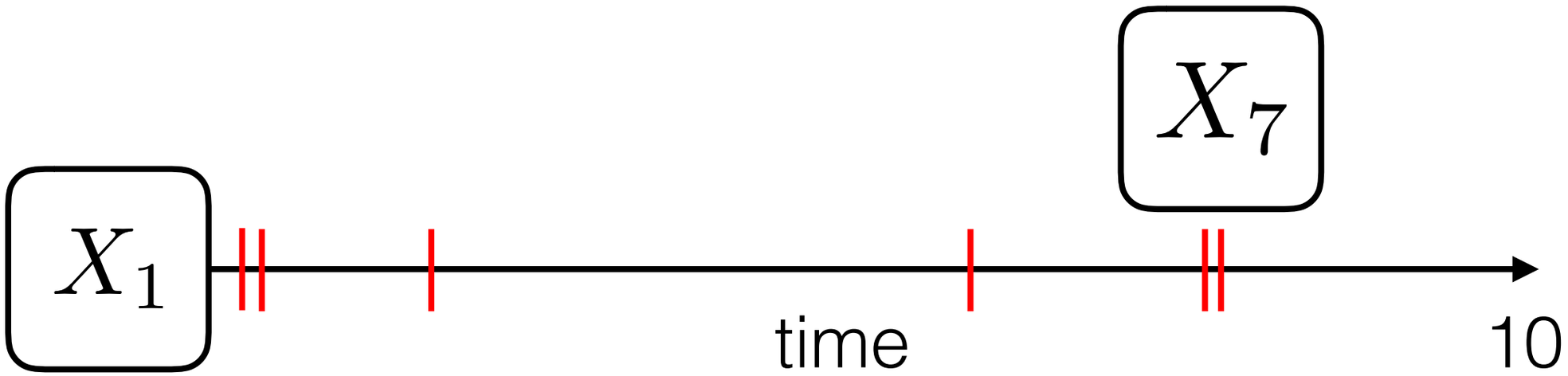
What does the stationary distribution tell us?

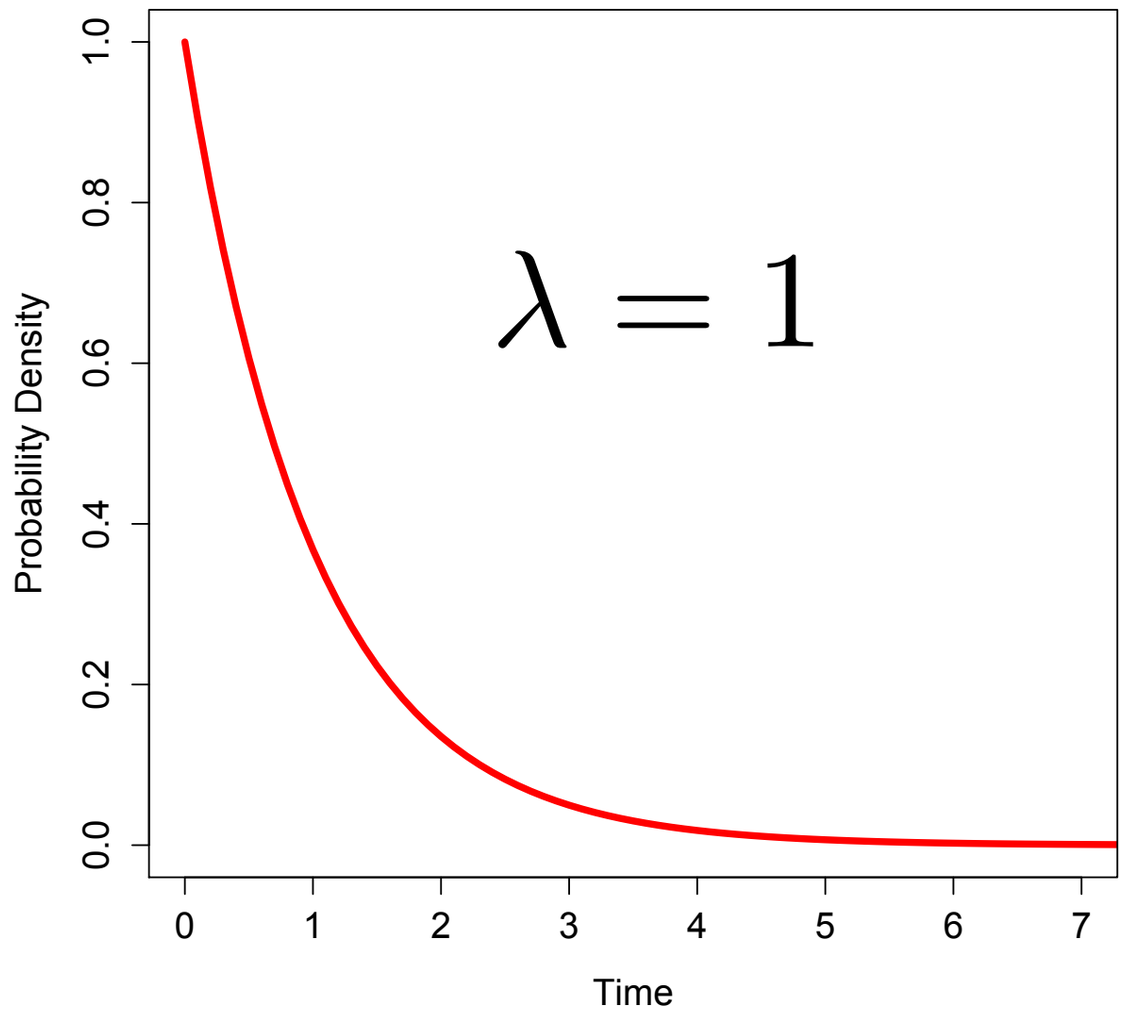
What does it mean for a chain to be reversible?

Simulating Sequence Evolution Assignment

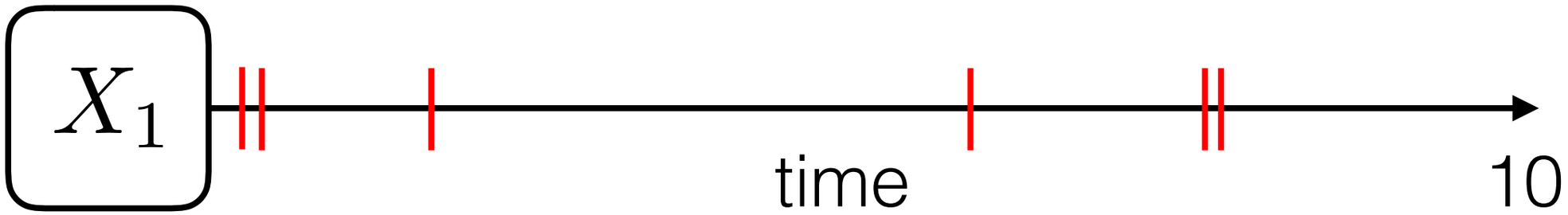
PSEUDOCODE! What are the steps?

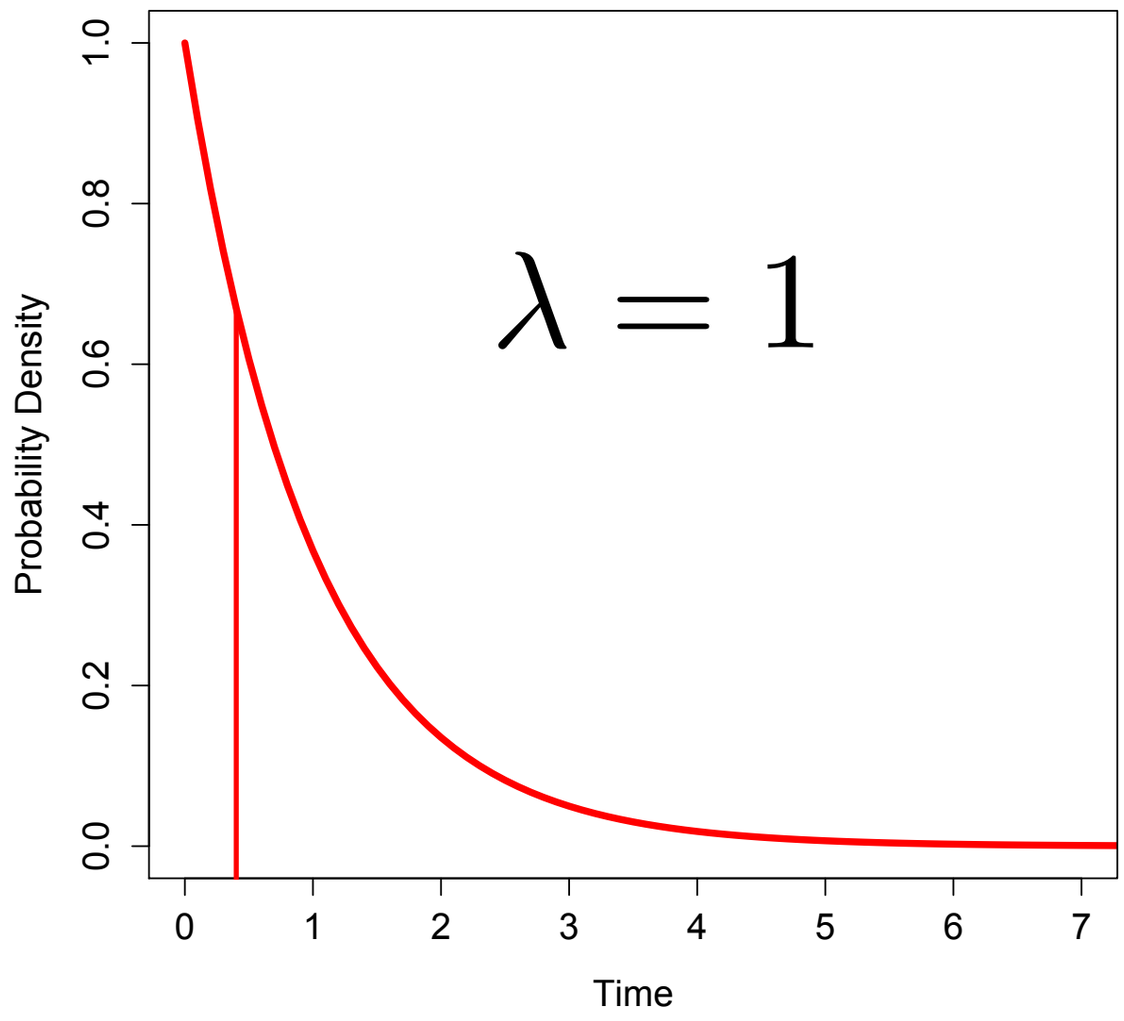
Continuous-Time Markov Chains



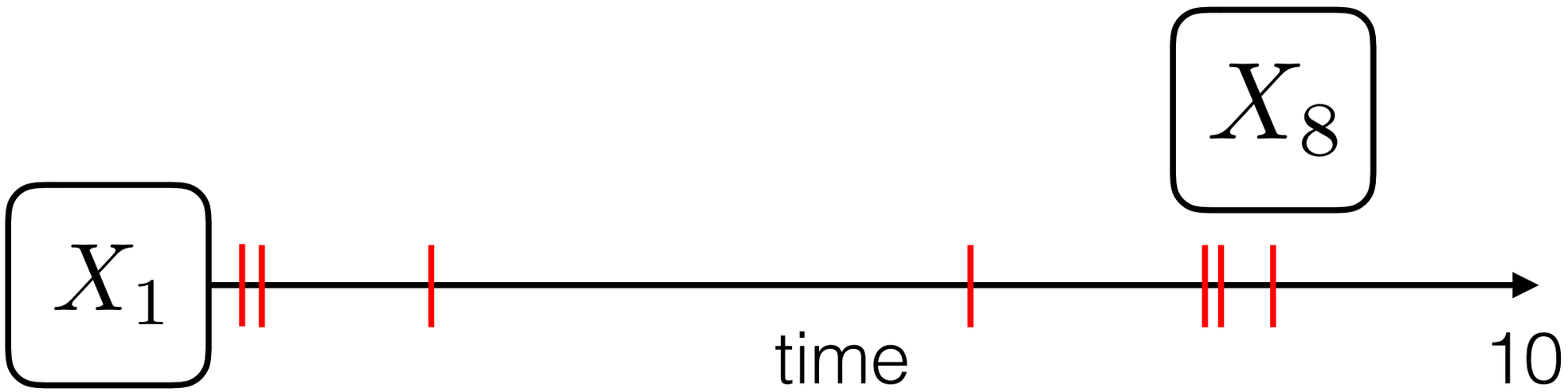


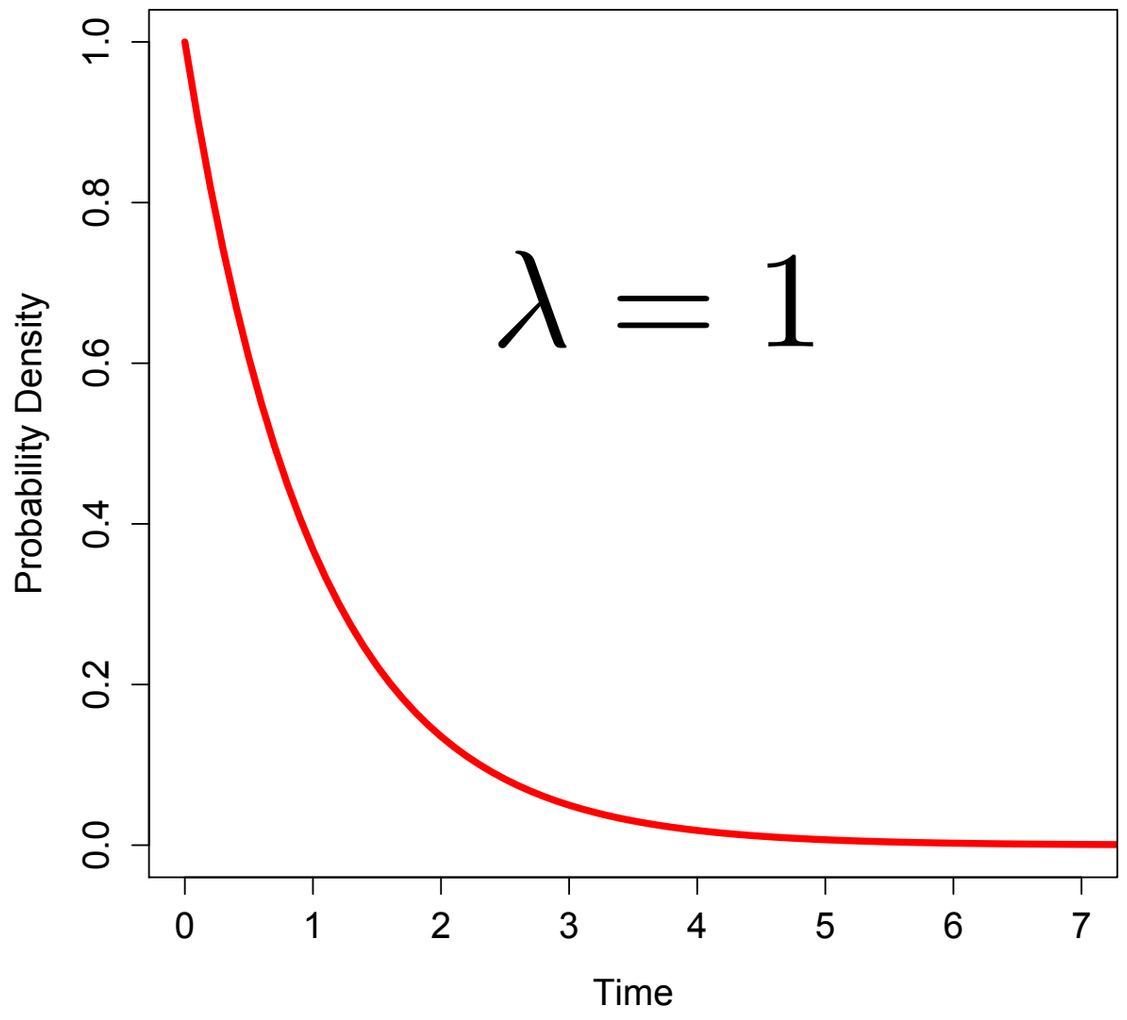
Continuous-Time Markov Chains



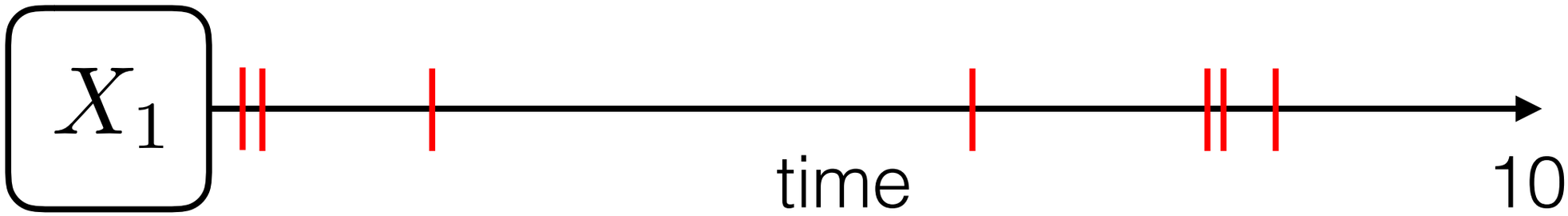


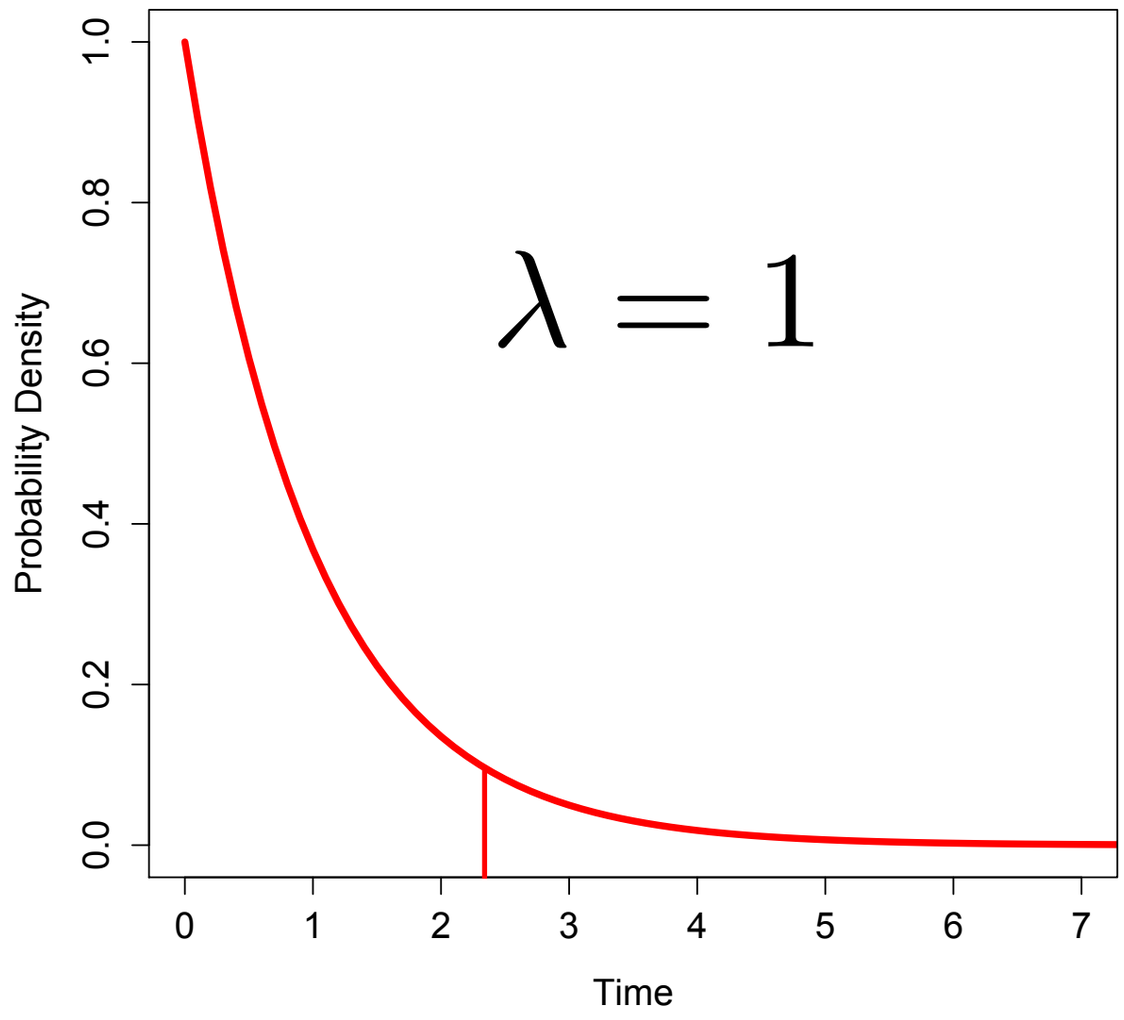
Continuous-Time Markov Chains



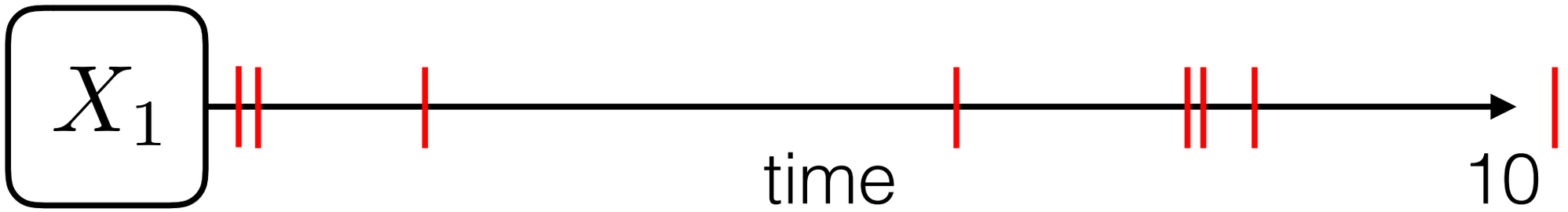


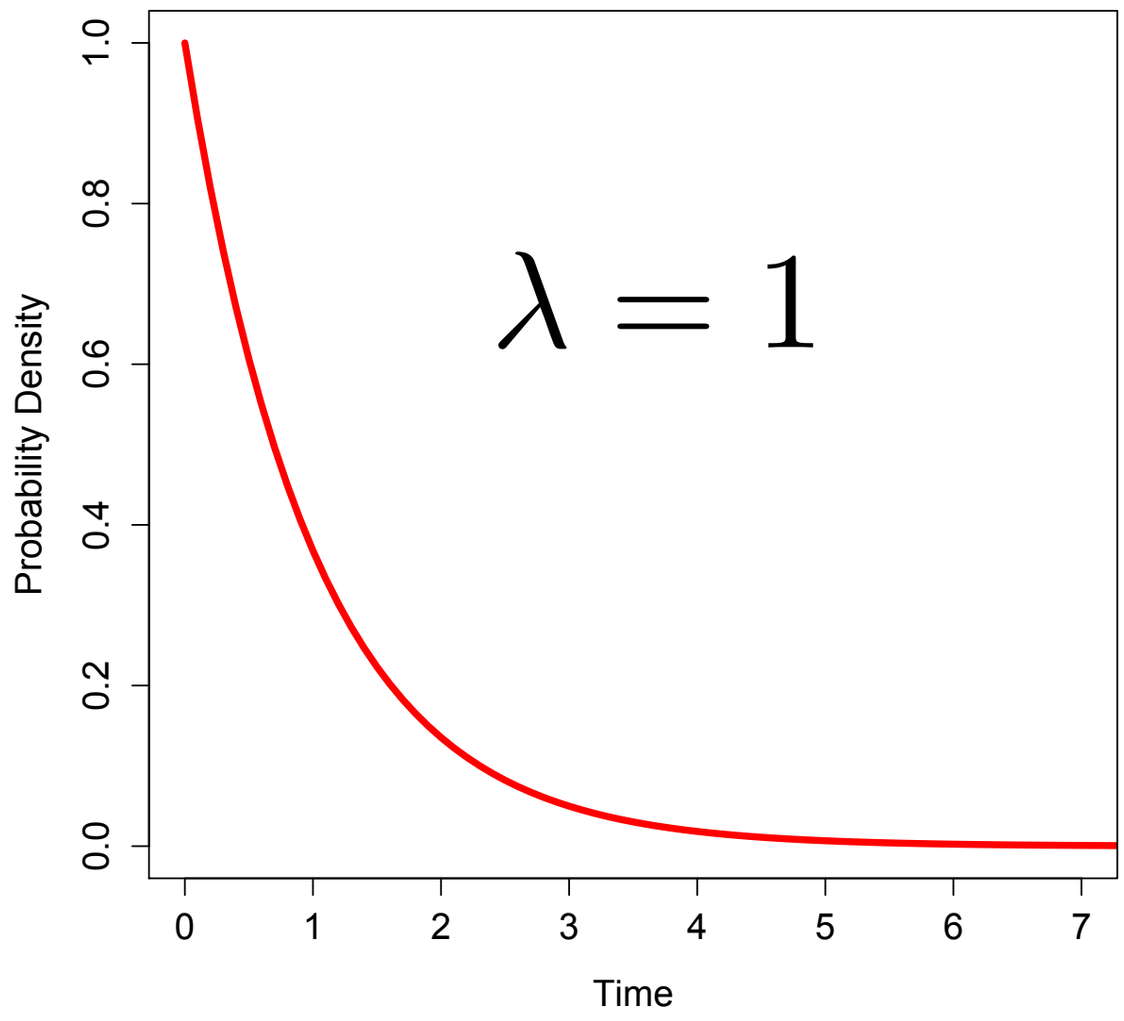
Continuous-Time Markov Chains



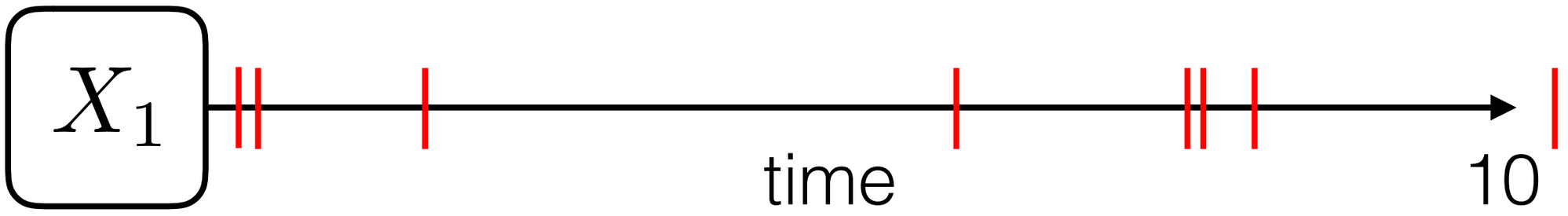


Continuous-Time Markov Chains





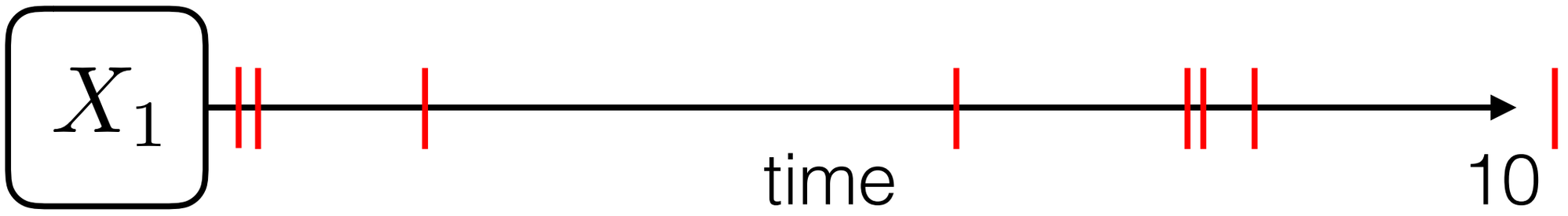
Continuous-Time Markov Chains



Simulation stops when the total waiting time exceeds the length of the simulation (for phylogenetics, think of this as a branch length).

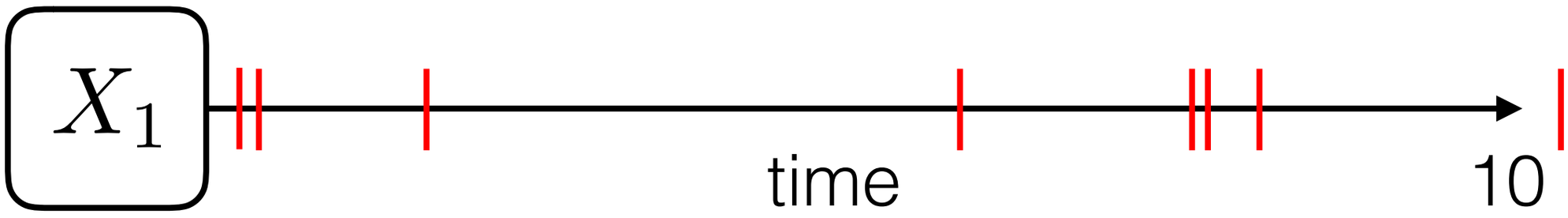
The state of the chain at the end of the simulation is the state last sampled before the branch length was exceeded (X_8, in this case).

Continuous-Time Markov Chains



If the rate parameter (λ) is 1, what is the expected number of events in a Poisson distribution? What does this mean if we're thinking about a phylogenetic tree with branch lengths?

Poisson Process



A continuous-time Markov chain is one type of a **Poisson process**, which has these properties:

(1) The number of events that occur in an interval of length t is a Pois(λt) random variable.

(2) The number of events that occur in disjoint intervals are independent of one another.

(3) The waiting times between events are distributed as: $T \sim Exp(\lambda)$

CTMC Rate Matrix

$$
\begin{array}{c} \\ A \\ C \\ G \\ T \end{array}
\begin{array}{c}
\begin{array}{cccc} A & C & G & T \end{array} \\
\begin{pmatrix}
-1 & 1/3 & 1/3 & 1/3 \\
1/3 & -1 & 1/3 & 1/3 \\
1/3 & 1/3 & -1 & 1/3 \\
1/3 & 1/3 & 1/3 & -1
\end{pmatrix}
\end{array}
$$

Jukes and Cantor (1969)

CTMC Rate Matrix

$$
\begin{array}{c} \\ A \\ C \\ G \\ T \end{array}
\begin{array}{c}
\begin{array}{cccc} A & C & G & T \end{array} \\
\begin{pmatrix}
-1 & 1/3 & 1/3 & 1/3 \\
1/3 & -1 & 1/3 & 1/3 \\
1/3 & 1/3 & -1 & 1/3 \\
1/3 & 1/3 & 1/3 & -1
\end{pmatrix}
\end{array}
$$

Jukes and Cantor (1969)

CTMC Rate Matrix

Exponential Rate for State A

$$
\begin{array}{c} \\ A \\ C \\ G \\ T \end{array}
\begin{array}{c}
\begin{array}{cccc} A & C & G & T \end{array} \\
\begin{pmatrix}
-1 & 1/3 & 1/3 & 1/3 \\
1/3 & -1 & 1/3 & 1/3 \\
1/3 & 1/3 & -1 & 1/3 \\
1/3 & 1/3 & 1/3 & -1
\end{pmatrix}
\end{array}
$$

Jukes and Cantor (1969)

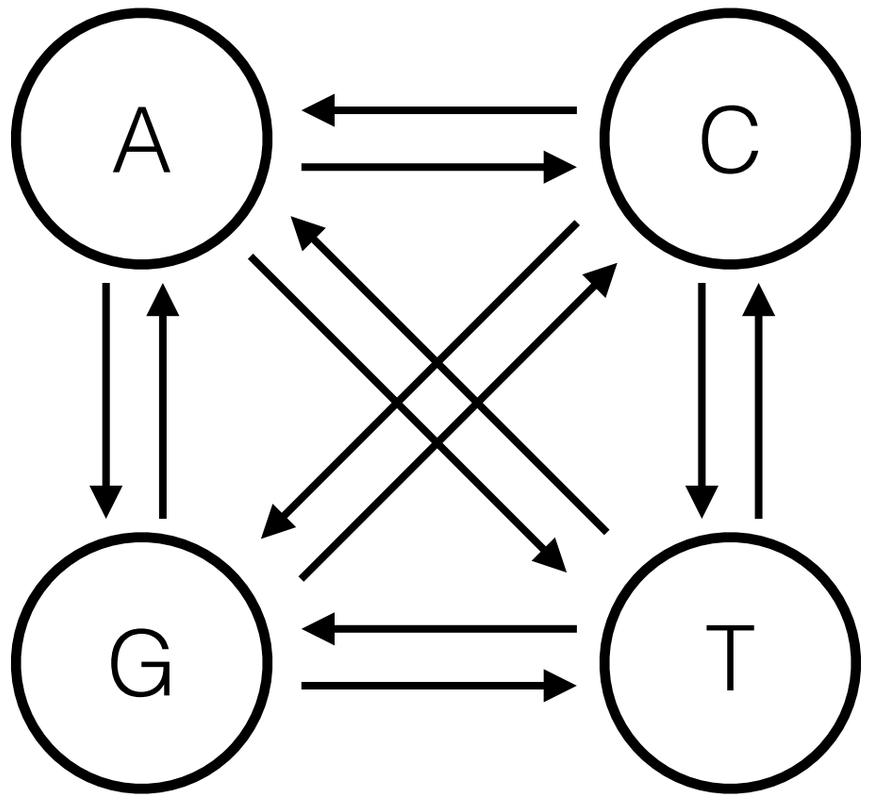
CTMC Rate Matrix

Relative Probabilities of Transition

$$
\begin{array}{c} \\ A \\ C \\ G \\ T \end{array}
\begin{array}{c}
\begin{array}{cccc} A & C & G & T \end{array} \\
\begin{pmatrix}
-1 & 1/3 & 1/3 & 1/3 \\
1/3 & -1 & 1/3 & 1/3 \\
1/3 & 1/3 & -1 & 1/3 \\
1/3 & 1/3 & 1/3 & -1
\end{pmatrix}
\end{array}
$$

Jukes and Cantor (1969)

Jukes and Cantor (1969)



CTMC Rate Matrix

$$
\begin{array}{c} \\ A \\ C \\ G \\ T \end{array}
\begin{array}{c}
\begin{array}{cccc} A & C & G & T \end{array} \\
\begin{pmatrix}
 & \pi_C r_{AC} & \pi_G r_{AG} & \pi_T r_{AT} \\
\pi_A r_{AC} & & \pi_G r_{CG} & \pi_T r_{CT} \\
\pi_A r_{AG} & \pi_C r_{CG} & & \pi_T r_{GT} \\
\pi_A r_{AT} & \pi_C r_{CT} & \pi_G r_{GT} &
\end{pmatrix}
\end{array}
$$

General Time Reversible (GTR; Tavaré 1986)

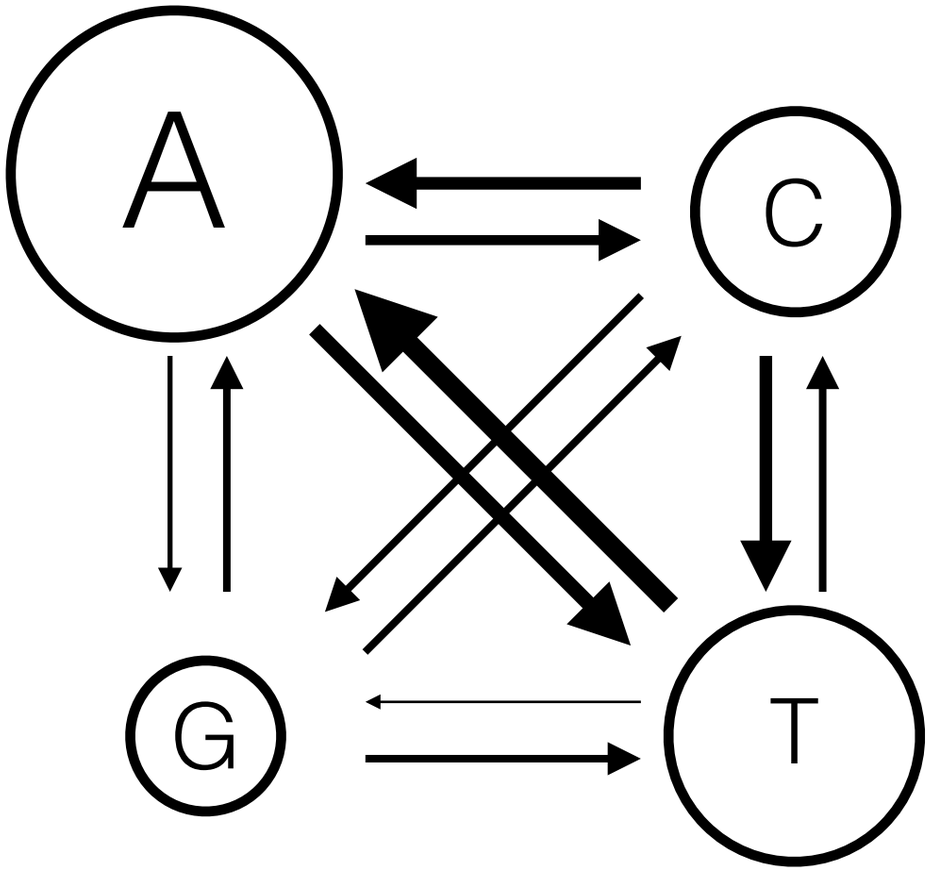
CTMC Rate Matrix

$$\pi = (\pi_A, \pi_C, \pi_G, \pi_T) \quad R = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} & r_{AC} & r_{AG} & r_{AT} \\ & & r_{CG} & r_{CT} \\ & & & r_{GT} \\ & & & \end{pmatrix} \end{array}$$

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} & \pi_C r_{AC} & \pi_G r_{AG} & \pi_T r_{AT} \\ \pi_A r_{AC} & & \pi_G r_{CG} & \pi_T r_{CT} \\ \pi_A r_{AG} & \pi_C r_{CG} & & \pi_T r_{GT} \\ \pi_A r_{AT} & \pi_C r_{CT} & \pi_G r_{GT} & \end{pmatrix} \end{array}$$

General Time Reversible (GTR; Tavaré 1986)

General Time Reversible (GTR)



Reversibility

$$\pi_i q_{ij} = \pi_j q_{ji}, \text{ for all } i \neq j$$

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} & \pi_C r_{AC} & \pi_G r_{AG} & \pi_T r_{AT} \\ \pi_A r_{AC} & & \pi_G r_{CG} & \pi_T r_{CT} \\ \pi_A r_{AG} & \pi_C r_{CG} & & \pi_T r_{GT} \\ \pi_A r_{AT} & \pi_C r_{CT} & \pi_G r_{GT} & \end{pmatrix} \end{array}$$

$$\text{if } i = \text{A and } j = \text{C},$$

$$\pi_A q_{AC} = \pi_C q_{CA}$$

$$\pi_A \pi_C r_{AC} = \pi_C \pi_A r_{AC}$$

Branch-Length Scaling

$$
\begin{array}{c} \\ A \\ C \\ G \\ T \end{array}
\begin{array}{c}
\begin{array}{cccc} A & C & G & T \end{array} \\
\begin{pmatrix}
 & \pi_C r_{AC} & \pi_G r_{AG} & \pi_T r_{AT} \\
\pi_A r_{AC} & & \pi_G r_{CG} & \pi_T r_{CT} \\
\pi_A r_{AG} & \pi_C r_{CG} & & \pi_T r_{GT} \\
\pi_A r_{AT} & \pi_C r_{CT} & \pi_G r_{GT} &
\end{pmatrix}
\end{array}
$$

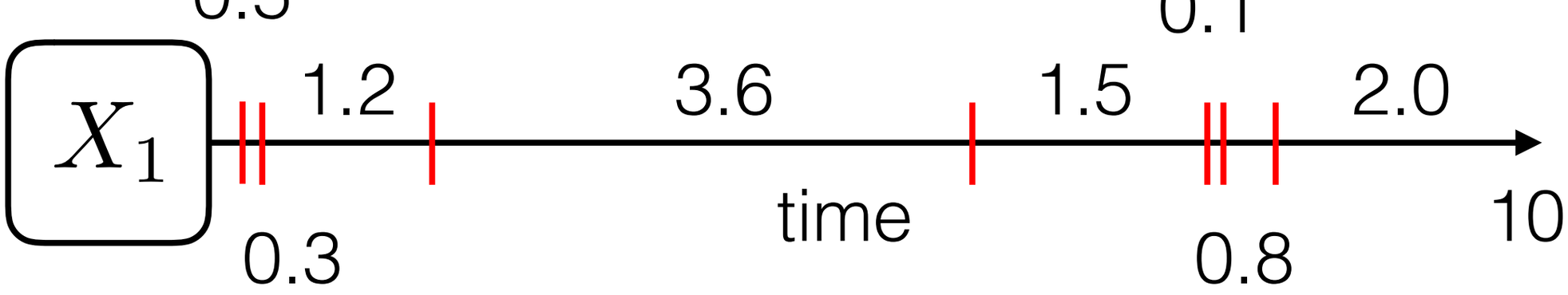
Must average to 1

Branch lengths typically denote **expected number of substitutions**. For this to be true, the weighted mean across all states must be 1.

Probability of a Character History

Jukes-Cantor

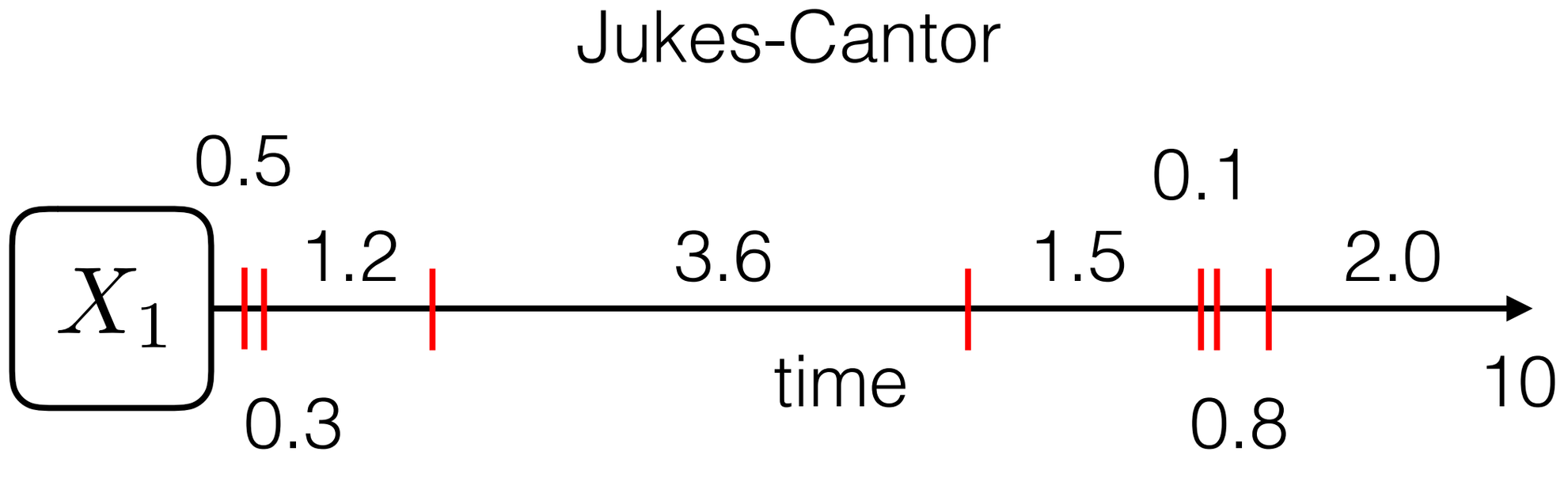
Instead of thinking about simulating a character history, let's calculate the probability of one that we've been given.



The events in this history are independent. When events are independent how do we combine their probabilities?

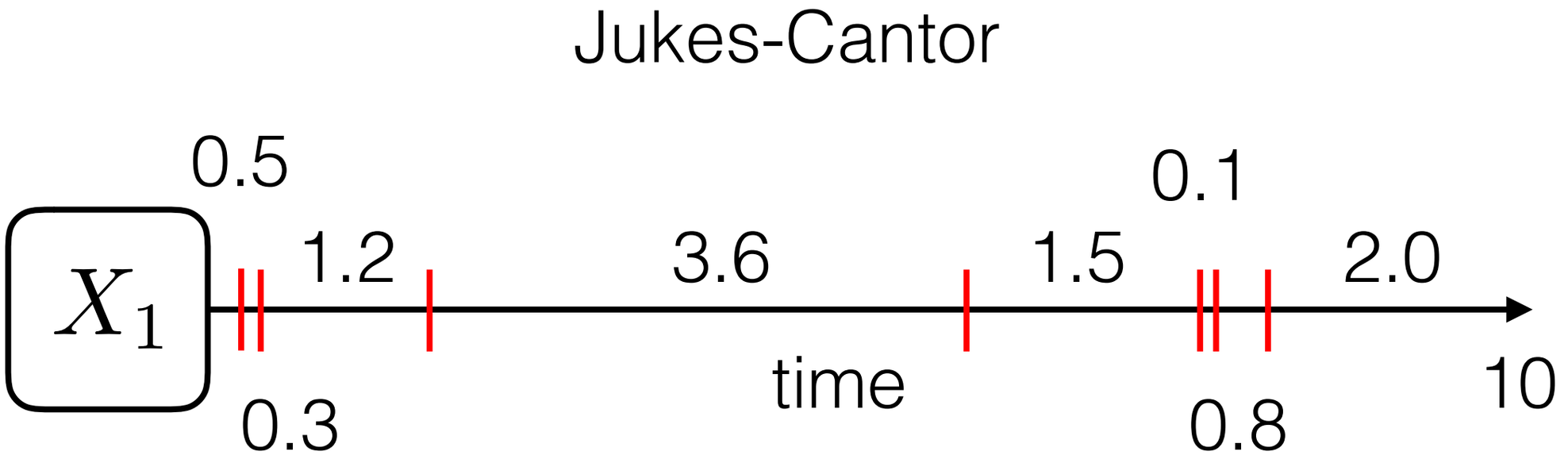
What are the relevant probabilities here?

Probability of a Character History



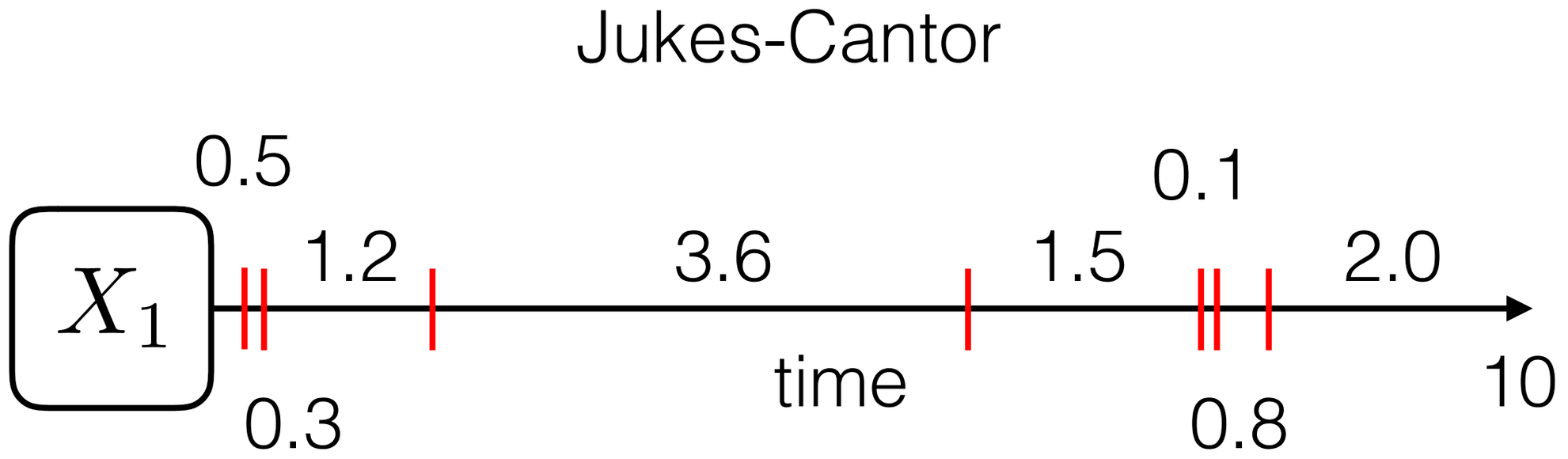
Prob(history) = $Exp(0.5)Exp(0.3)Exp(1.2)Exp(3.6)Exp(1.5)Exp(0.1)Exp(0.8)e^{-2.0}(1/3)^7$

Probability of a Character History



$$\text{Prob(history)} = \underbrace{Exp(0.5)Exp(0.3)Exp(1.2)Exp(3.6)Exp(1.5)Exp(0.1)Exp(0.8)}_{\text{Waiting Times}}e^{-2.0}(1/3)^7$$

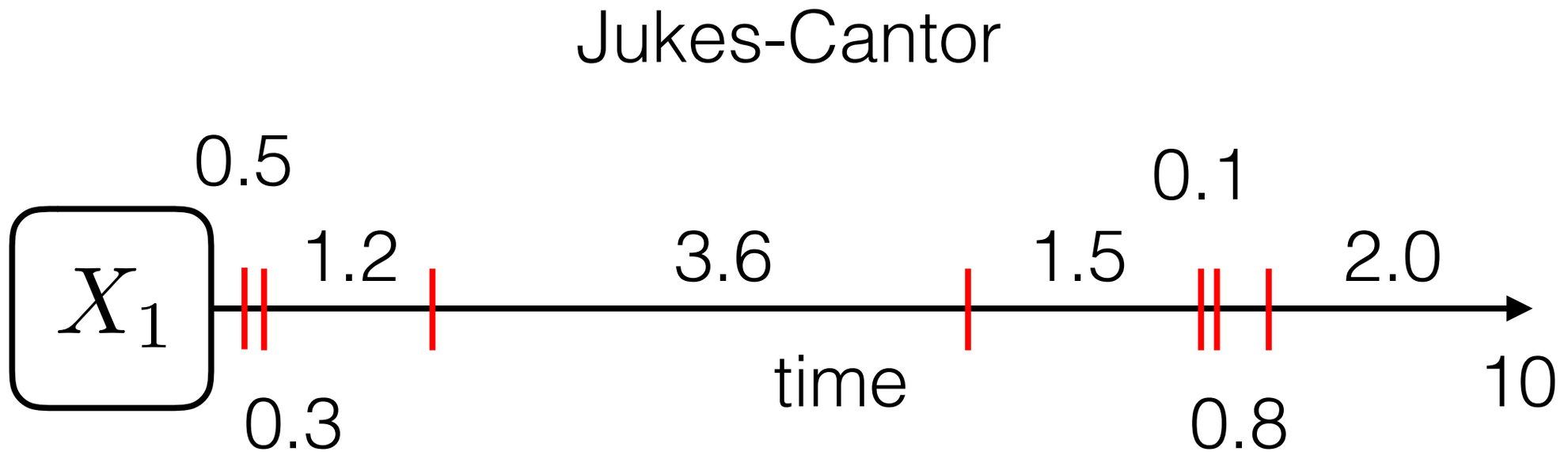
Probability of a Character History



$$\text{Prob(history)} = \underbrace{Exp(0.5)Exp(0.3)Exp(1.2)Exp(3.6)Exp(1.5)Exp(0.1)Exp(0.8)}_{\text{Waiting Times}}\underbrace{e^{-2.0}}_{\substack{\text{Prob No} \\ \text{Change} \\ \text{at End}}}(1/3)^7$$

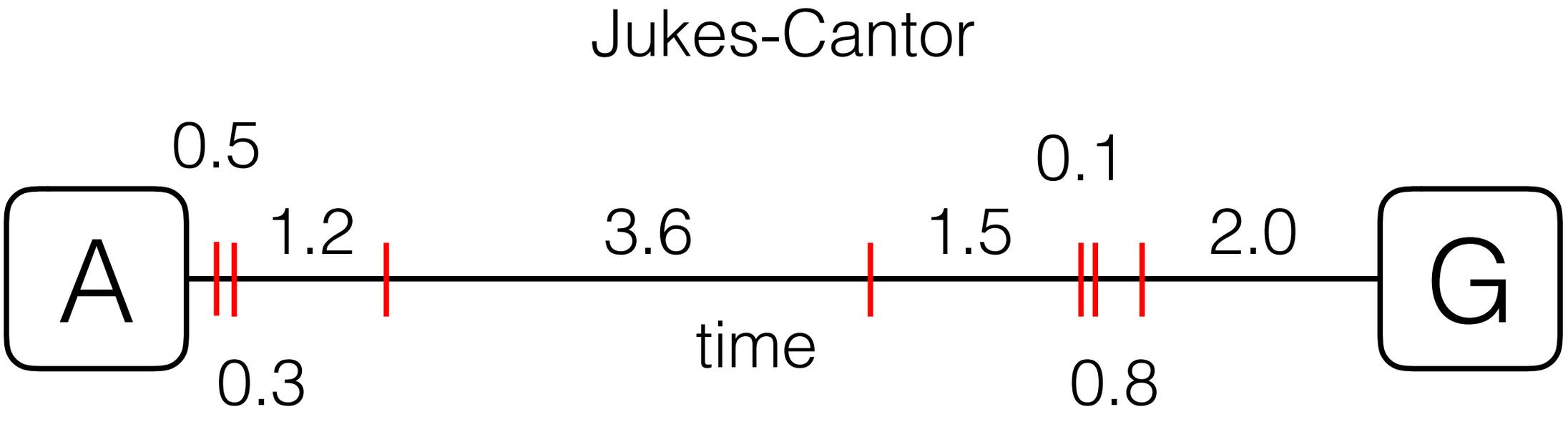
Integral of *Exp*(1) from 2.0 to infinity

Probability of a Character History



$$\text{Prob(history)} = \underbrace{Exp(0.5)Exp(0.3)Exp(1.2)Exp(3.6)Exp(1.5)Exp(0.1)Exp(0.8)}_{\text{Waiting Times}}\underbrace{e^{-2.0}}_{\substack{\text{Prob No} \\ \text{Change} \\ \text{at End}}}\underbrace{(1/3)^7}_{\substack{\text{Character} \\ \text{State} \\ \text{Changes}}}$$

Probability of Starting and Ending



Let’s say we simulated along our branch of length 10. We started in A and ended in G, with those specific events marked.

Probability of Starting and Ending

Jukes-Cantor

Let's say we simulated along our branch of length 10. We started in A and ended in G, with those specific events marked.

But now let's say we don't actually know the precise character history. How do we calculate

$$P(A \text{ to } G \mid t = 10)\,?$$

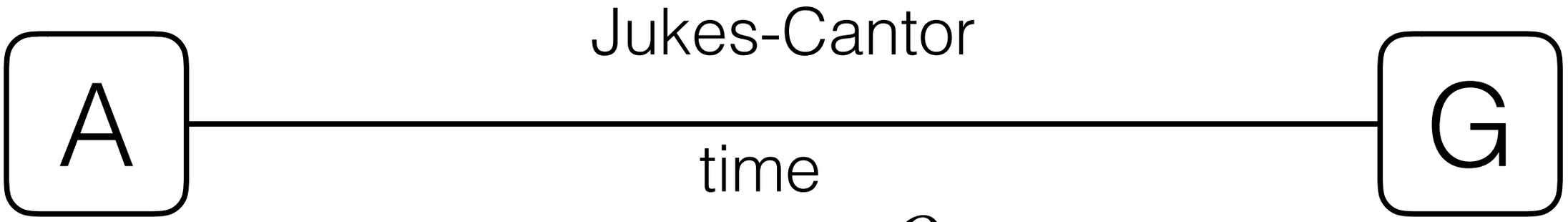
Probability of Starting and Ending

Jukes-Cantor

$$P(t) = e^{Qt}$$

This is the continuous-time equivalent of the matrix exponentiation we did with the discrete-time matrix!

Probability of Starting and Ending

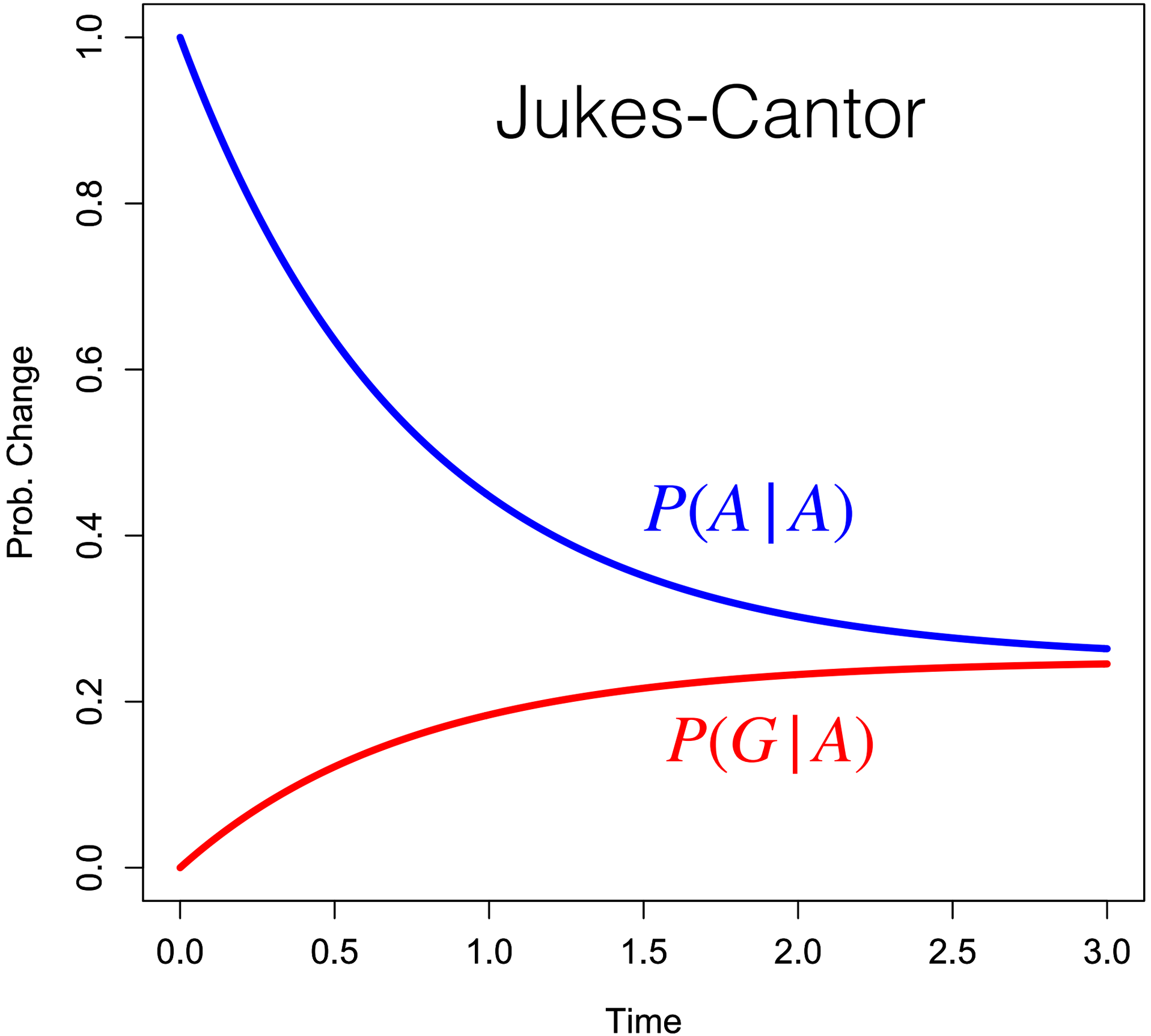


$$P(t) = e^{Qt}$$

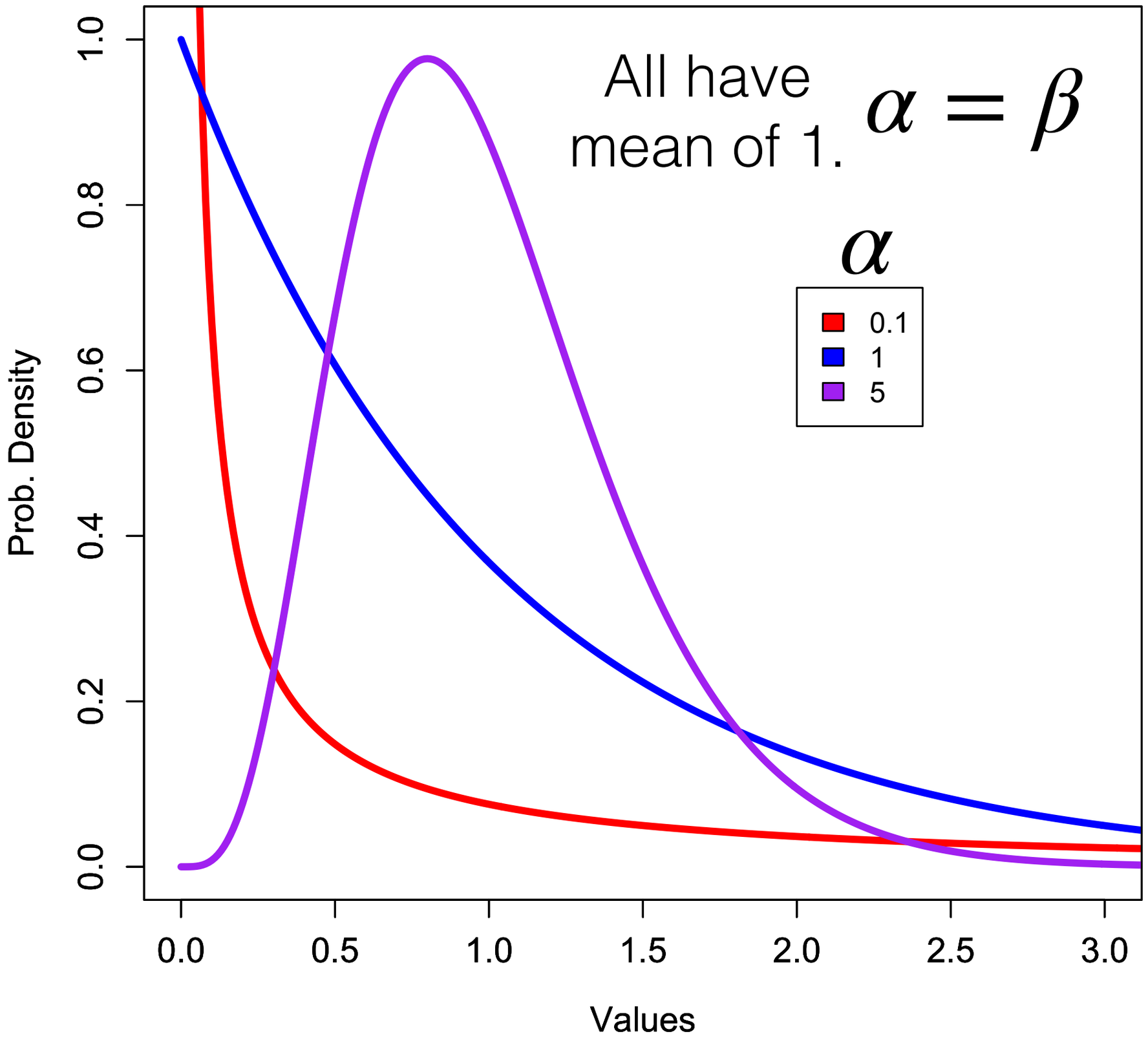
$$P(0) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad P(0.01) = \begin{bmatrix} 0.9901 & 0.0033 & 0.0033 & 0.0033 \\ 0.0033 & 0.9901 & 0.0033 & 0.0033 \\ 0.0033 & 0.0033 & 0.9901 & 0.0033 \\ 0.0033 & 0.0033 & 0.0033 & 0.9901 \end{bmatrix}$$

$$P(0.1) = \begin{bmatrix} 0.9064 & 0.0312 & 0.0312 & 0.0312 \\ 0.0312 & 0.9064 & 0.0312 & 0.0312 \\ 0.0312 & 0.0312 & 0.9064 & 0.0312 \\ 0.0312 & 0.0312 & 0.0312 & 0.9064 \end{bmatrix} \quad P(1) = \begin{bmatrix} 0.4477 & 0.1841 & 0.1841 & 0.1841 \\ 0.1841 & 0.4477 & 0.1841 & 0.1841 \\ 0.1841 & 0.1841 & 0.4477 & 0.1841 \\ 0.1841 & 0.1841 & 0.1841 & 0.4477 \end{bmatrix}$$

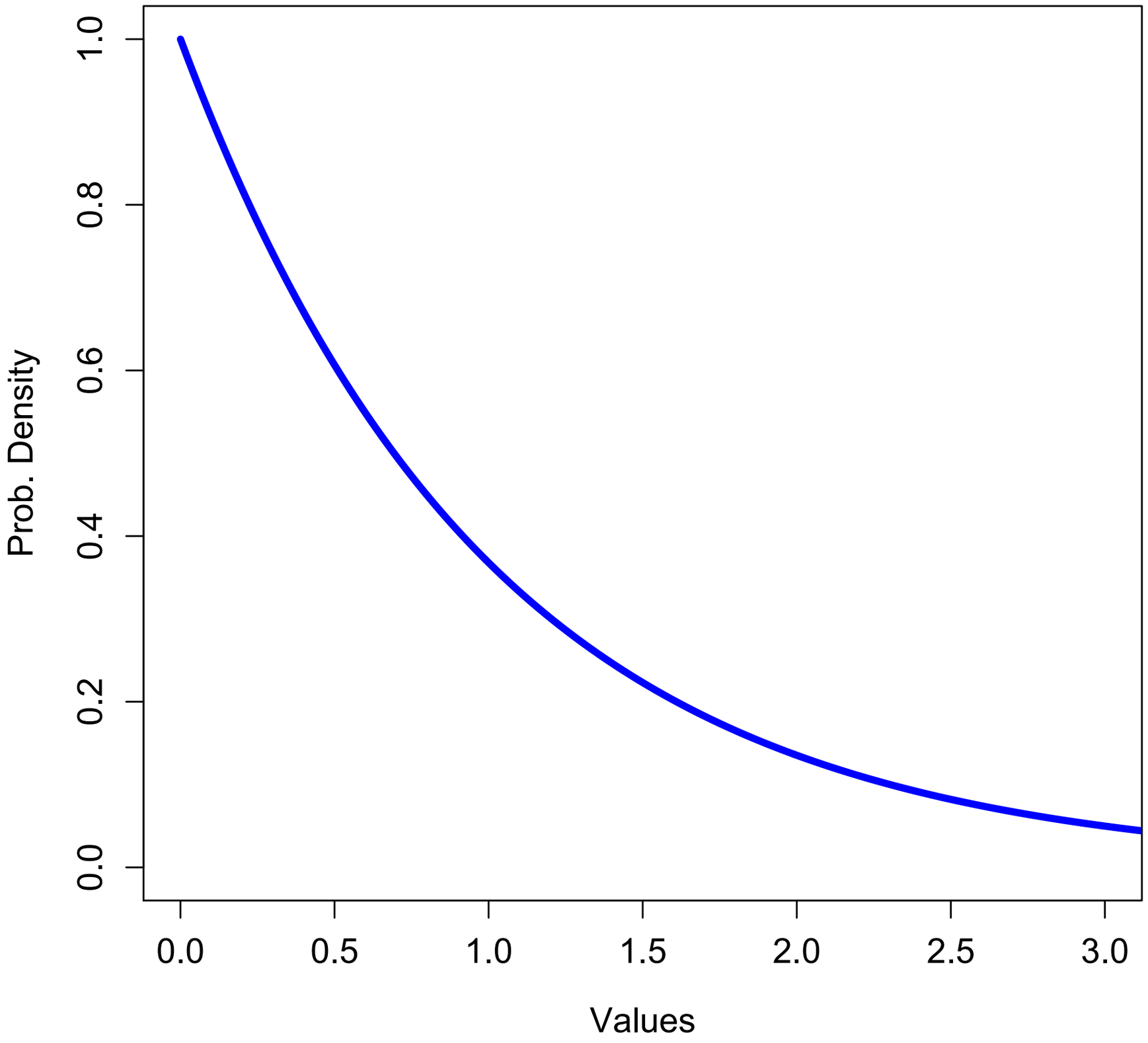
$$P(3) = \begin{bmatrix} 0.2637 & 0.2454 & 0.2454 & 0.2454 \\ 0.2454 & 0.2637 & 0.2454 & 0.2454 \\ 0.2454 & 0.2454 & 0.2637 & 0.2454 \\ 0.2454 & 0.2454 & 0.2454 & 0.2637 \end{bmatrix}$$


Jukes-Cantor
$P(A \mid A)$
$P(G \mid A)$
Prob. Change
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Time

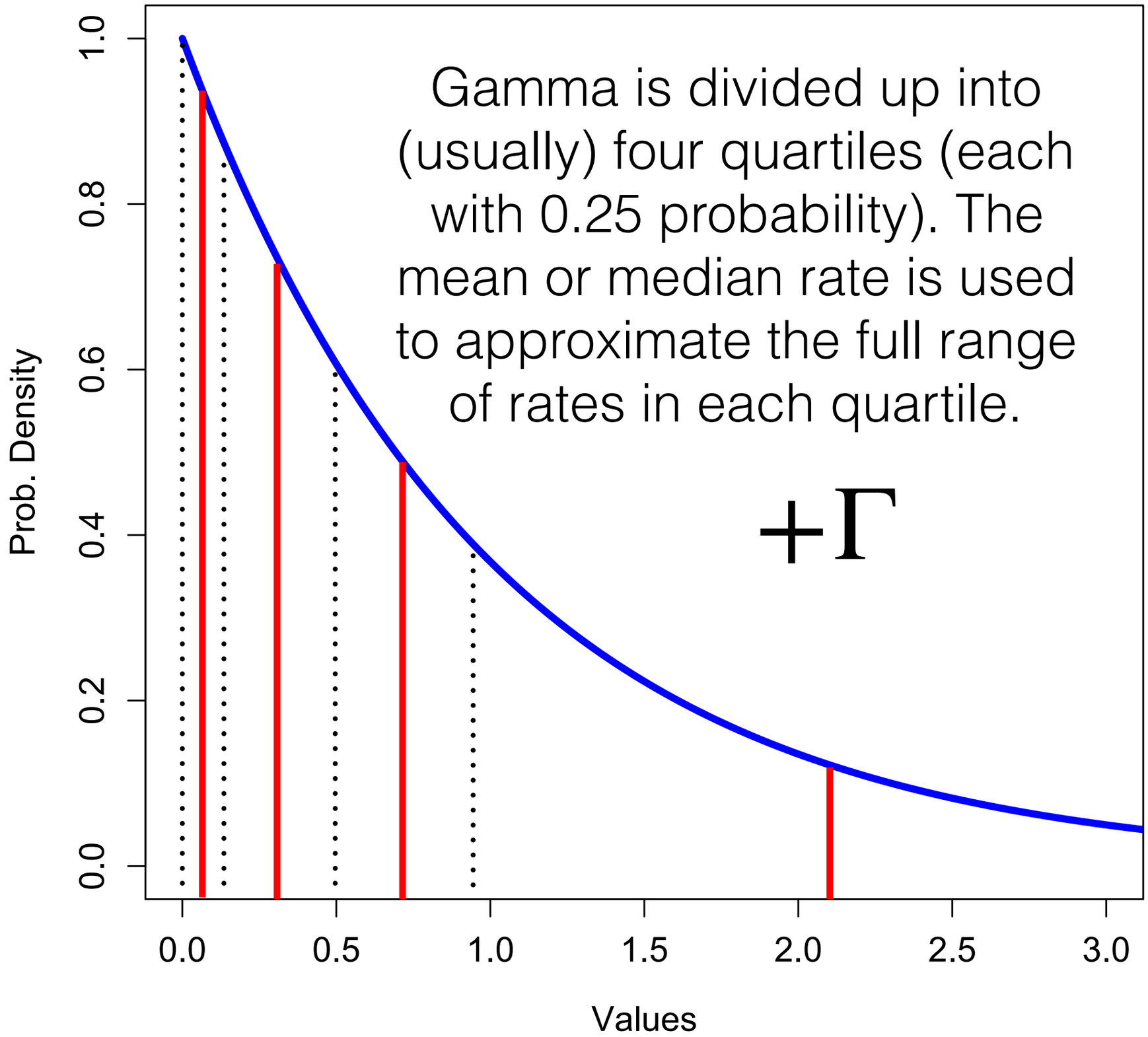
Rate Variation Across Sites (Gamma)



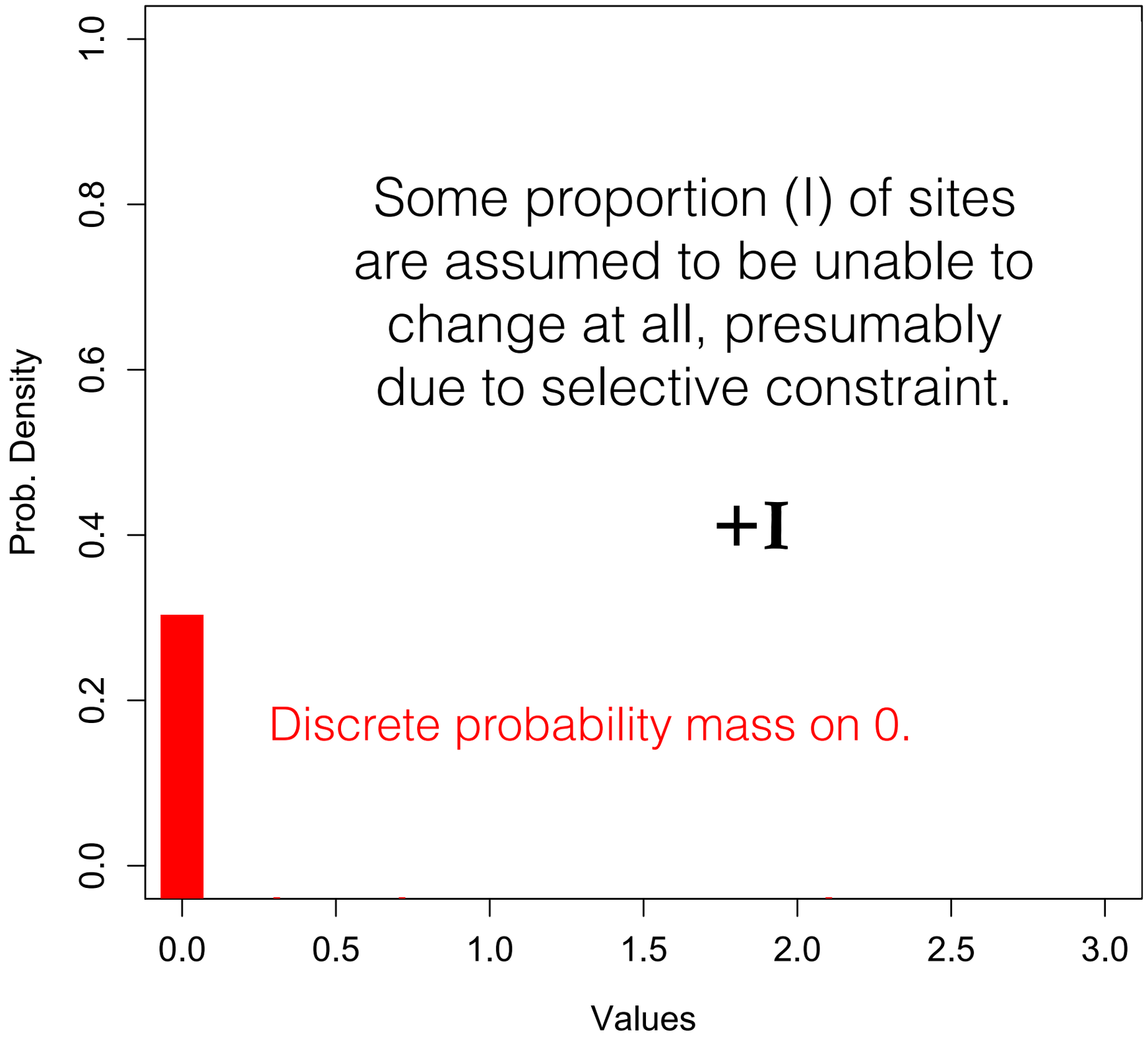
Rate Variation Across Sites (Gamma)



Rate Variation Across Sites (Discrete Gamma)

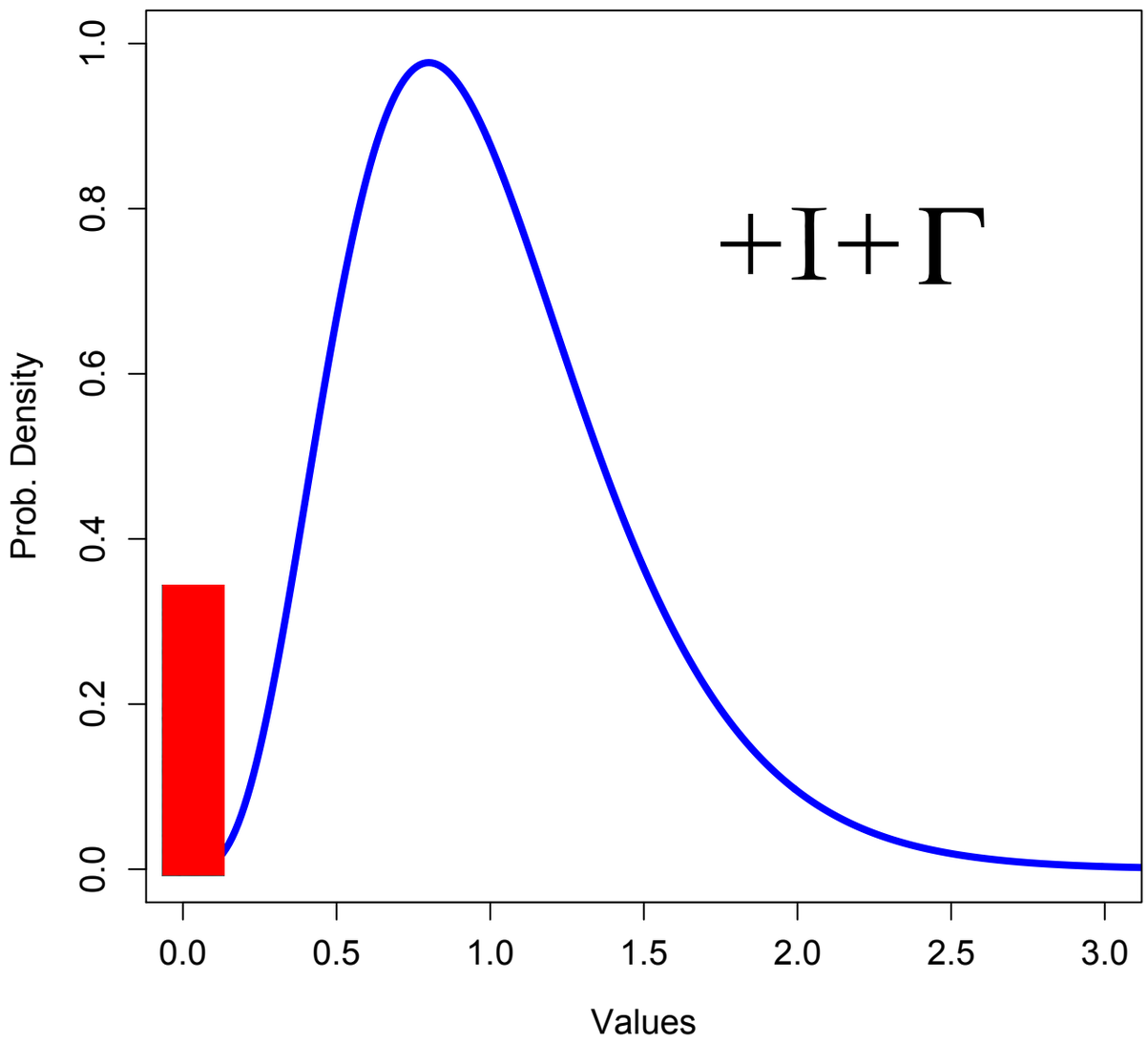


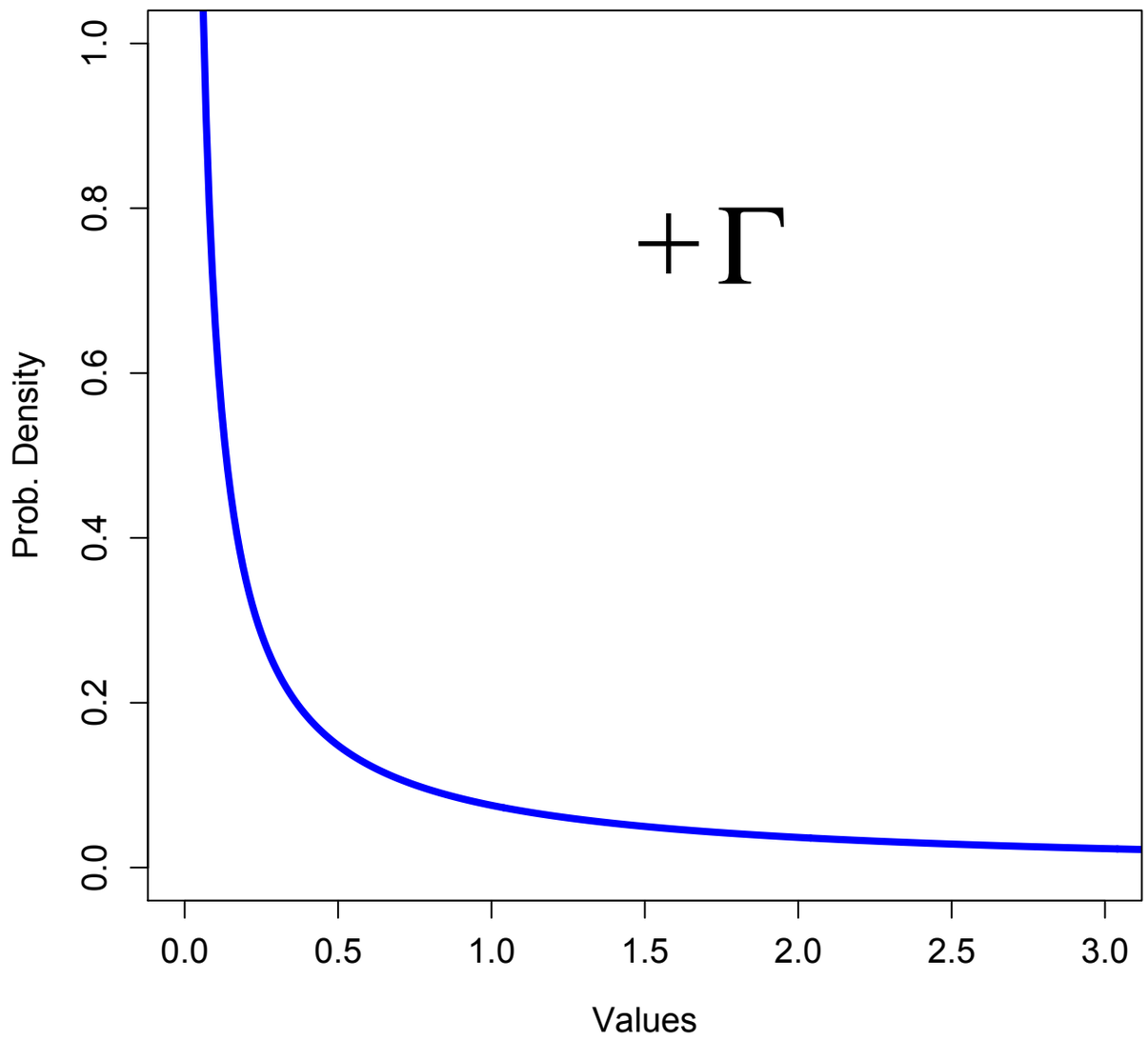
Rate Variation Across Sites (Proportion Invariable Sites)



Do these make **distinct** predictions about how rates of evolution vary across sites?

Challenge of **identifiability**.





Using models for: **Prediction** v Inference

If you know you have a fair coin (p=0.5), you can predict whether or not you should expect to observe different numbers of heads and tails.

$$P(k = 0 \,|\, n = 10, p = 0.5) = 0.001$$
$$P(k = 1 \,|\, n = 10, p = 0.5) = 0.010$$
$$P(k = 2 \,|\, n = 10, p = 0.5) = 0.044$$
$$P(k = 3 \,|\, n = 10, p = 0.5) = 0.117$$
$$P(k = 4 \,|\, n = 10, p = 0.5) = 0.205$$

...

Using models for: Prediction v **Inference**

But, what if instead of knowing (p=0.5), you know that k=3 and now you want to learn something about p?

Using models for: Prediction v **Inference**

But, what if instead of knowing (p=0.5), you know that k=3 and now you want to learn something about p?

Learning about some aspect of the process that is generating the data based on observed outcomes of that process is known as **inference**!

Even when we don't formalize it, we do this every day.

Using models for: Prediction v **Inference**

One way to think about this is to flip the conditionality in our previous probability statement:

$$P(p = 0.5 \,|\, n = 10, k = 3)$$

Using models for: Prediction v **Inference**

One way to think about this is to flip the conditionality in our previous probability statement:

$$P(p = 0.5 \,|\, n = 10, k = 3)$$

This is Bayesian inference!

However, because Bayesian inference associates probabilities with “belief” and requires statements of “prior belief”, it has made many people uncomfortable (R.A. Fisher included).

Maximum **Likelihood** Inference

To avoid the uncomfortable parts of Bayesian inference, Fisher invented *likelihood*.

Note that while *likelihood* and *probability* are used interchangeably in every day language, they mean distinct things in statistical inference.

Maximum **Likelihood** Inference

$$\mathscr{L}(p; n, k) = P(k \mid n, p)$$

Looking at these likelihoods, what is different than when we looked at the list of conditional probabilities a few slides ago?

$$\mathscr{L}(p = 0.1; n = 10, k = 3) = 0.057$$
$$\mathscr{L}(p = 0.2; n = 10, k = 3) = 0.201$$
$$\mathscr{L}(p = 0.3; n = 10, k = 3) = 0.267$$
$$\mathscr{L}(p = 0.4; n = 10, k = 3) = 0.215$$
$$\mathscr{L}(p = 0.5; n = 10, k = 3) = 0.117$$
$$\ldots$$

Maximum **Likelihood** Inference

$$\mathscr{L}(p; n, k) = P(k \mid n, p)$$

Looking at these likelihoods, what is different than when we looked at the list of conditional probabilities a few slides ago?

$$\mathscr{L}(p = 0.1; n = 10, k = 3) = 0.057$$
$$\mathscr{L}(p = 0.2; n = 10, k = 3) = 0.201$$
$$\mathscr{L}(p = 0.3; n = 10, k = 3) = 0.267$$
$$\mathscr{L}(p = 0.4; n = 10, k = 3) = 0.215$$
$$\mathscr{L}(p = 0.5; n = 10, k = 3) = 0.117$$
$$\ldots$$

Now, *k* stays the same and *p* changes.

Maximum **Likelihood** Inference

$$\mathscr{L}(p; n, k) = P(k \mid n, p)$$

Looking at these likelihoods, what is different than when we looked at the list of conditional probabilities a few slides ago?

$$\begin{aligned}
\mathscr{L}(p = 0.1; n = 10, k = 3) &= 0.057 \\
\mathscr{L}(p = 0.2; n = 10, k = 3) &= 0.201 \\
\mathscr{L}(p = 0.3; n = 10, k = 3) &= 0.267 \\
\mathscr{L}(p = 0.4; n = 10, k = 3) &= 0.215 \\
\mathscr{L}(p = 0.5; n = 10, k = 3) &= 0.117
\end{aligned}$$

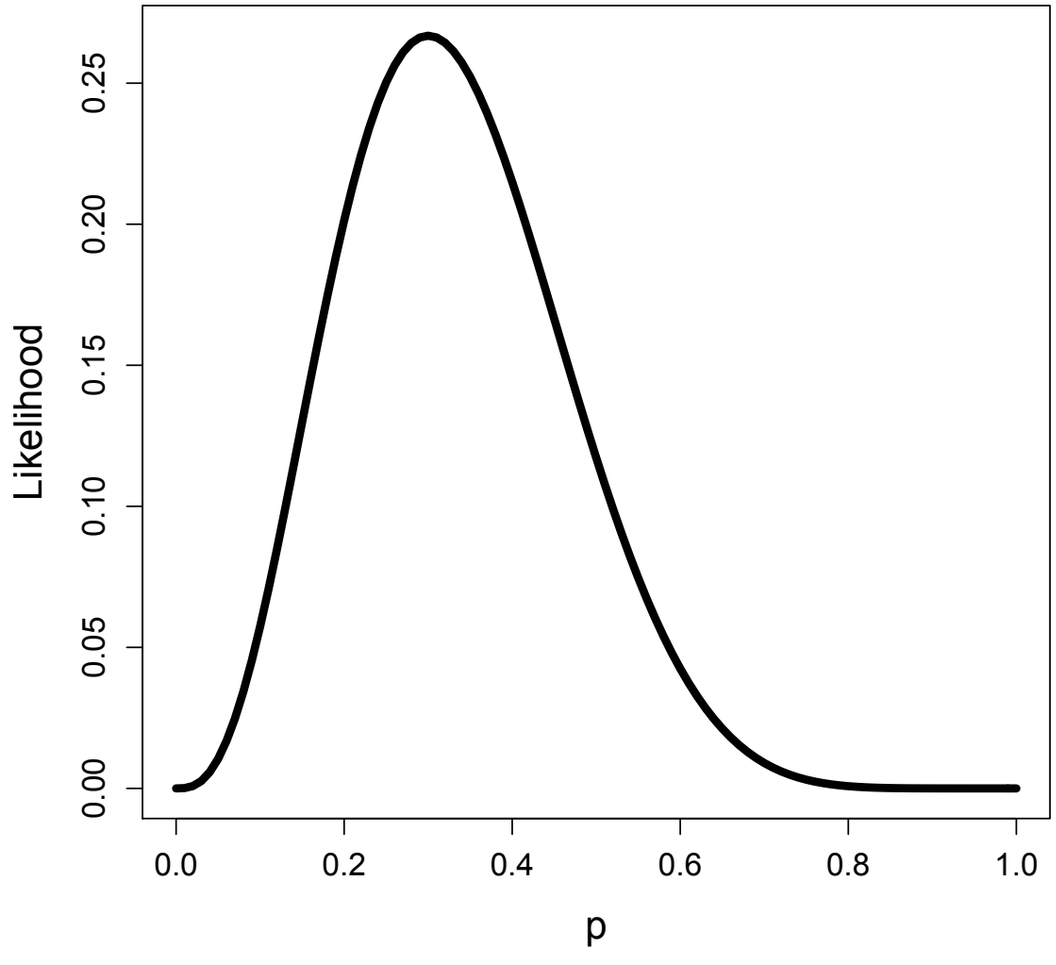
...

Also, likelihoods don't sum or integrate to 1.

Maximum **Likelihood** Inference

$$\mathscr{L}(p; n, k) = P(k \mid n, p)$$

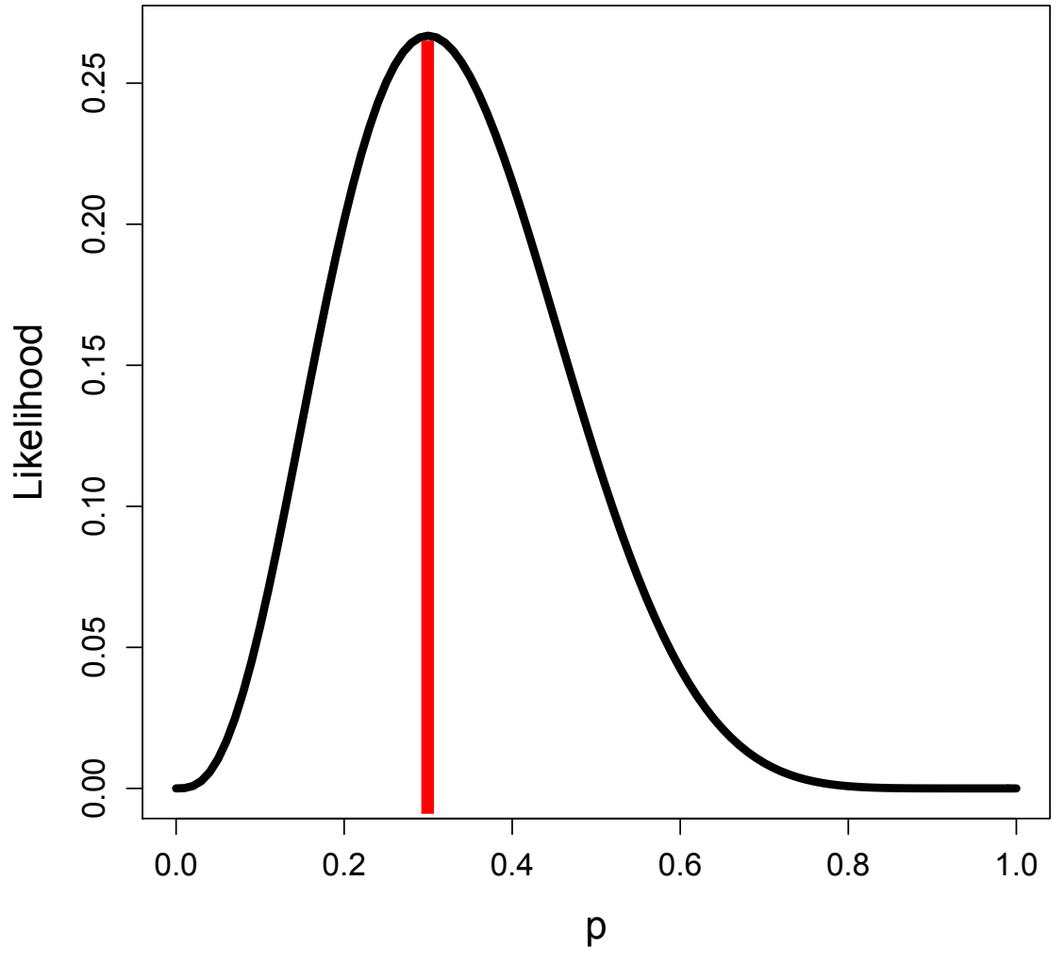
Fisher (and others) argued that likelihood is the most objective and principled way to evaluate evidence about parameters and models.



Maximum **Likelihood** Inference

$$\mathscr{L}(p; n, k) = P(k \mid n, p)$$

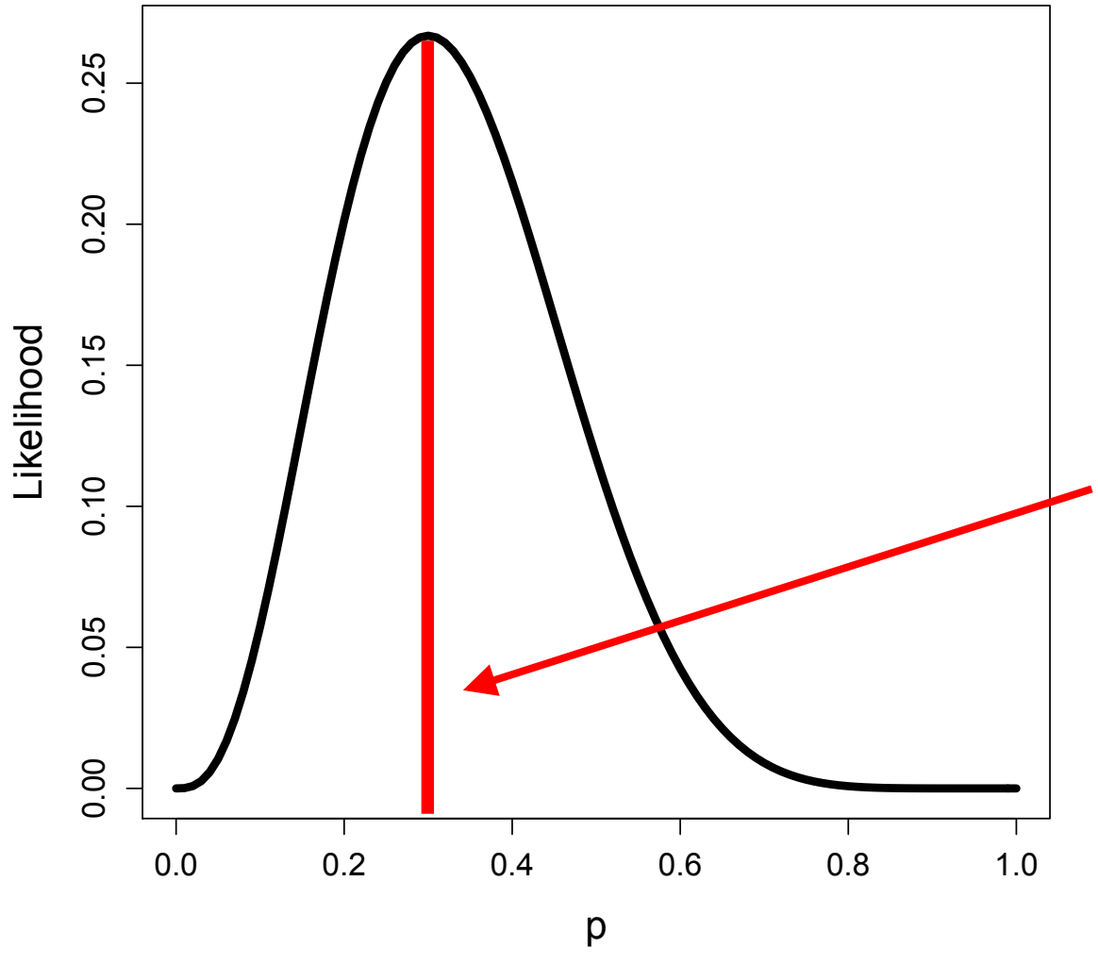
The value of a parameter that maximizes the likelihood is known as the **maximum-likelihood estimate (MLE)** of the "true" parameter value.



Maximum **Likelihood** Inference

$$\mathscr{L}(p; n, k) = P(k \mid n, p)$$

The value of a parameter that maximizes the likelihood is known as the **maximum-likelihood estimate (MLE)** of the "true" parameter value.

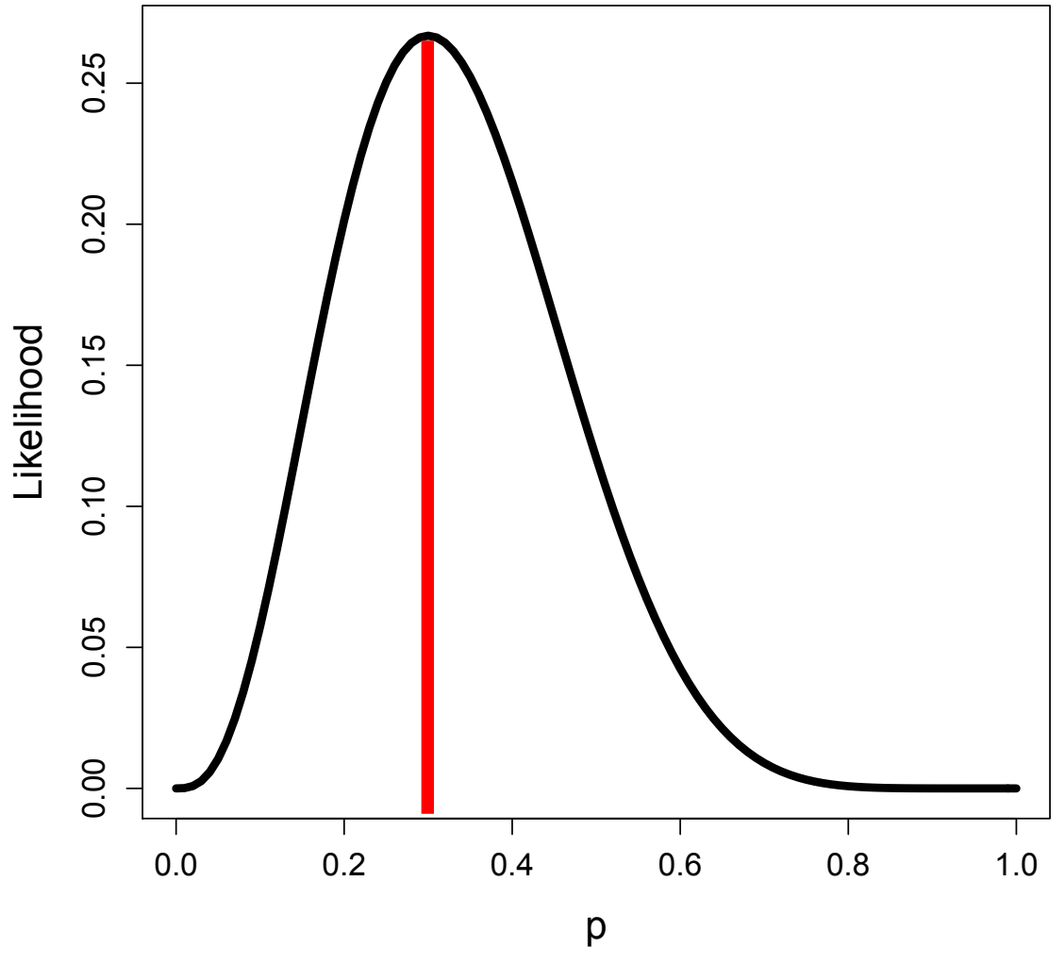


Can you guess the MLE for 3 heads in 10 coin flips?

Hill Climbing

$$\mathscr{L}(p; n, k) = P(k \mid n, p)$$

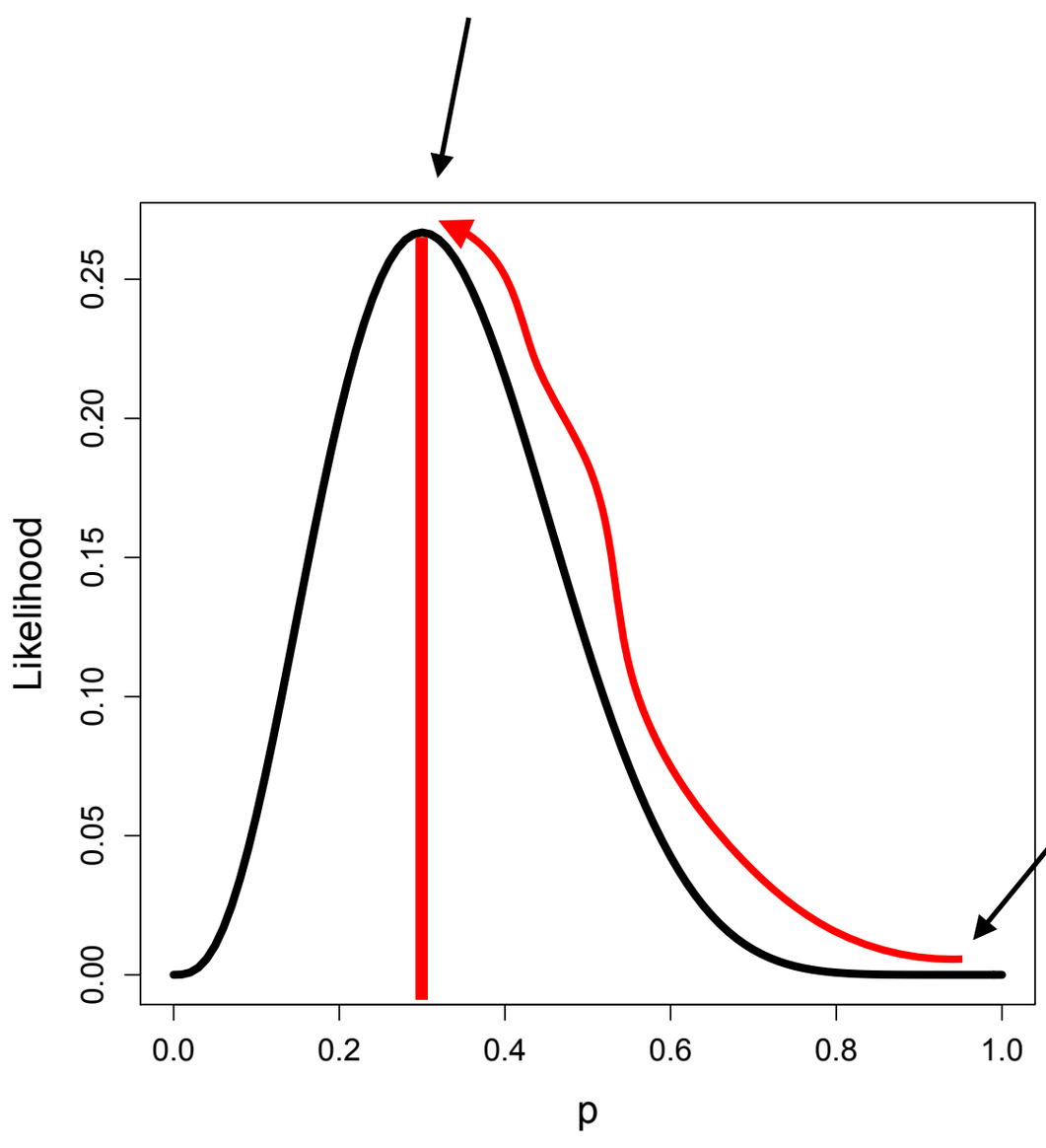
If we have a likelihood function, we will generally use some kind of hill climbing algorithm (there are lots!) to find the MLE.



Hill Climbing

$$\mathscr{L}(p; n, k) = P(k \mid n, p)$$

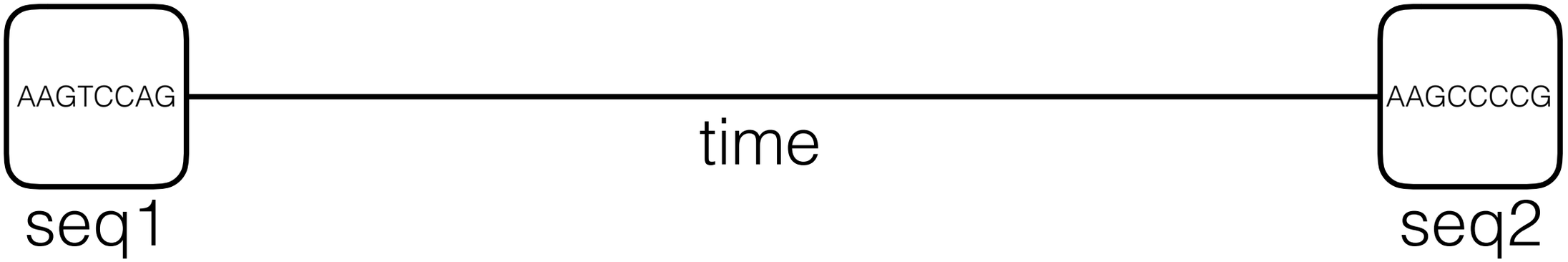
Estimate of the MLE



Initial guess or starting *p*.

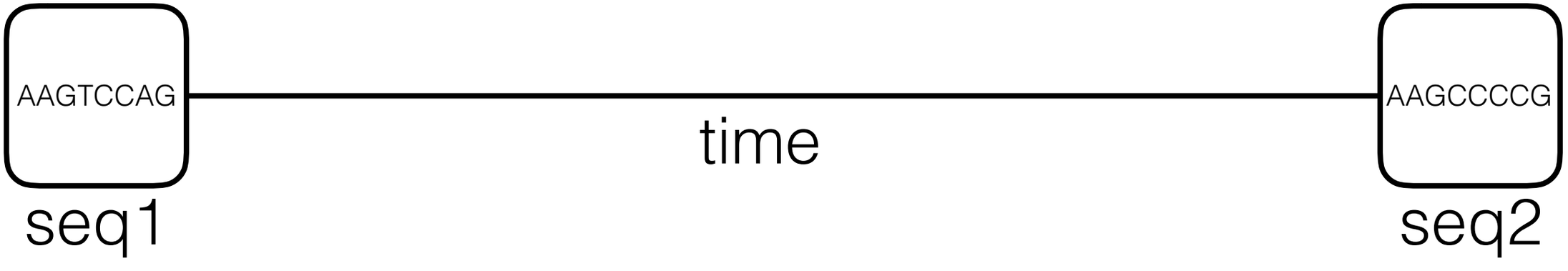
Maximum-likehood genetic distance estimation

$$\mathscr{L}(t; seq1, seq2) = P(seq1, seq2 \mid t)$$



Maximum-likehood genetic distance estimation

$$\mathscr{L}(t; seq1, seq2) = P(seq1, seq2 \mid t)$$



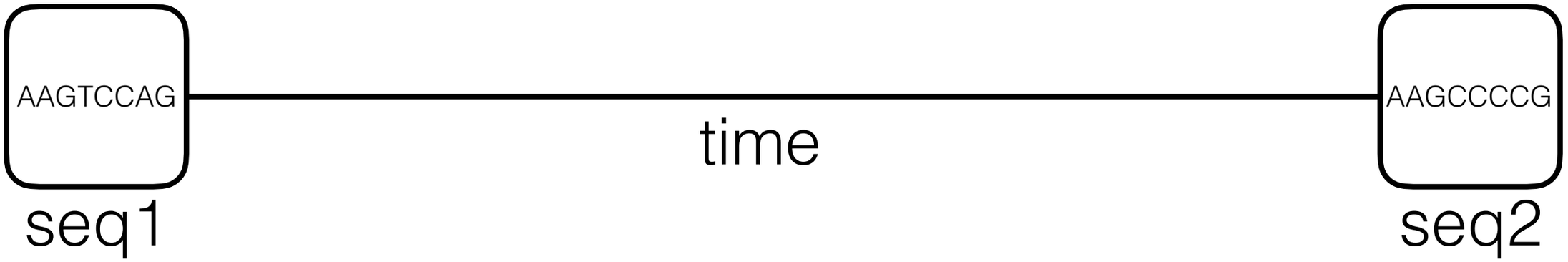
$$P(0) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

What is $\mathscr{L}(0; AAGTCCAG, AAGCCCCG)$?

Remember, sites are independent of one another!

Maximum-likehood genetic distance estimation

$$\mathscr{L}(t; seq1, seq2) = P(seq1, seq2 \mid t)$$

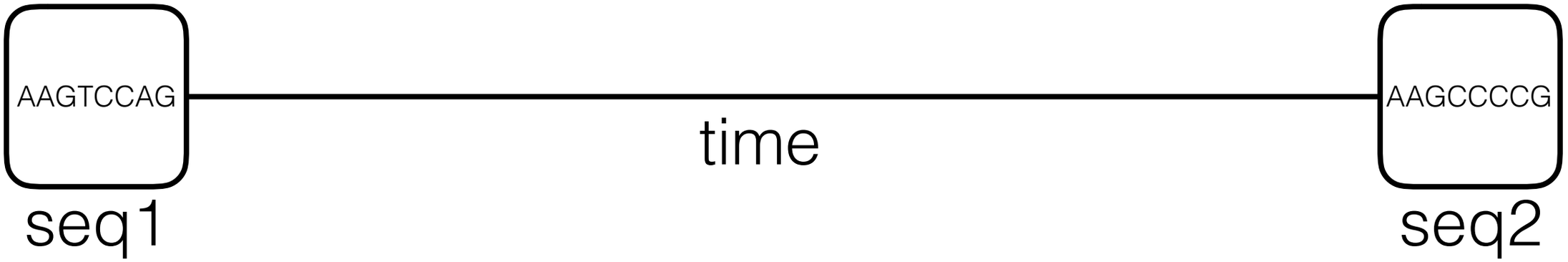


$$P(0.1) = \begin{bmatrix} 0.9064 & 0.0312 & 0.0312 & 0.0312 \\ 0.0312 & 0.9064 & 0.0312 & 0.0312 \\ 0.0312 & 0.0312 & 0.9064 & 0.0312 \\ 0.0312 & 0.0312 & 0.0312 & 0.9064 \end{bmatrix}$$

What is $\mathscr{L}(0.1; AAGTCCAG, AAGCCCCG)$?

Maximum-likehood genetic distance estimation

$$\mathscr{L}(t; seq1, seq2) = P(seq1, seq2 \mid t)$$



Based on visual inspection of these sequences, what do you think is the MLE for genetic distance?