Bayesian Phylogenetics

Český Krumlov Phylogenomics Workshop 2019

Mario dos Reis

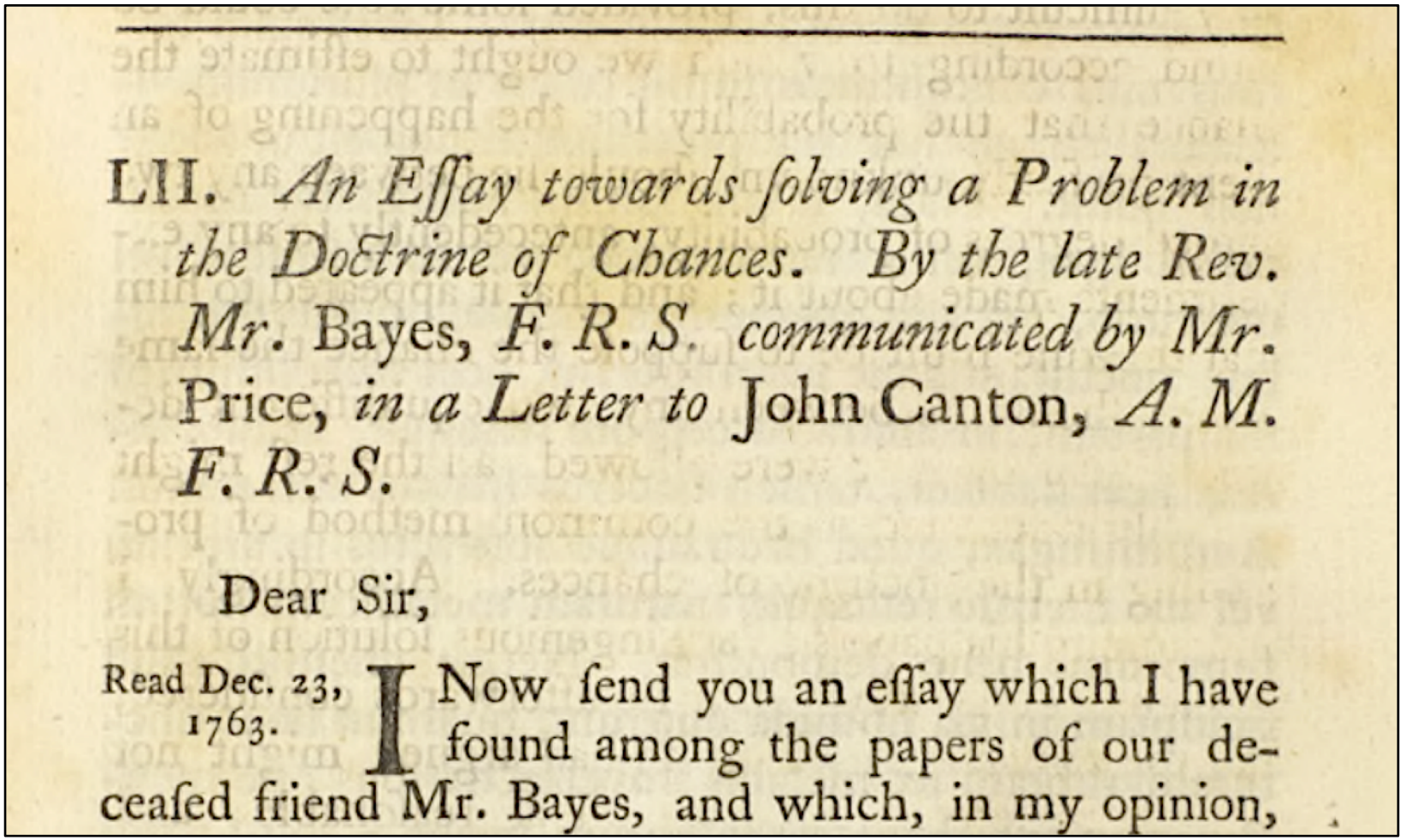
@mariodosreis

Queen Mary University of London

Bayesian Phylogenetics

- Brief review of probability
- Bayes theorem
- Bayesian inference – Rev Bayes experiment
- Introduction to MCMC
- Understanding MCMC output
  - Summarising the posterior distribution
  - Convergence diagnostics
- Bayesian asymptotics (approximating the likelihood for large data)

Rev Thomas Bayes Paper



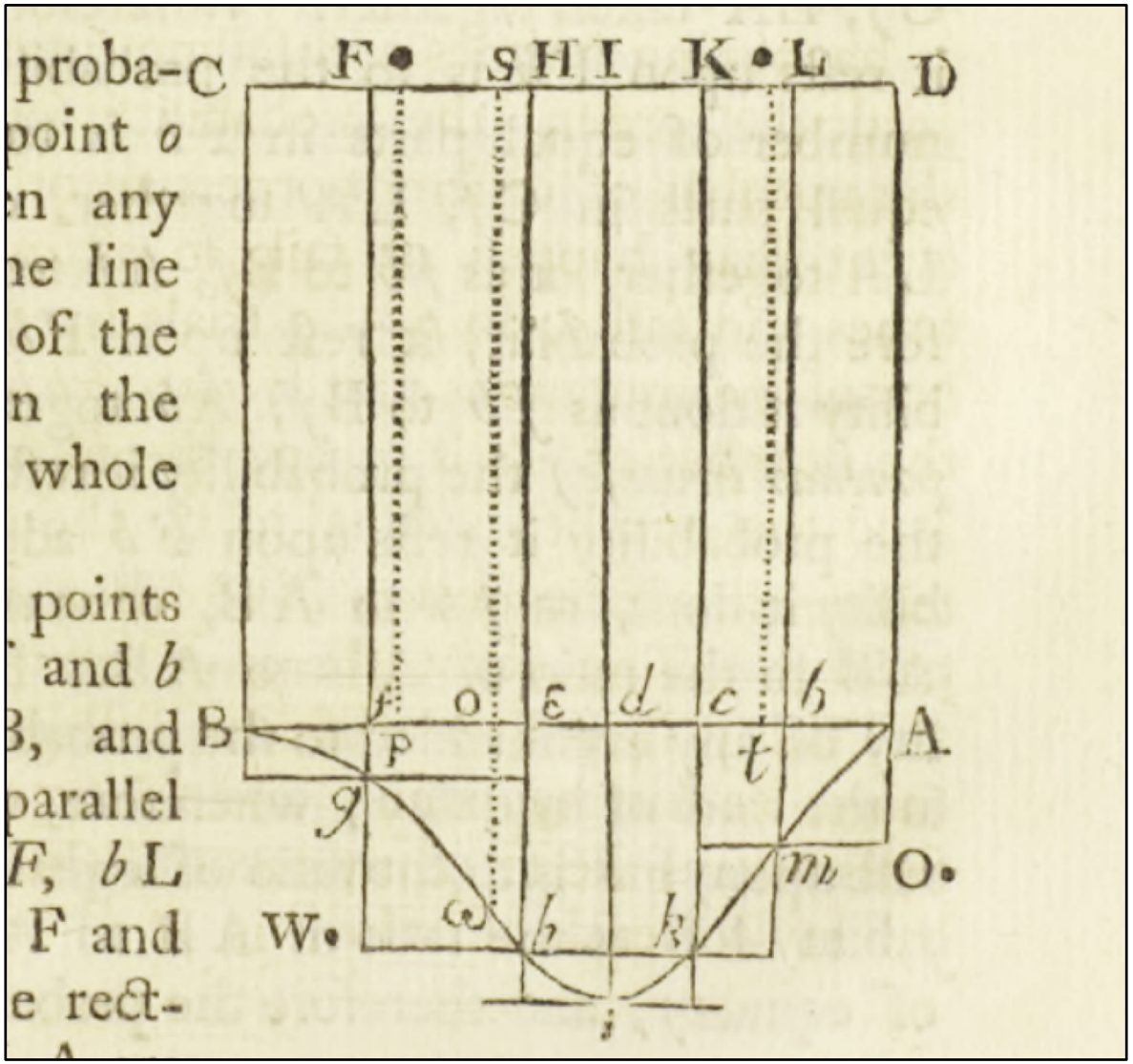
LII. *An Eſſay towards ſolving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

Read Dec. 23, 1763. I Now ſend you an eſſay which I have found among the papers of our deceaſed friend Mr. Bayes, and which, in my opinion,

Philosophical Transactions of the Royal Society of London, (1763) 53: 370–418. doi:10.1098/rstl.1763.0053.

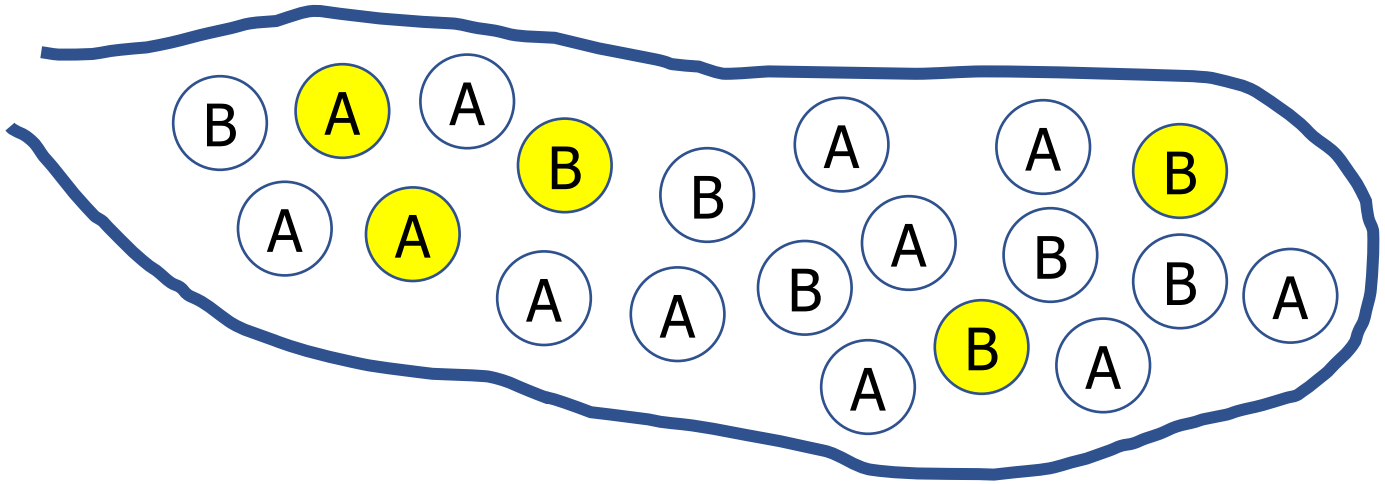
Rev Thomas Bayes Paper



Golf Balls in a Bag

- Suppose you place twenty golf balls in a dark bag:

	Brand A	Brand B	
Yellow	2	3	
White	10	5	
			20



Golf Balls in a Bag

- Suppose you place twenty golf balls in a dark bag:

	Brand A	**Brand B**	
Yellow	2	3	
White	10	5	
			20

- You take one ball randomly out of the bag
- What is the probability that it is yellow and made by A?
- P(Y, A) = ?

Golf Balls in a Bag

- Suppose you place twenty golf balls in a dark bag:

	Brand A	**Brand B**	
Yellow	2	3	
White	10	5	
			20

- P(Y, A) = 2 / 20 = 0.1 or 10%
- P(Y, A) is known as the **joint probability** of Y and A

Golf Balls in a Bag

- Suppose you place twenty golf balls in a dark bag:

	Brand A	**Brand B**	
Yellow	2	3	
White	10	5	
			20

- You place the ball back in the bag, mix and take out another ball
- What is the probability that it is white?
- P(W) = ?

Golf Balls in a Bag

- Suppose you place twenty golf balls in a dark bag:

Table margin!

	Brand A	Brand B	
Yellow	2	3	5
White	10	5	15
Table margin! →	12	8	20

- P(W) = 15 / 20 = 0.75
- P(W) is known as the **marginal probability** of W

Golf Balls in a Bag

- Suppose you place twenty golf balls in a dark bag:

	Brand A	**Brand B**	
Yellow	2	3	5
White	10	5	15
	12	8	20

- But note that:
- P(W) = 10 / 20 + 5 / 20 = 0.75 or
- P(W) = P(W, A) + P(W, B)
- The **marginal** is the sum over the **joints**!

Golf Balls in a Bag

- Suppose you place twenty golf balls in a dark bag:

	Brand A	**Brand B**	
Yellow	2	3	5
White	10	5	15
	12	8	20

- Assume you took out a white ball, what is the probability that it was made by A?
- P(A | W) = ?

Golf Balls in a Bag

- Suppose you place twenty golf balls in a dark bag:

	Brand A	**Brand B**	
Yellow	2	3	5
White	10	5	15
	12	8	20

- P(A | W) = 10 / 15 = 0.666…
- P(A | W) is the **conditional probability** of A given W

Golf Balls in a Bag

- Suppose you place twenty golf balls in a dark bag:

	Brand A	**Brand B**	
Yellow	2	3	5
White	10	5	15
	12	8	20

- But note that:
- P(A | W) = (10 / 20) / (15 / 20) = 0.666 … or
- P(A | W) = P(W, A) / P(W)
- The **conditional** is the **joint** over the **marginal**

Golf Balls in a Bag

- Suppose you place twenty golf balls in a dark bag:

	Brand A	**Brand B**	
Yellow	2	3	5
White	10	5	15
	12	8	20

- Note we can reverse the conditional:
- P(W | A) = P(W, A) / P(A)
- P(W | A) = (10 / 20) / (12 / 20) = 0.833…

Bayes Theorem

- $P(W \mid A) = P(W, A) / P(A)$
- $P(A \mid W) = P(W, A) / P(W)$
- This means that:
- $P(W, A) = P(A) \times P(W \mid A)$
- $P(W, A) = P(W) \times P(A \mid W)$
- Thus

$$P(W \mid A) = P(W) \times P(A \mid W) / P(A)$$

- This is **Bayes theorem**!

Marginal Probability

- $P(A) = P(W, A) + P(Y, A)$
- $P(A) = P(A \mid W)P(W) + P(A \mid Y)P(Y)$
- Suppose there are balls of *n* different colours in the bag, then
- $P(A) = P(A \mid C_1)P(C_1) + \ldots + P(A \mid C_n)$
- $P(A) = \Sigma_i P(A \mid C_i)P(C_i)$

Rev Bayes Thought Experiment (modified)

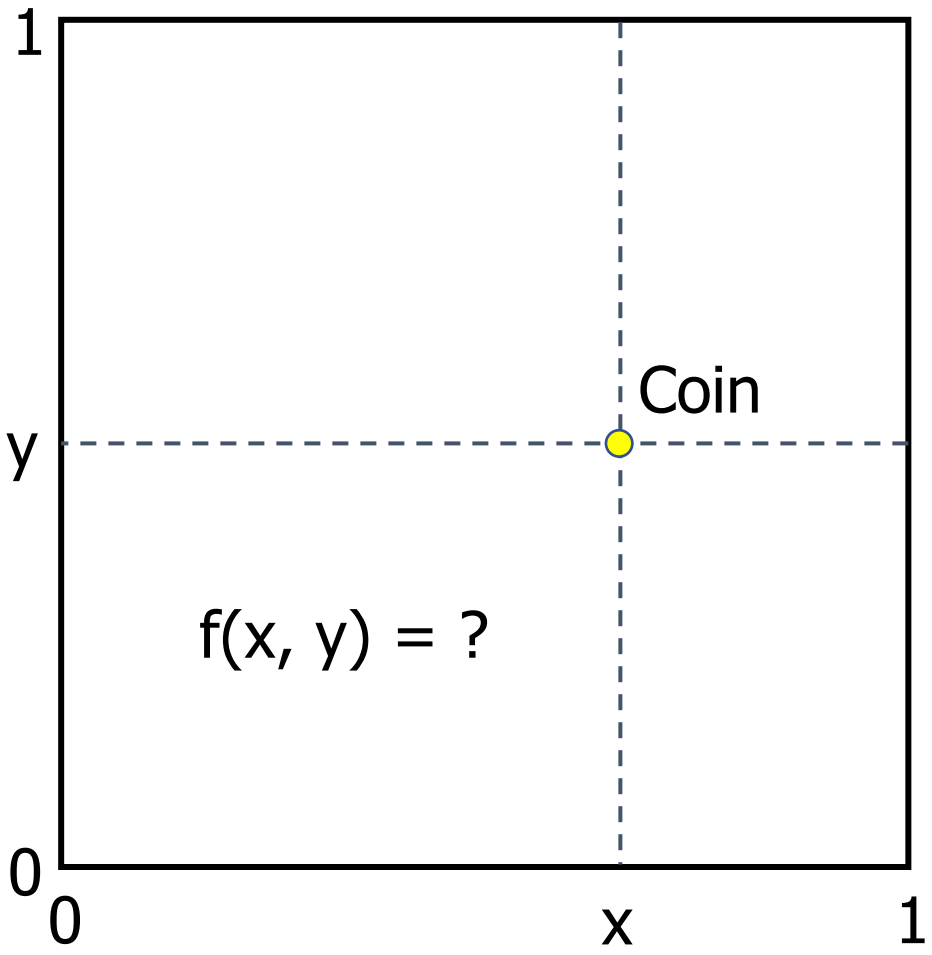
- We have a perfectly even, horizontal table with raised edges
- A ball is thrown onto the table and its resting position marked with a coin
- We are never allowed to see the ball or coin
- The ball is thrown again and we are told if:
  - The ball landed left or right of the coin
  - The ball landed in front or behind the coin
- After *n* throws, can we guess the position of the coin?

Rev Bayes Thought Experiment (modified)

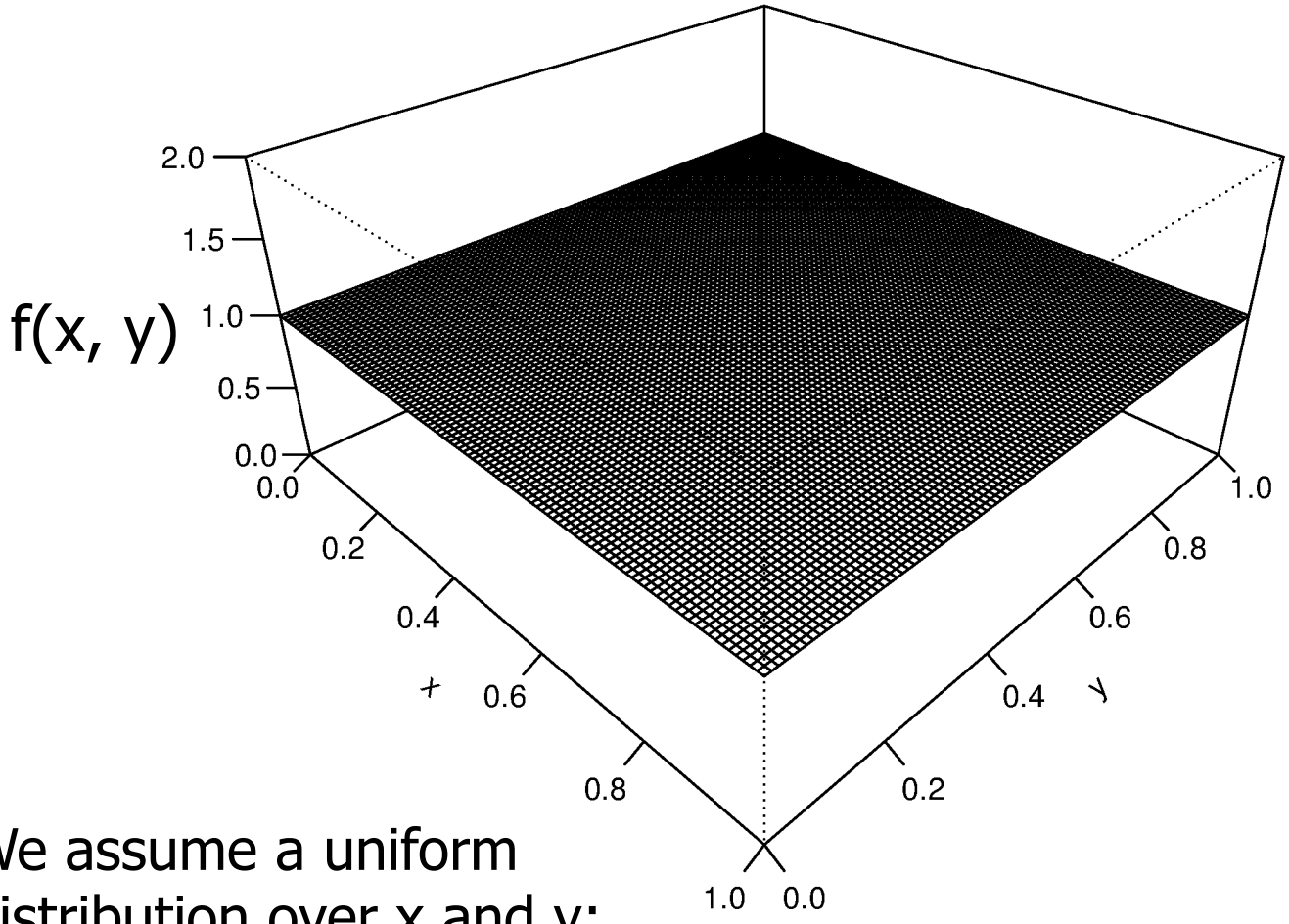
- Revered Bayes showed how to calculate a reasonable guess
- Not only that, he showed that with sufficient throws, we would eventually become **almost certain** of the coin's position!
- We will learn his method

Rev Bayes Thought Experiment (modified)

The coin marks the position of the ball for the initial throw

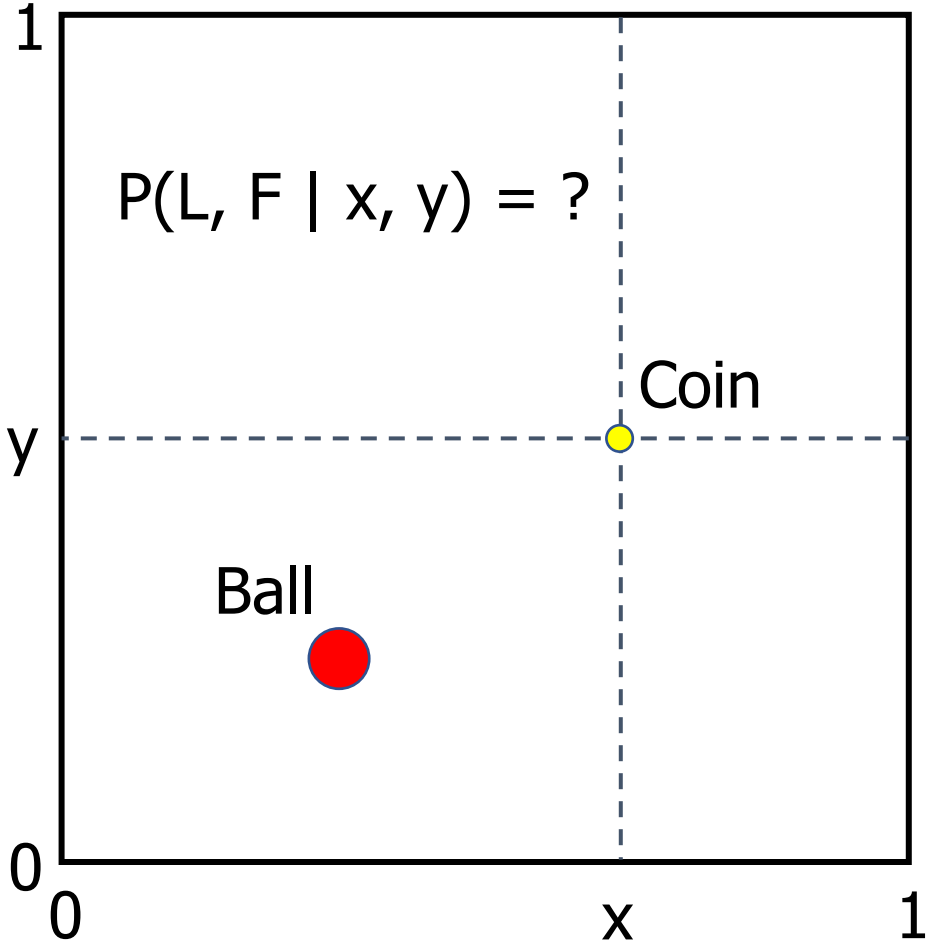


Rev Bayes Thought Experiment (modified)



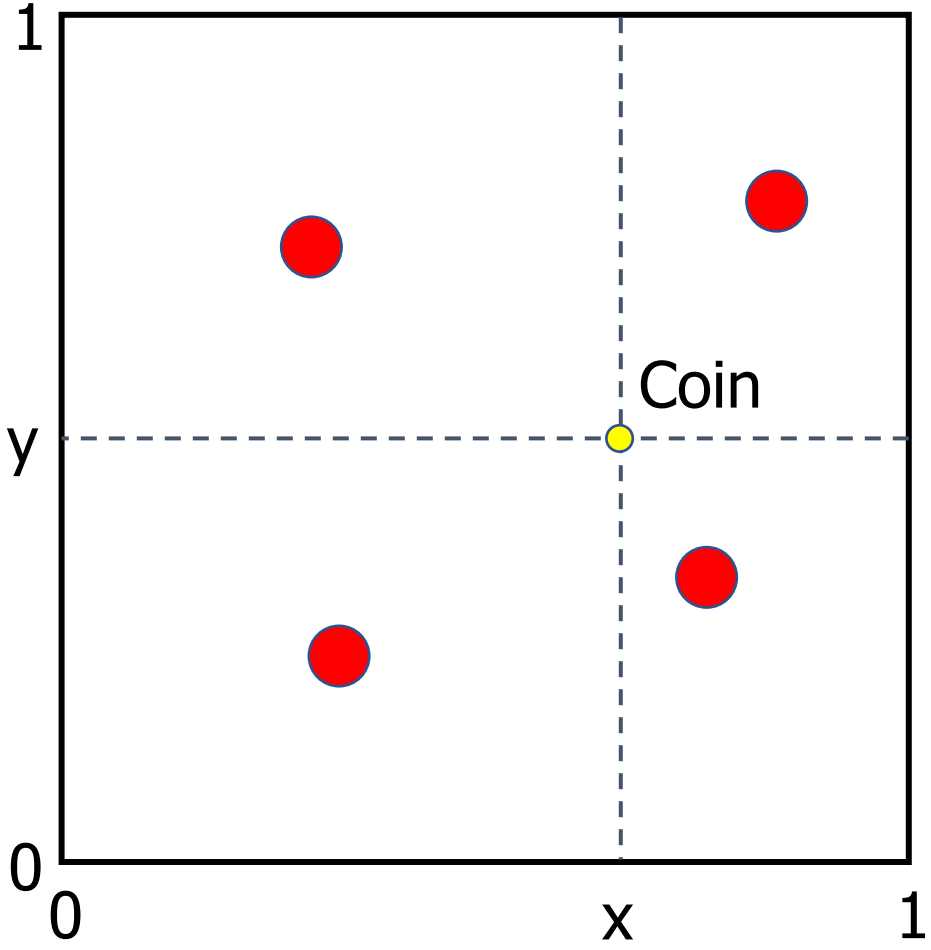
We assume a uniform distribution over x and y:
f(x, y) = 1

Rev Bayes Thought Experiment (modified)



- L: left
- R: right
- F: front
- B: back

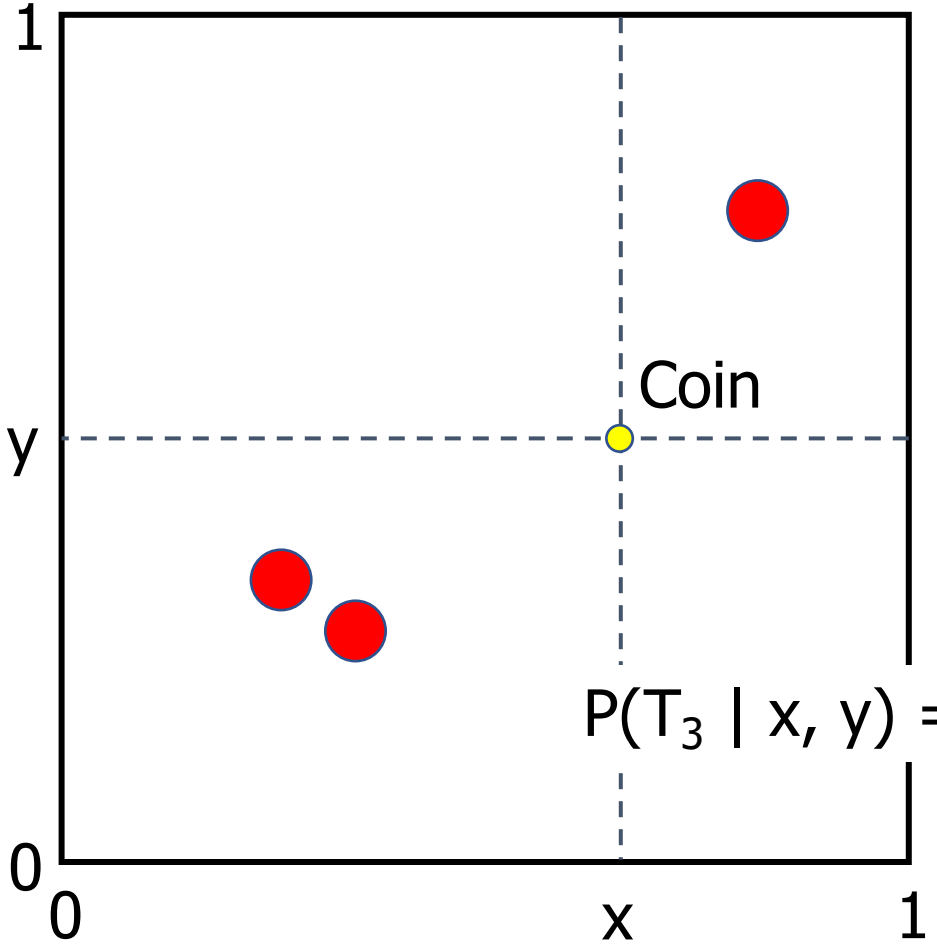
Rev Bayes Thought Experiment (modified)



The probability, after **one throw**, is the landing area:

- P(L, F | x, y) = xy
- P(L, B | x, y) = x (1 – y)
- P(R, F | x, y) = (1 – x) y
- P(R, B | x, y) = (1–x)(1–y)

Rev Bayes Thought Experiment (modified)



The probability, of a **sequence of throws**, is the product of the single throw probabilities:

- $T_3 = (\{L,F\}, \{L,F\}, \{R,B\})$

$$P(T_3 \mid x, y) = P(L, F \mid x, y)^2 \, P(R, B \mid x, y)$$

Rev Bayes Thought Experiment (modified)

The probability, of a **sequence of throws**, are the product of the single throw probabilities:

- $T_3 = (\{L, F\}, \{L, F\}, \{R, B\})$
- $P(T_3 \mid x, y) = P(L, F \mid x, y)^2 \, P(R, B \mid x, y)$
- $P(T_3 \mid x, y) = (xy)^2(1 - x)(1 - y)$

In general, the probability after **n throws** is:

- $P(T_n \mid x, y) = x^a \, (1 - x)^{(n - a)} \, y^b \, (1 - y)^{(n - b)}$
- a and b: number of left and front landings
- n: total number of throws

Rev Bayes Thought Experiment (modified)

We have defined the marginal density of x and y, and calculated the conditional probability of T_n given x, y:

- $f(x, y) = 1$
- $P(T_n \mid x, y) = x^a (1 - x)^{(n - a)} y^b (1 - y)^{(n - b)}$

Thus, we now have **the joint density** of T_n, x, y:

- $f(T_n, x, y) = f(x, y) P(T_n \mid x, y)$

Rev Bayes Thought Experiment (modified)

Thus, according to the **Bayes theorem**:

- $f(x, y \mid T_n) = f(x, y)\, P(T_n \mid x, y) / P(T_n)$

Our problem is calculating the marginal probability $P(T_n)$

Recall that the marginal probability is the sum over the joint probabilities. Here, x and y are continuous, so instead of a double sum, we have a double integral:

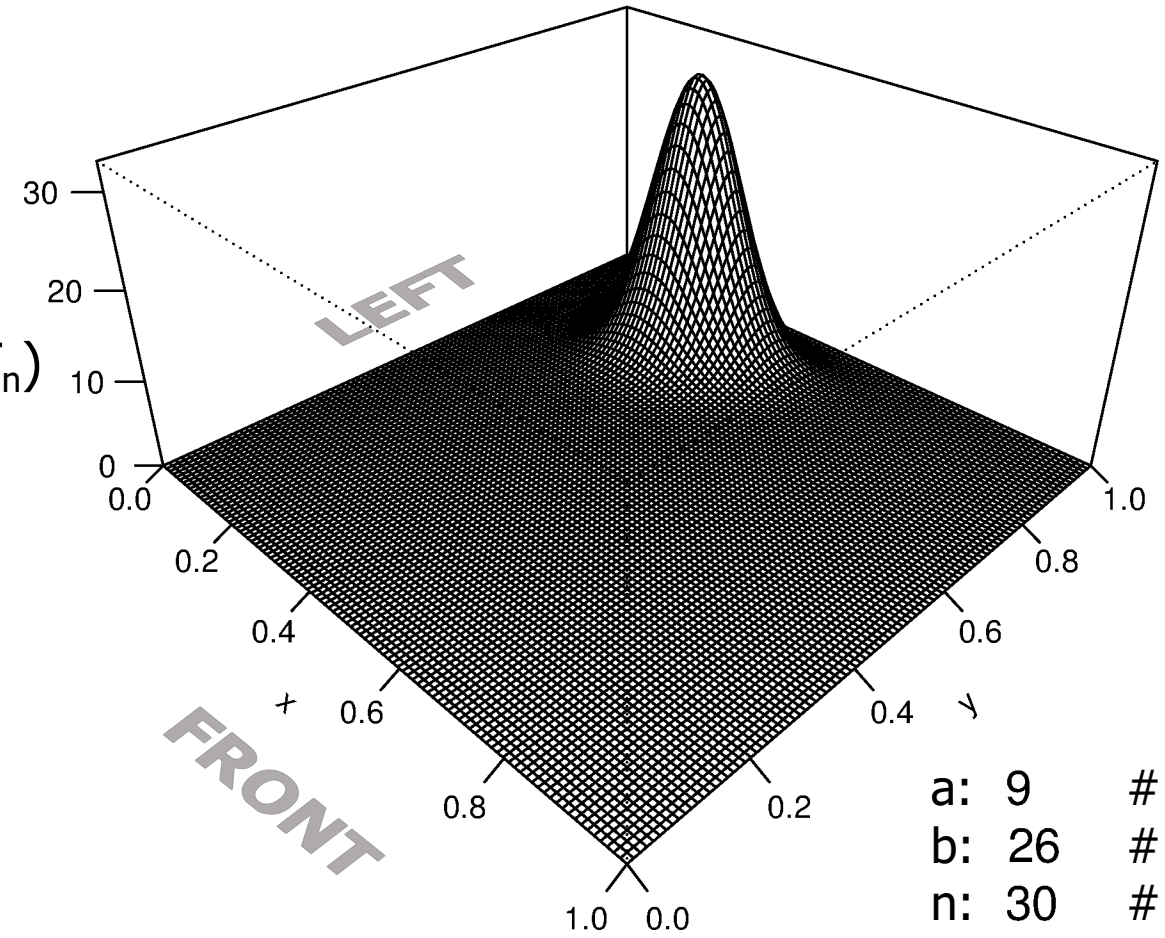
- $P(T_n) = \iint f(T_n, x, y)\, dx\, dy$
- $P(T_n) = [a!\,(n - a)!\,b!\,(n - b)!] / [(n + 1)!]^2$

Rev Bayes Thought Experiment (simulation)

Computer simulation of modified Bayes experiment:

1. Sample x and y from the joint uniform f(x, y). This is the position of the coin
2. Set a = b = n = 0
3. Sample two numbers, w and z, from the joint uniform. This is the position of the ball after one throw
4. Set a = a + 1 if w < x (ball is at left)
5. Set b = b + 1 if z < y (ball is at front)
6. Repeat steps 3-5 n times
7. Calculate $f(x, y \mid T_n)$

Rev Bayes Thought Experiment (simulation)

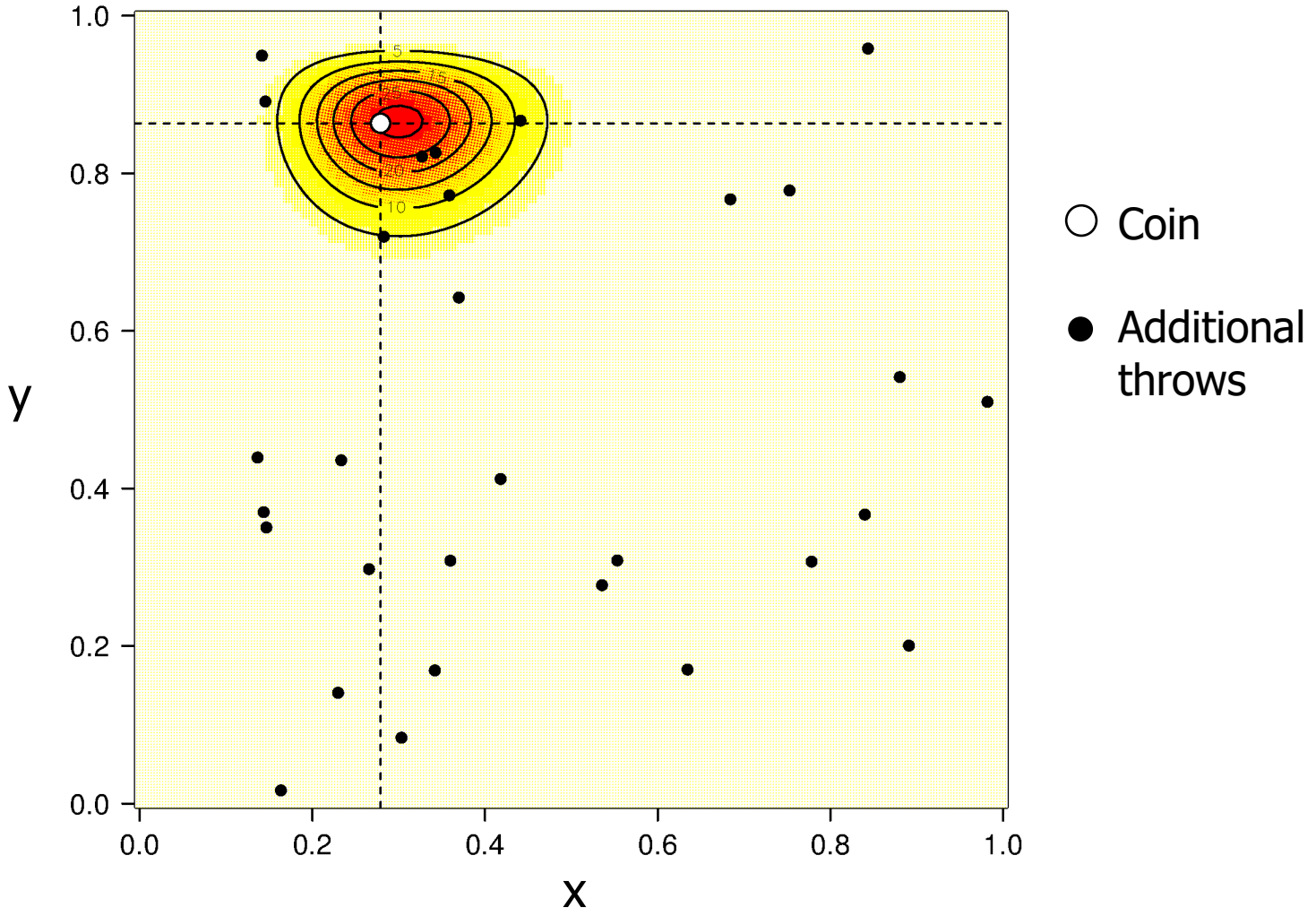


a: 9 # lefts
b: 26 # fronts
n: 30 # throws

$f(x, y \mid T_n) = x^a (1 - x)^{(n-a)} y^b (1 - y)^{(n-b)} / P(T_n)$

R code: https://dosreislab.github.io

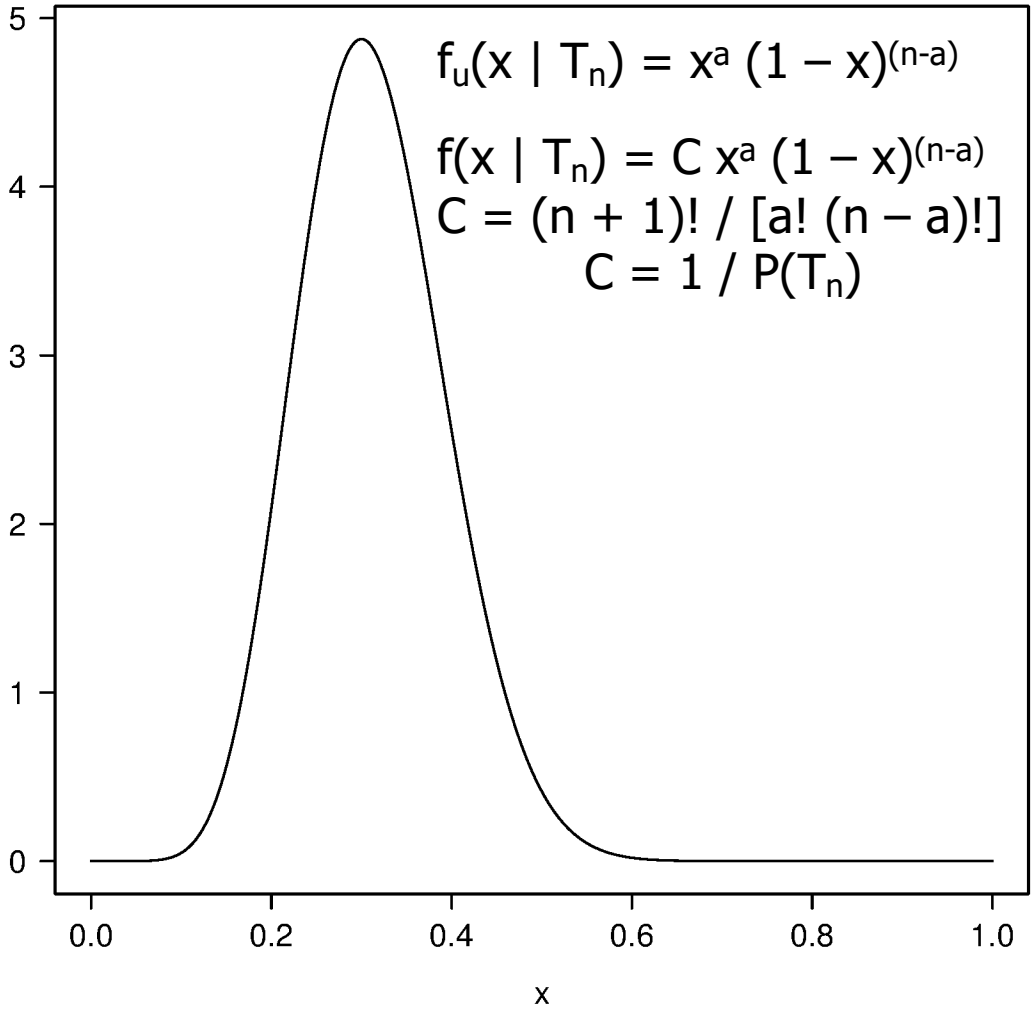
Rev Bayes Thought Experiment (simulation)



Bayesian Terminology

- The marginal of x and y, f(x, y), is known as the **prior distribution** of x and y
- This is because f(x, y) reflects our prior knowledge before any data have been observed
- The conditional $f(T_n \mid x, y)$ is known as the **likelihood** of T_n (the data)
- $P(T_n)$ is known as the **marginal likelihood**
- $f(x, y \mid T_n)$ is known as the **posterior distribution** of x and y
- This is because $f(x, y \mid T_n)$ reflects our posterior knowledge after the data have been observed

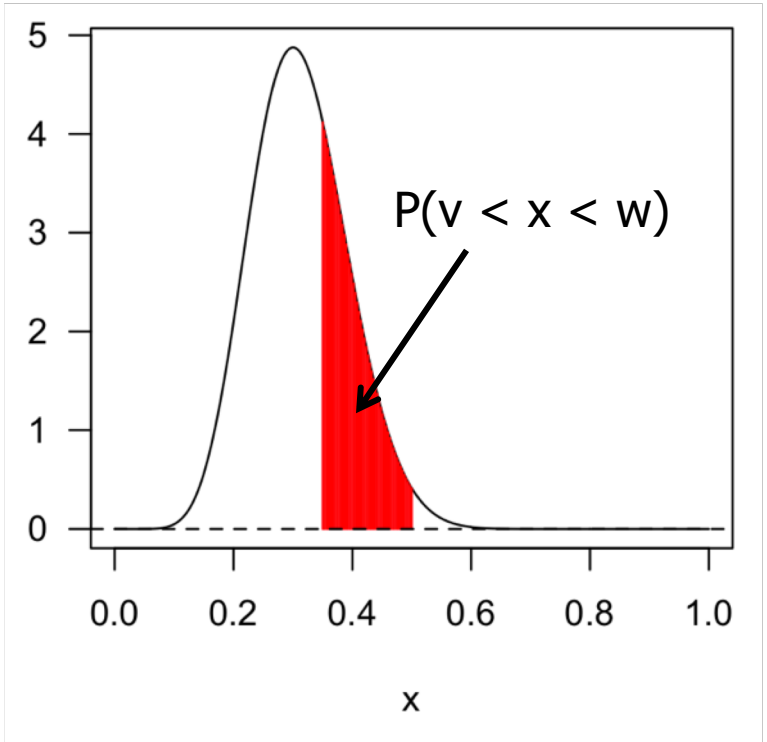
Rev Bayes Thought Experiment



f_u: unnormalized density – has the same shape as the normalised density f

Rev Bayes Thought Experiment

- So, can we ignore the marginal likelihood, $P(T_n)$?
- No.
- The density must be normalised because the probability is the area under the curve:
- $P(v < x < w) = \int_v^w f(x \mid T_n)\, dx$



Note:

- $P(0 < x < 1) = \int_0^1 f(x \mid T_n)\, dx = 1$

- For multi-dimensional densities, the probability is the volume under the surface

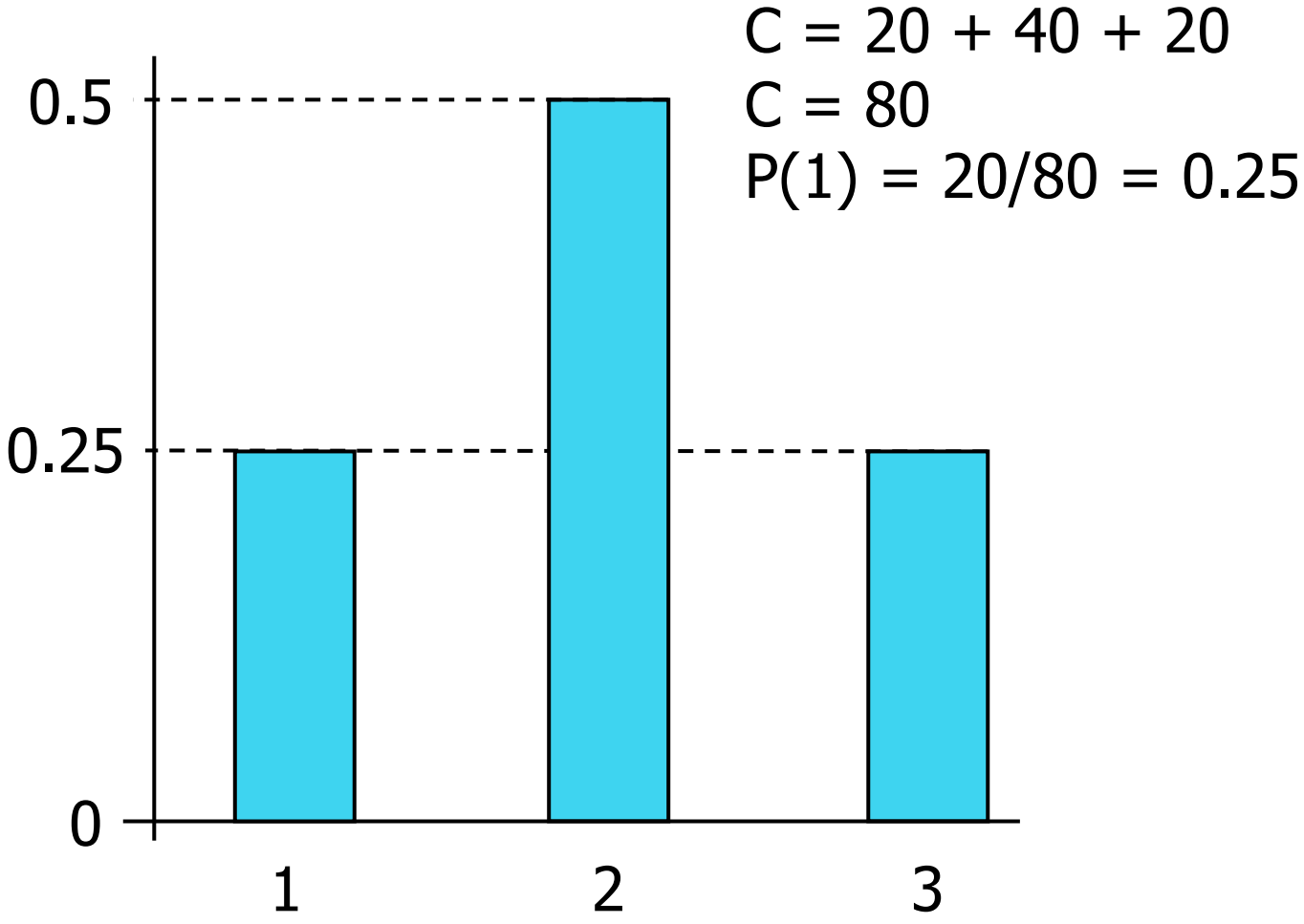
General Bayesian Model

$$f(\theta|D) = f(\theta)f(D|\theta)/f(D)$$

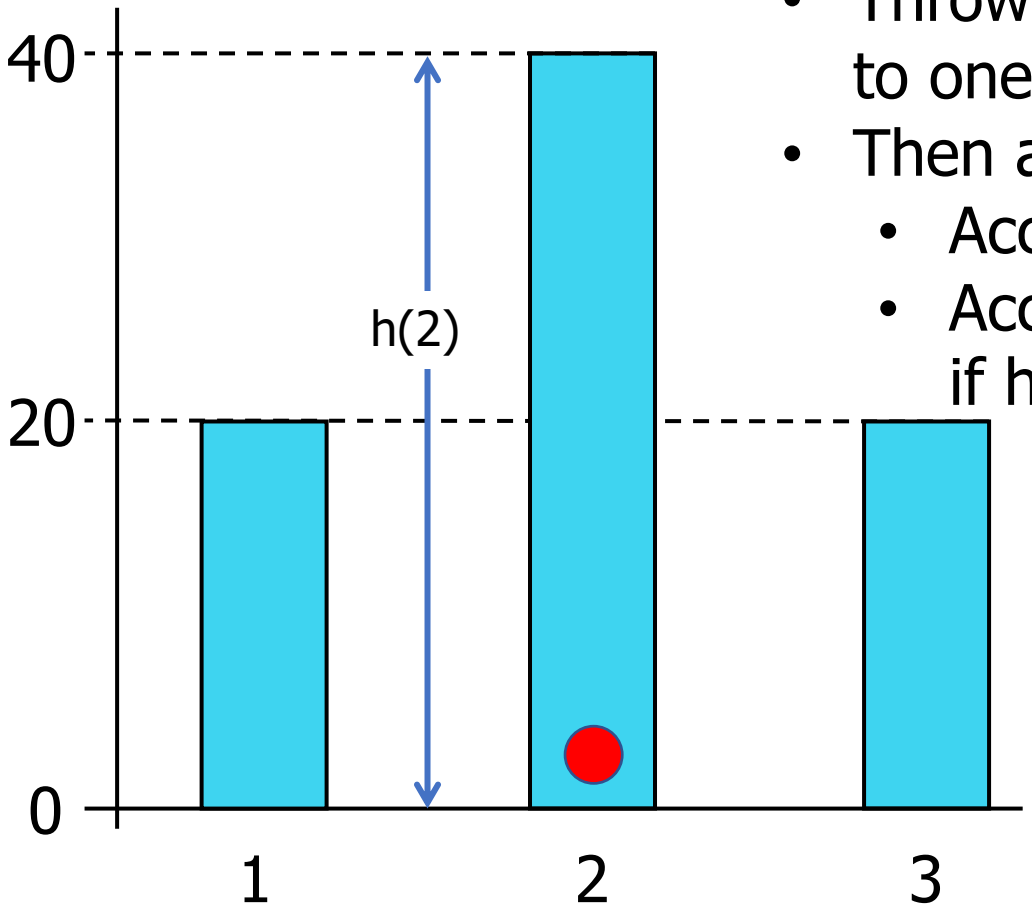
Posterior **Prior** **Likelihood** **Marginal L**

- D: data
- $\theta = (\theta_1, \dots, \theta_n)$: model parameters
- $f(D) = \int f(\theta)f(D|\theta)\mathrm{d}\theta$
- $f(D)$ is an n-dimensional integral
- Usually, this integral **does not** have an analytical solution
- What do we do?

Sampling from Histograms



Sampling from Histograms



Visit-my-neighbour game:

- Select a starting point
- Throw a coin to propose a visit to one of my neighbours
- Then accept/reject visit:
  - Accept* if h(n) > h(m)
  - Accept with P = h(n)/h(m) if h(n) < h(m)

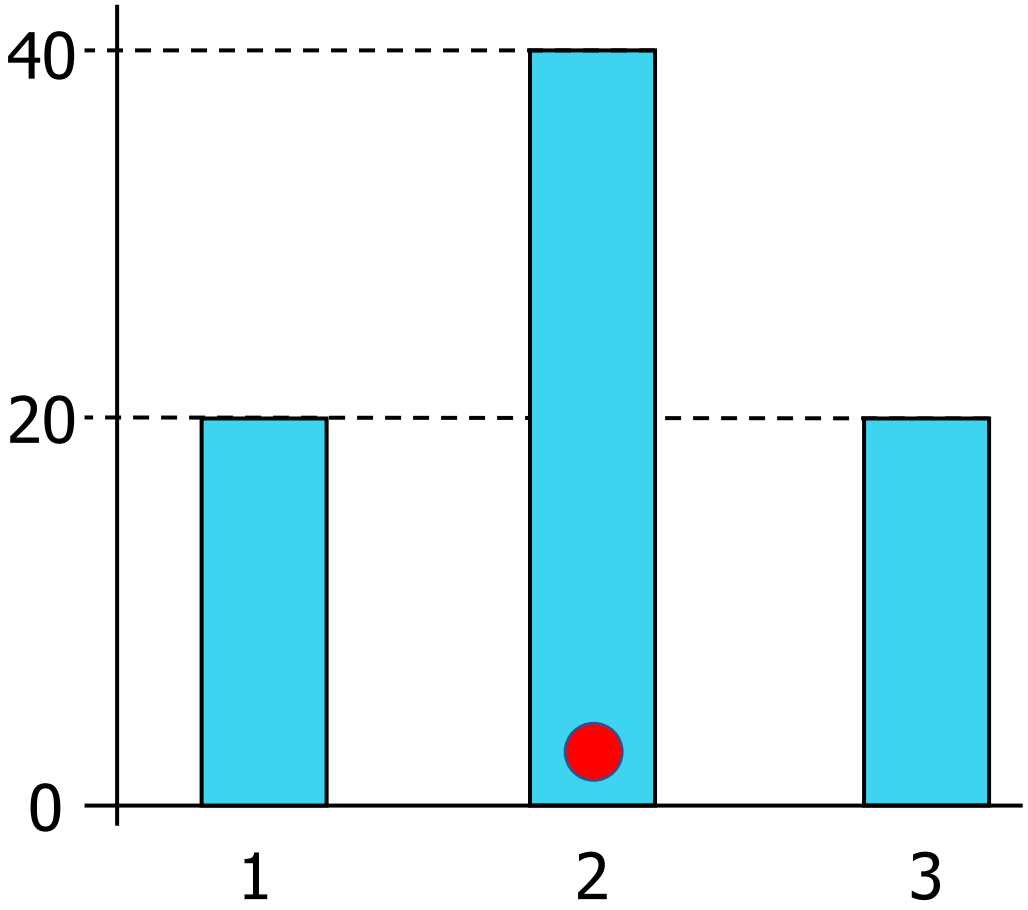
Note:

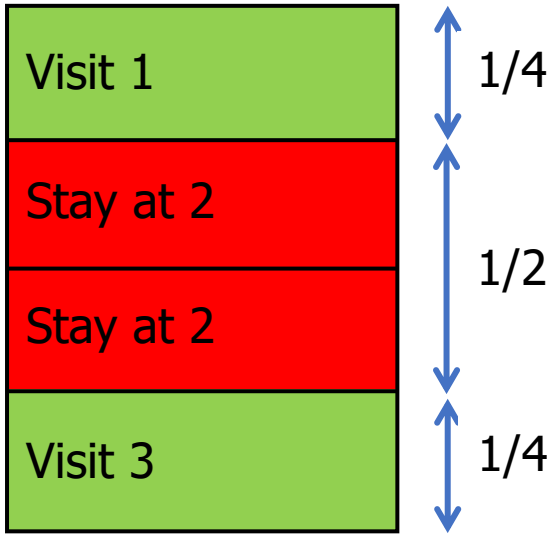
- If I am missing a neighbour, h(n) = 0
- h(n)/h(m) = P(n)/P(m)
- To play this game, we don't need to know C
- Play as long as you want

*n: neighbour, m: myself

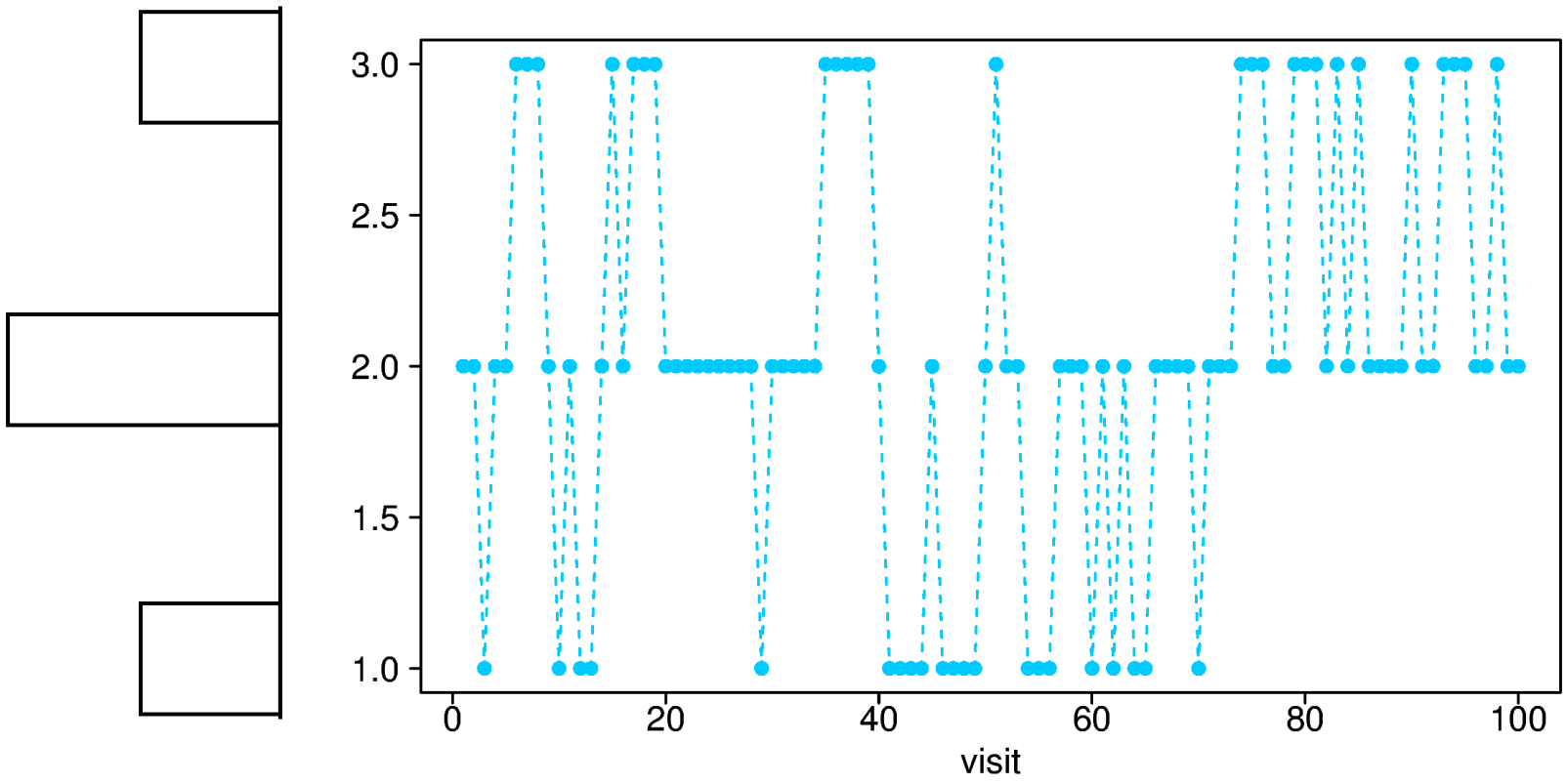
Sampling from Histograms

Visit-my-neighbour game:
Given I'm currently at 2:

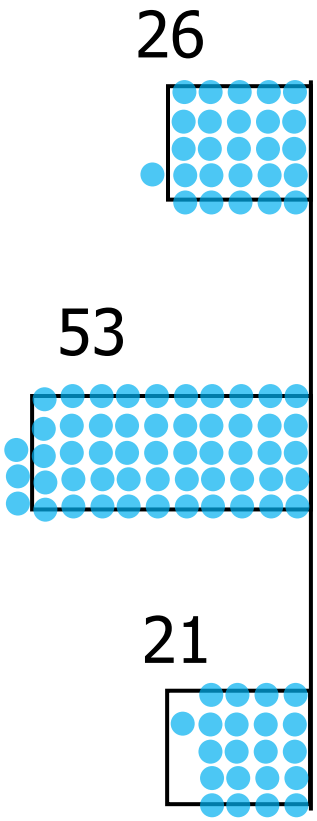


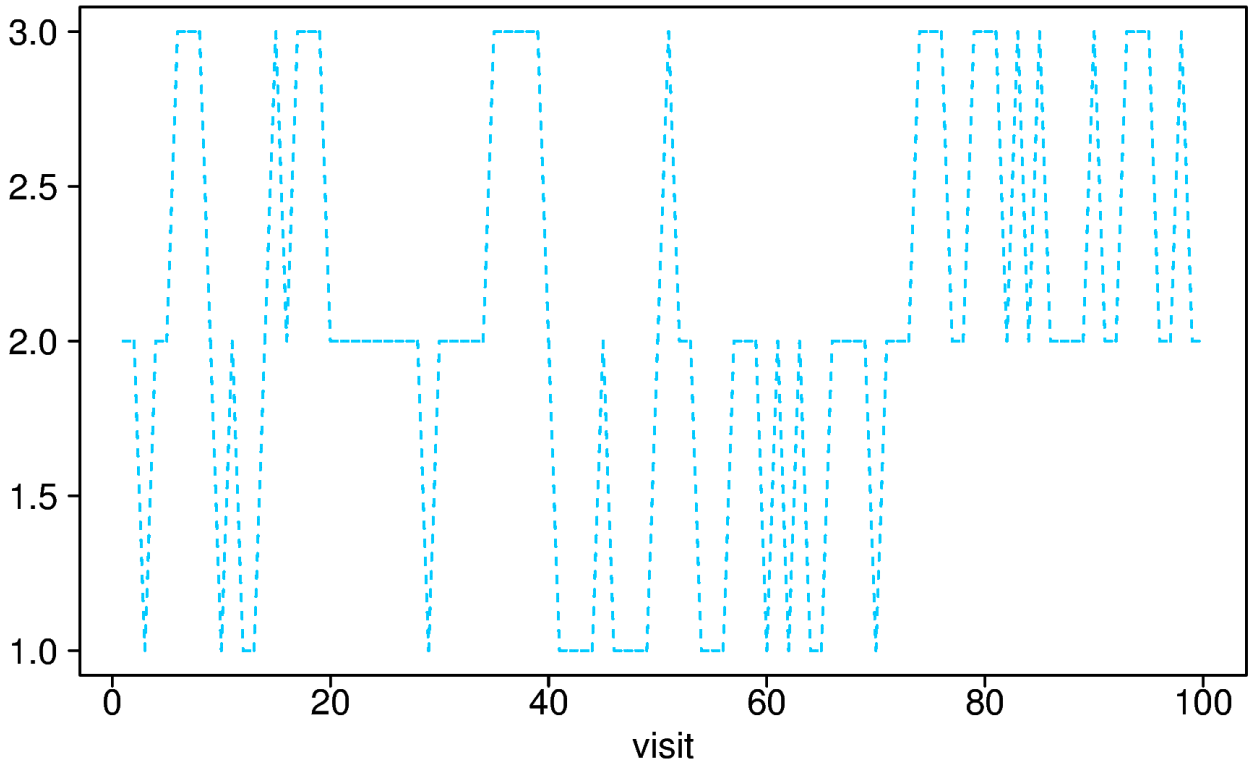


Sampling from Histograms



Sampling from Histograms

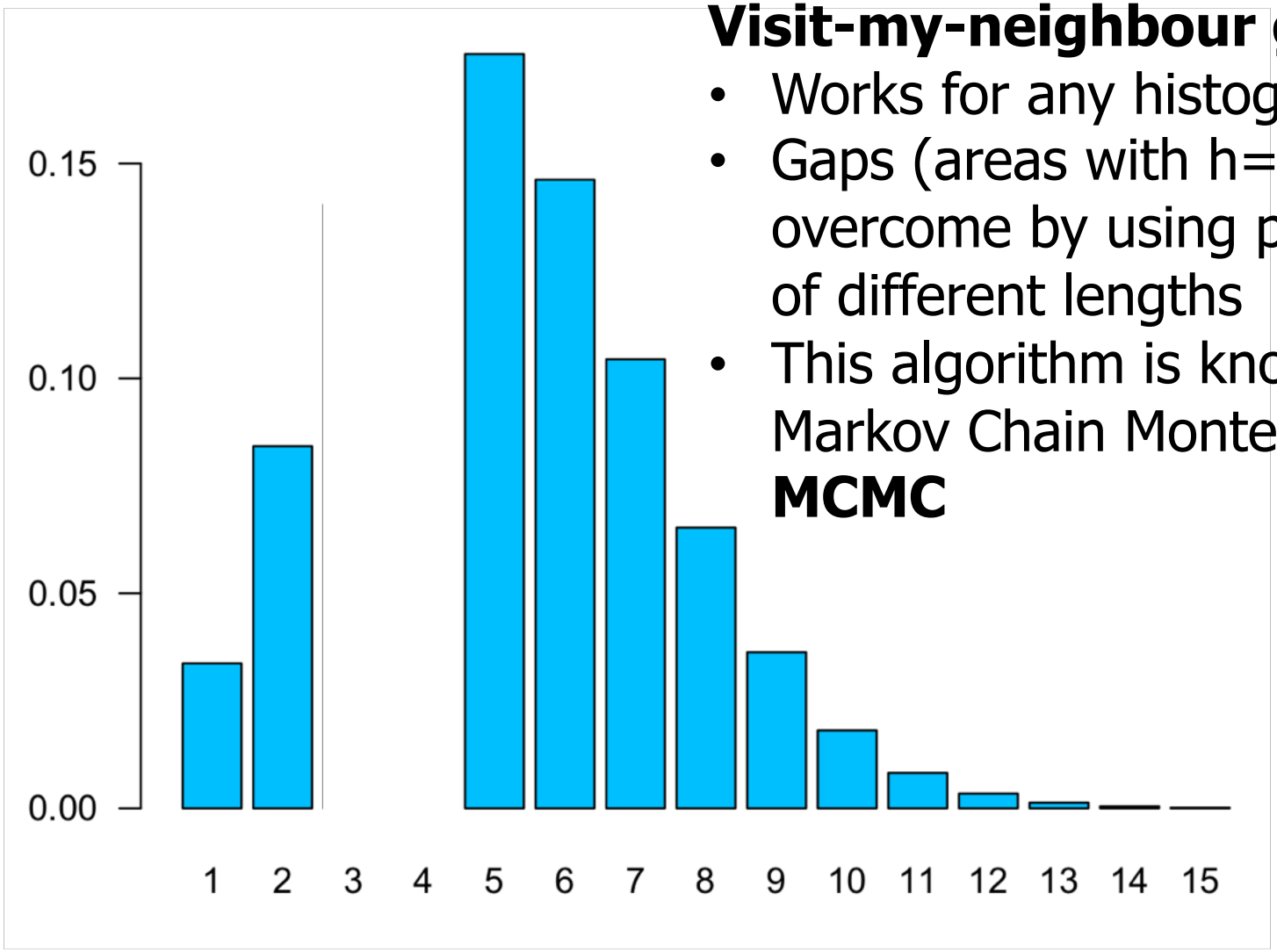




Expected: 25:50:25

In this game, time spent in a site is proportional to the site's probability

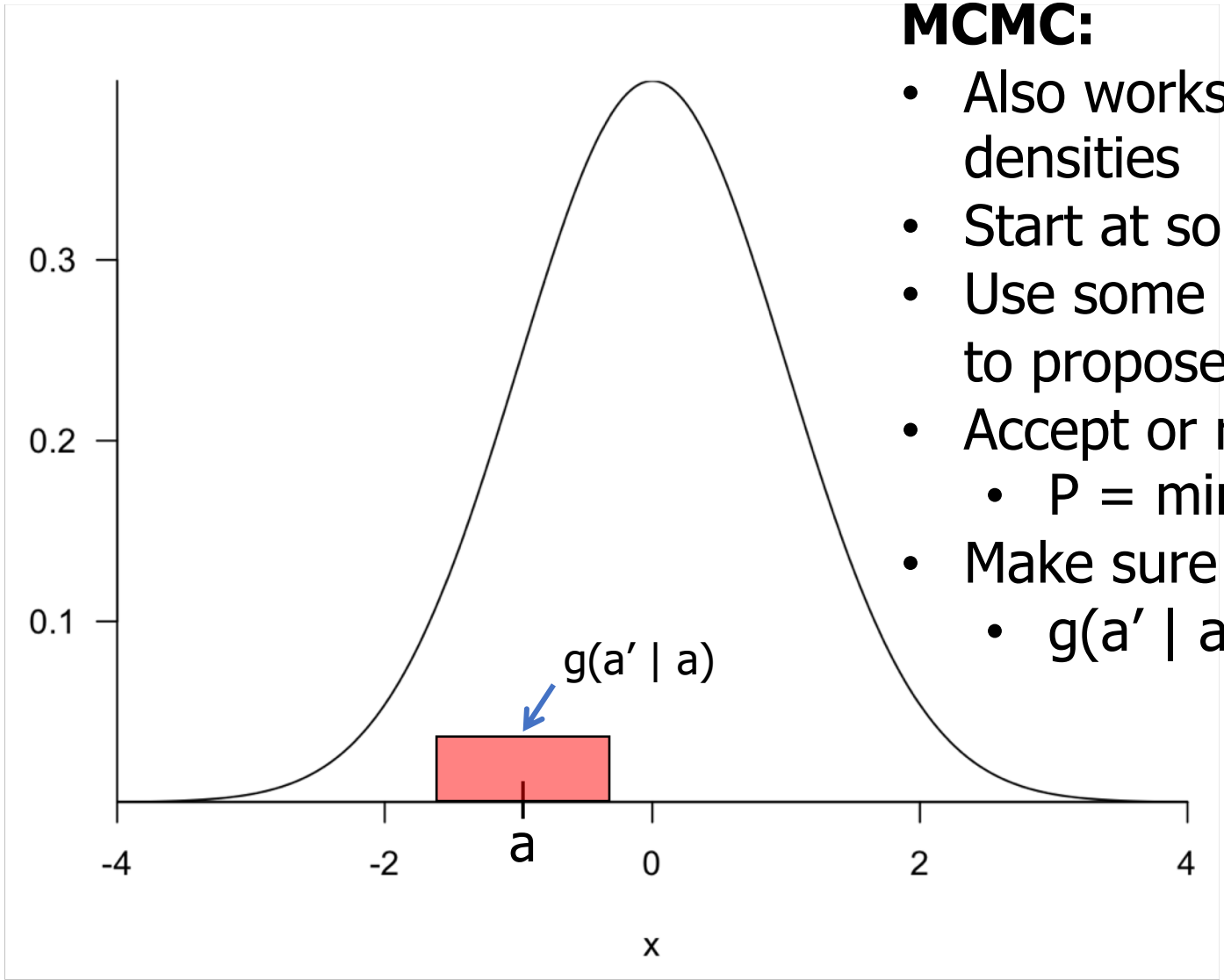
Sampling from Histograms



Visit-my-neighbour game:

- Works for any histogram
- Gaps (areas with h=0) are overcome by using proposals of different lengths
- This algorithm is known as Markov Chain Monte Carlo or **MCMC**

Sampling from Densities



MCMC:

- Also works for continuous densities
- Start at some point a
- Use some density g(a' | a) to propose neighbour a'
- Accept or reject with
  - P = min {1, f(a')/f(a)}
- Make sure:
  - g(a' | a) = g(a | a')

Markov Chain Monte Carlo

THE JOURNAL OF CHEMICAL PHYSICS VOLUME 21, NUMBER 6 JUNE, 1953

Equation of State Calculations by Fast Computing Machines

NICHOLAS METROPOLIS, ARIANNA W. ROSENBLUTH, MARSHALL N. ROSENBLUTH, AND AUGUSTA H. TELLER,
Los Alamos Scientific Laboratory, Los Alamos, New Mexico

AND

EDWARD TELLER,* *Department of Physics, University of Chicago, Chicago, Illinois*

(Received March 6, 1953)

A general method, suitable for fast computing machines, for investigating such properties as equations of state for substances consisting of interacting individual molecules is described. The method consists of a modified Monte Carlo integration over configuration space. Results for the two-dimensional rigid-sphere system have been obtained on the Los Alamos MANIAC and are presented here. These results are compared to the free volume equation of state and to a four-term virial coefficient expansion.

I. INTRODUCTION

THE purpose of this paper is to describe a general method, suitable for fast electronic computing machines, of calculating the properties of any substance

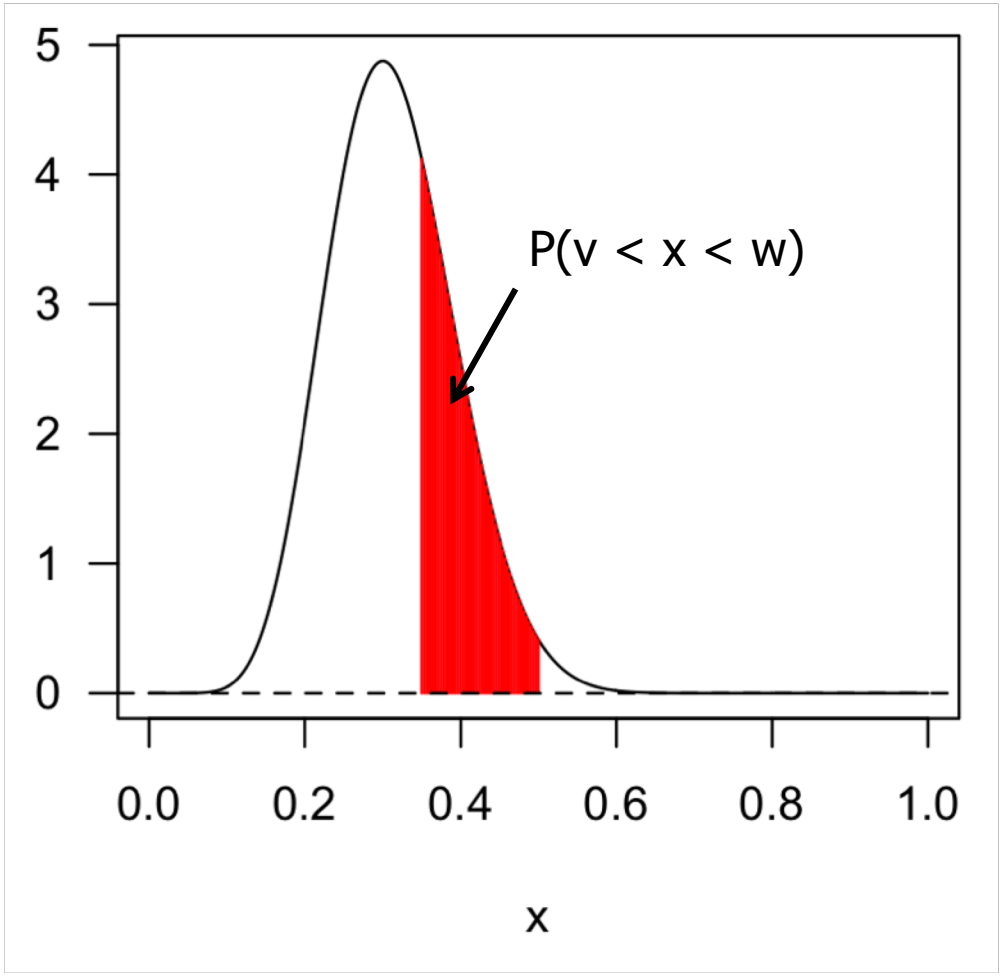
II. THE GENERAL METHOD FOR AN ARBITRARY POTENTIAL BETWEEN THE PARTICLES

In order to reduce the problem to a feasible size for numerical work, we can, of course, consider only a finite

- J. Chem. Phys., (1953) 21: 1087–1092.
- Hastings, Biometrika, (1970) 57: 97–109.
- Monte Carlo: Stan Ulam and John von Newmann

Markov Chain Monte Carlo

- So, how do I calculate $P(v < x < w) = \int_v^w f(x)\mathrm{d}x$?



- $P(v < x < w) \approx \frac{n_a}{N}$
- n_a: times red area was visited
- N: total number of visits
- $\bar{x} = \int_0^\infty x f(x)\mathrm{d}x$
- $\bar{x} \approx \sum_{i=1}^{N} x_i / N$
- x_i: values visited
- MCMC gives you an approximate answer
- Answer gets better with large N

Bayesian Phylogenomics

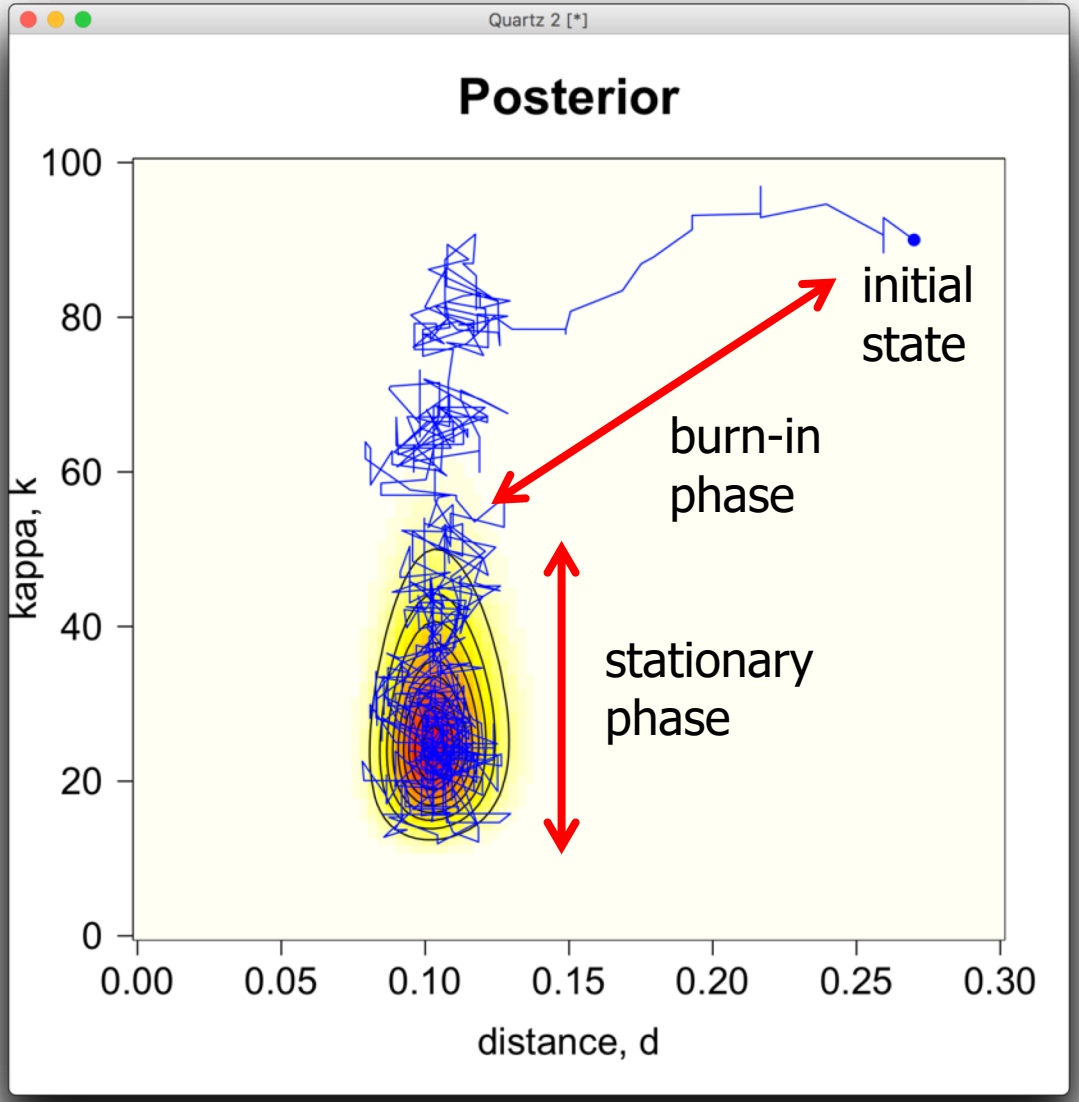
- In phylogenomics our interest may be, for example, in estimating:
  - A tree topology, T
  - The branch lengths, b, given the topology T
  - Some model parameters, θ
  - Given a genomic alignment matrix (our data) **G**

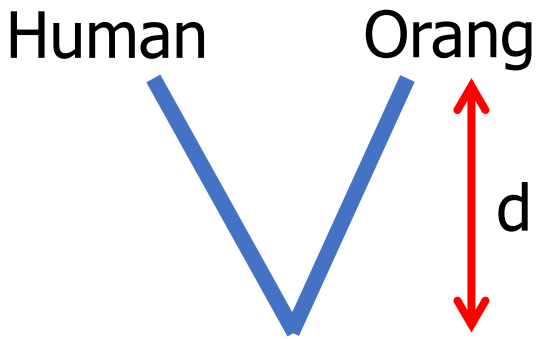
- Posterior distribution of T, b, θ given **G**:

- $f(T, b, \theta|\mathbf{G}) = f(\theta)P(T)f(b|T)\times P(\mathbf{G}|\theta, T, b)/P(\mathbf{G})$

- $P(\mathbf{G}) = \sum_i \int f(\theta, T_i, b)\mathrm{d}\theta\mathrm{d}b$
- $P(\mathbf{G})$ is impossible to calculate, so we need MCMC
- E.g. $P(T|\mathbf{G}) \approx n_T/N$

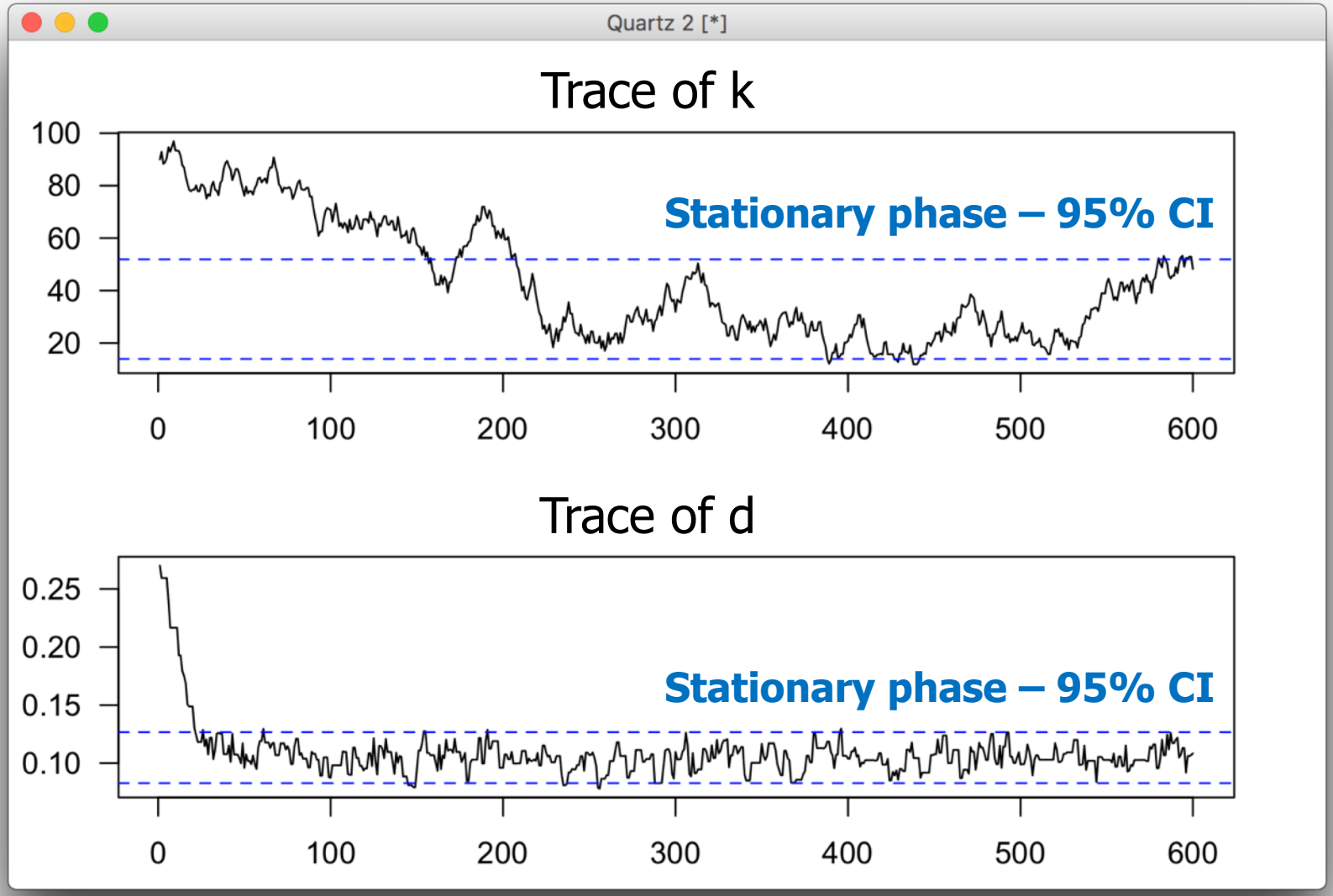
Example: 2s K80 model





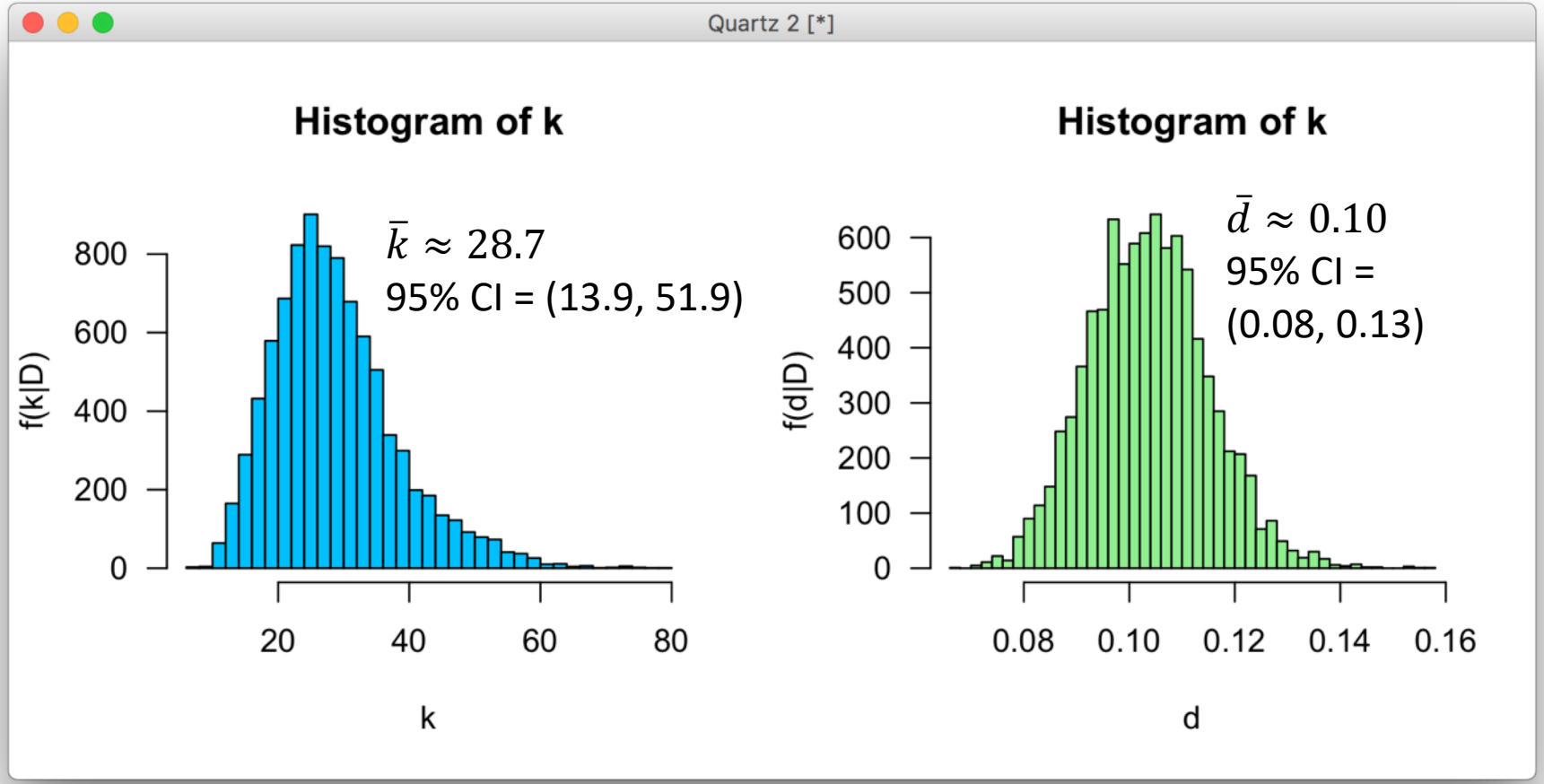
- d: molecular distance
- k: trans/transv ratio
- Kimura (1980) substitution model
- Alignment: 948 mit sites, 84 trans, 6 transv
- (2017) Nat. Ecol. Evol., 1: 1446.

Example: 2s K80 model



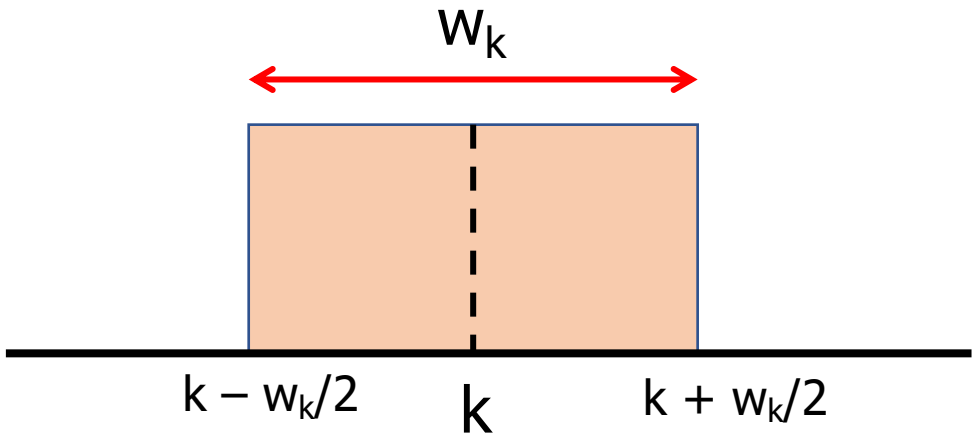
Example: 2s K80 model

- The sample from the stationary phase can be summarised to obtain the approximation to the posterior distribution



Proposal Step Size

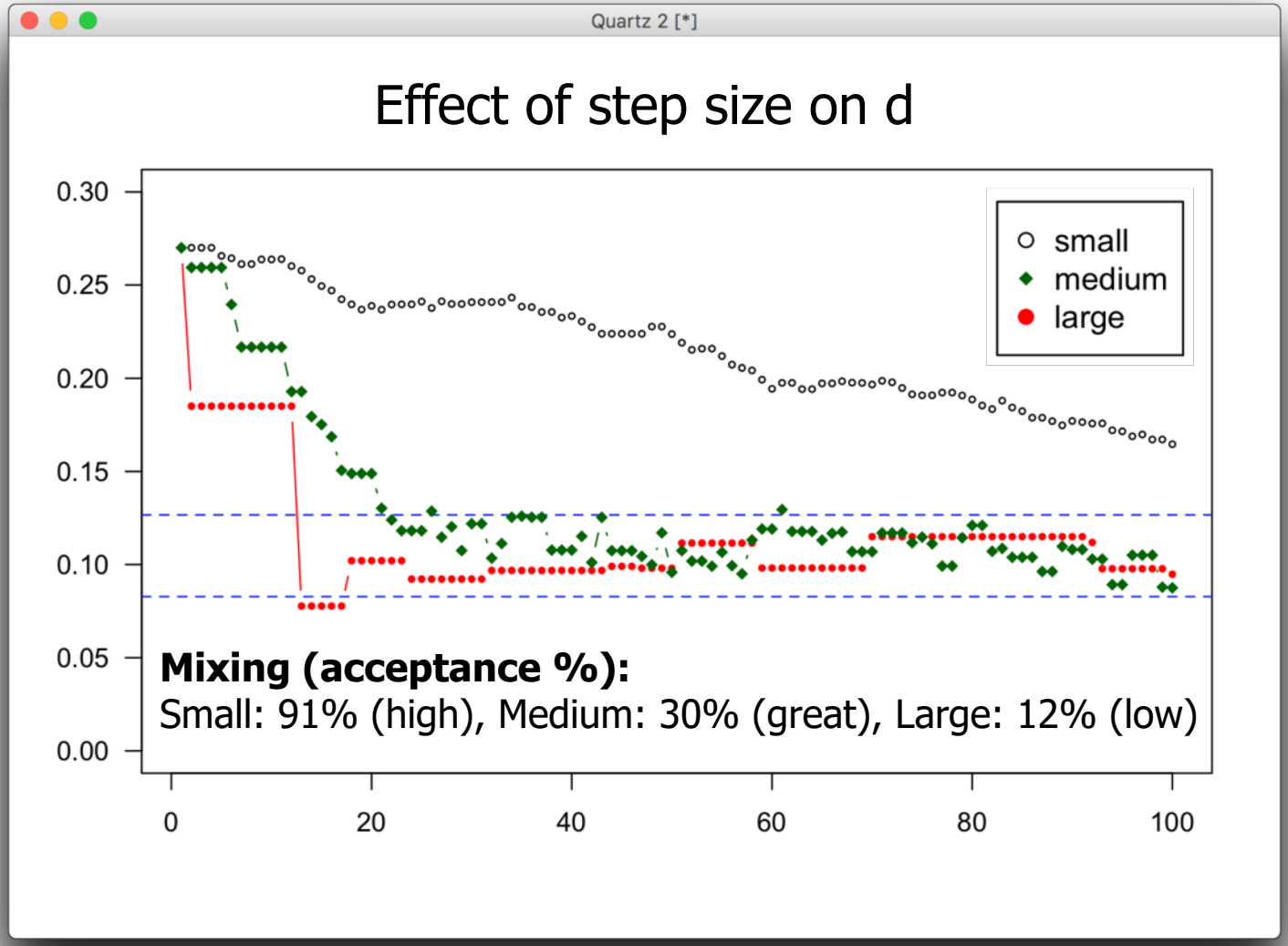
- In this example, we use uniform distributions to propose new values:
  - $d' \sim U(d - w_d/2, d + w_d/2)$
  - $k' \sim U(k - w_k/2, k + w_k/2)$
  - w_d, w_k are known as the proposal step sizes



Mixing and Convergence Rate

- **Mixing:** the ability of the chain to explore state-space quickly
  - If you reject too many proposals you stay in the same place too long
  - If you accept too many proposals you (usually) move slowly and stay in the same region too long
- Proposal step size affect mixing:
  - Step is too big: you reject most proposals
  - Step is too small: you make baby steps
- **Convergence rate:** how quickly the chain moves into the stationary phase
- Proposal step size also affects convergence rate
  - Small sizes lead to low convergence rate

Example: 2s K80 model

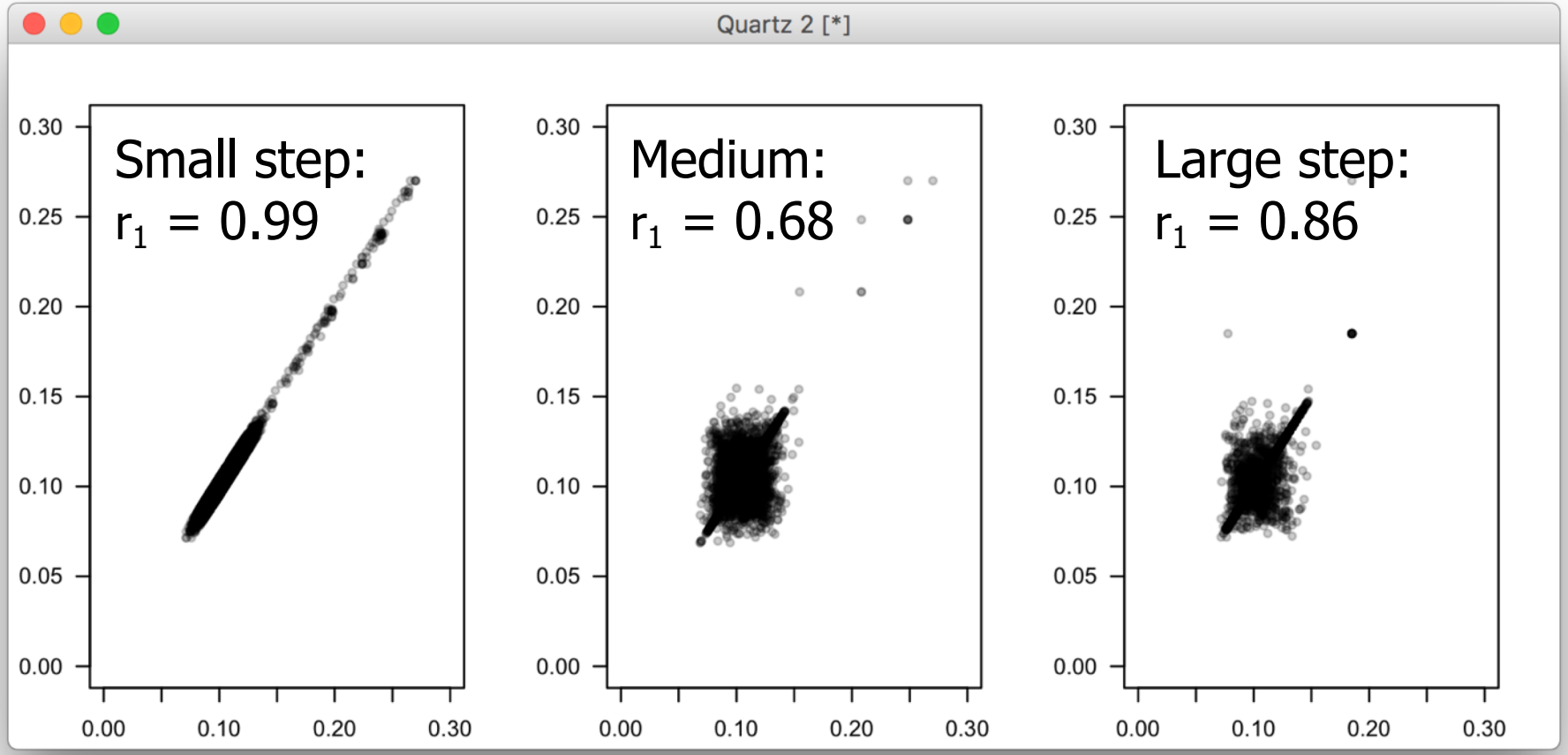


Mixing and Fine-tuning

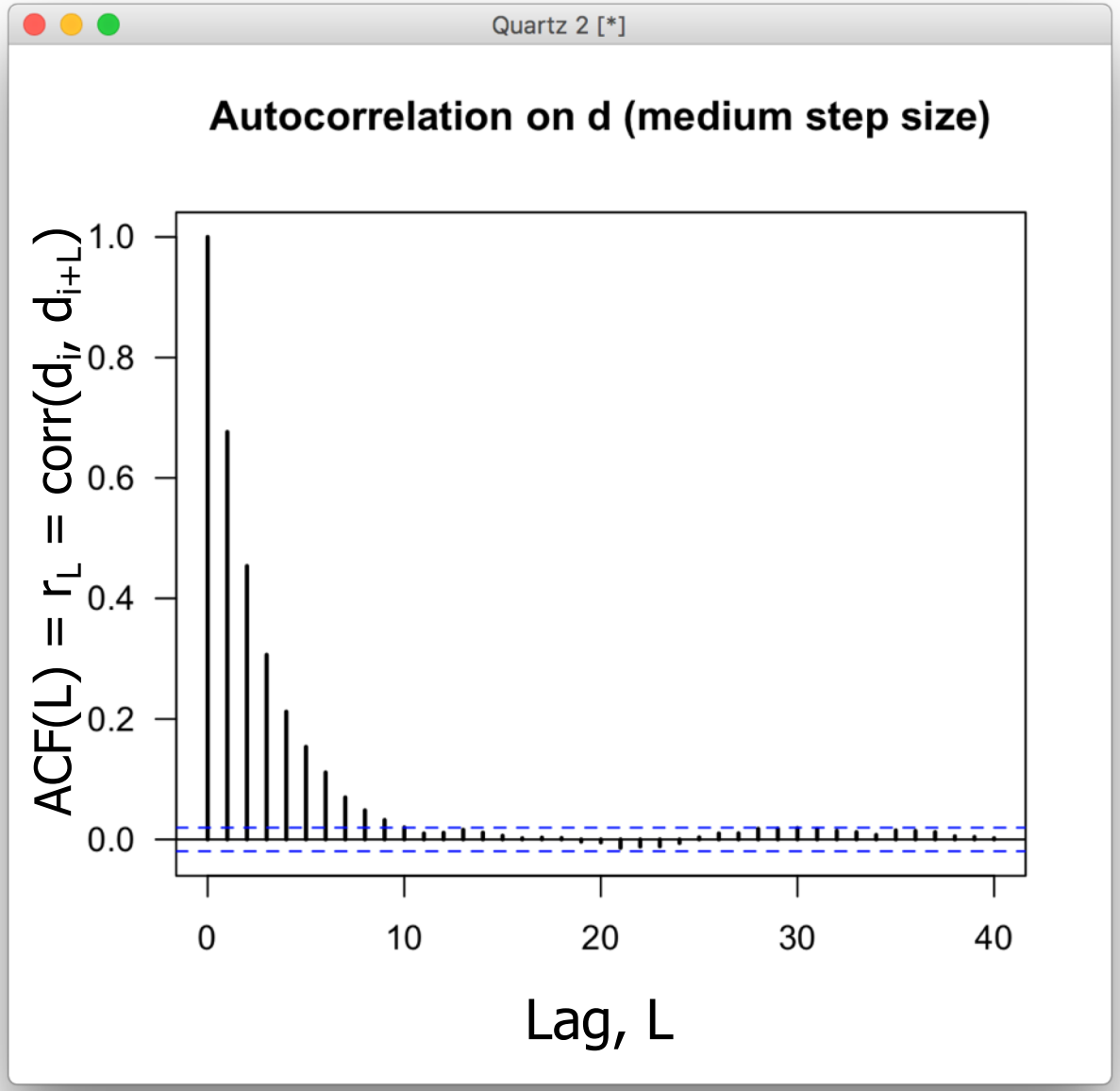
- Analysis of normal distribution indicate that mixing is best at 30% (20% – 40%)
- **Fine-tuning:** Adjusting the step sizes to achieve optimal mixing
- Most MCMC software will do this automatically for you, but sometimes you may need to fix it manually:
  - % is too high: increase step size
  - % is too low: decrease step size
- Remember MCMC estimates are approximate:
  - $\bar{d} \approx \sum_i d_i / N$
- For two chains with the same length, the errors in the estimates are larger for the chain with poorest mixing
- Remember calculations are done after removing burn-in samples

Autocorrelation

- MCMC samples are autocorrelated because accepted values are modifications of the previous values
- 2s K80 example, $r_1 = \text{corr}(d_i, d_{i+1})$:



Autocorrelation Function



- Chains that mix well have ACF that decay fast

Efficiency

- Chains that lead to estimates with small errors with respect to the chain's size are said to be **efficient**
- Efficiency relates to the autocorrelation of the chain:
  - **High (+) autocorrelation:** Low efficiency
  - **Moderate (+) autocorrelation:** Efficient chain
  - **No autocorrelation:** Independent sampling (very efficient)
  - **(-) autocorrelation:** Super efficient chain

- Efficiency:
  - $\text{Eff} = 1/[1 + 2(r_1 + r_2 + r_3 \ldots)]$
  - Eff = 1: as efficient as independent sampling
  - Eff = 0.2: 20% as efficient as independent sampling

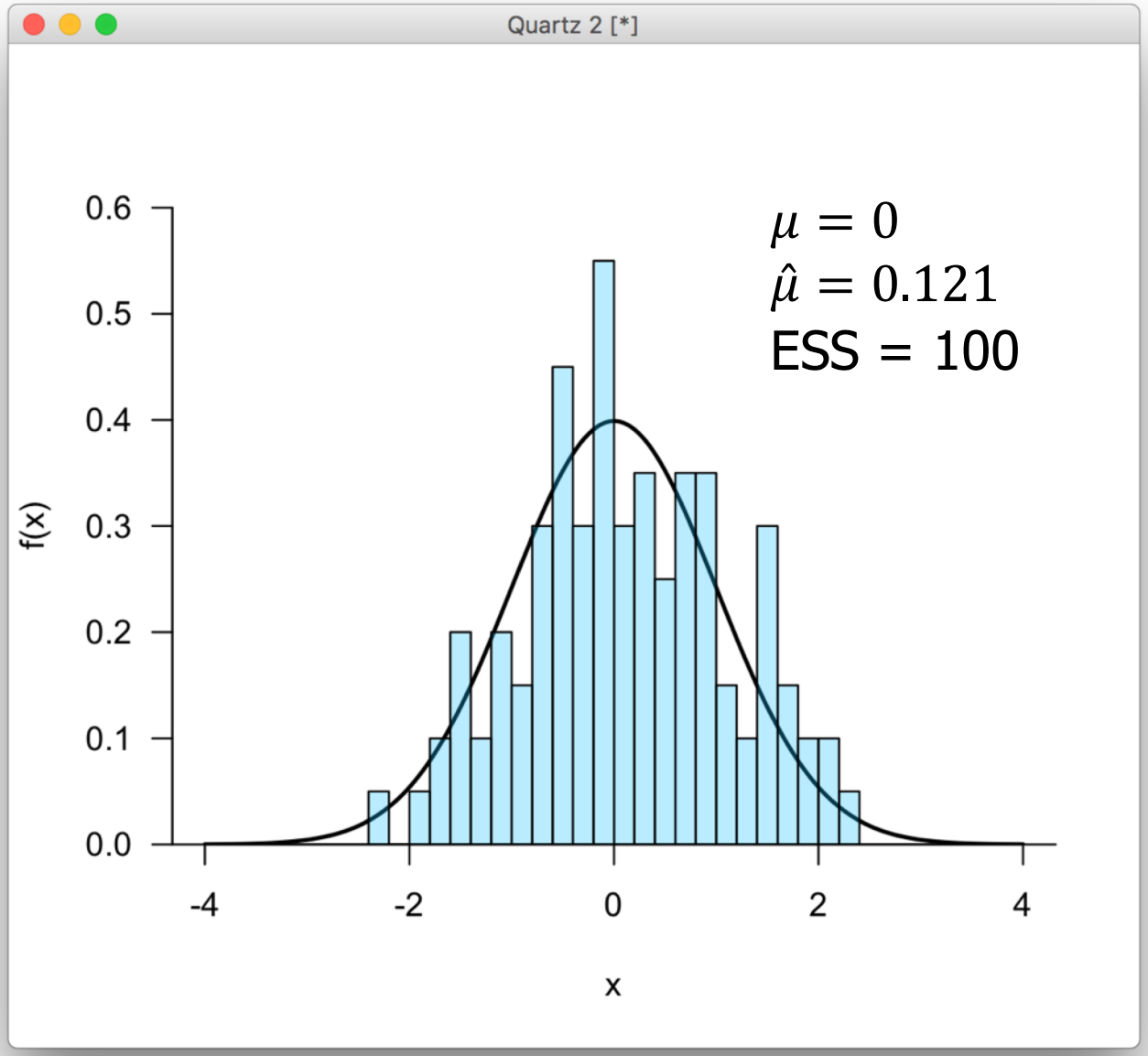
Effective Sample Size (ESS)

- Effective Sample Size = Chain Size × Efficiency
- $\mathrm{ESS} = N \times \mathrm{Eff}$
- For example, MCMC chain with N=1,000 and Eff=20%
- Then ESS = 200, meaning the chain has the same estimate error as an equivalent, independent chain of size 200
- Stochastic simulation theory recommendation:
  - N should be between 1,000 to 10,000 for independent sampling
  - Thus, ESS should be between 1,000 to 10,000
  - This is hard to achieve in Bayesian phylogenomics
  - You must try to have at least ESS > 200

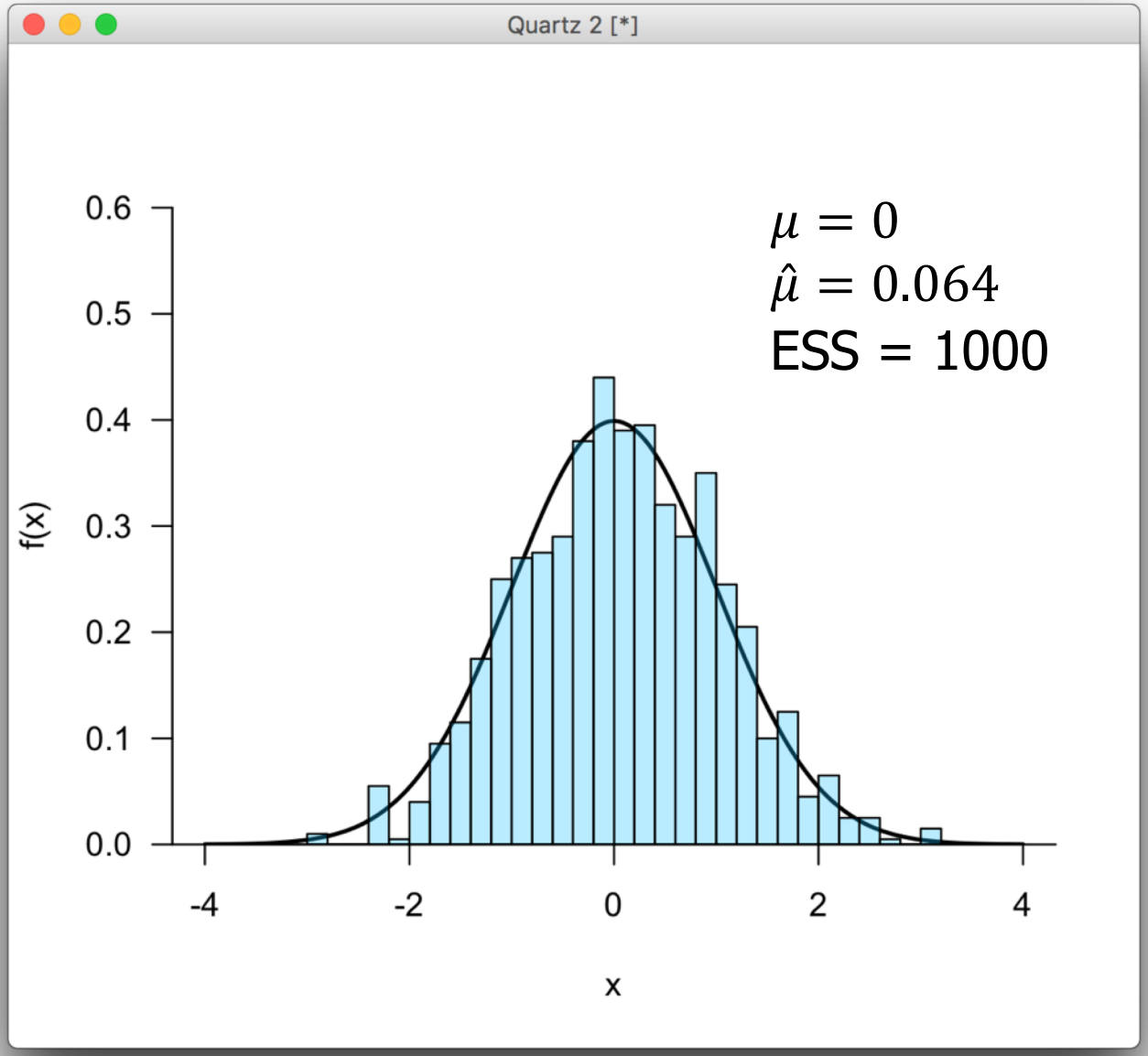
Convergence

- MCMC is an stochastic algorithm
- This means an MCMC histogram is an approximation of the posterior density
- This approximation improves as $N \rightarrow \infty$
- You must use **convergence diagnostics** to assess whether the MCMC sample has converged to the posterior

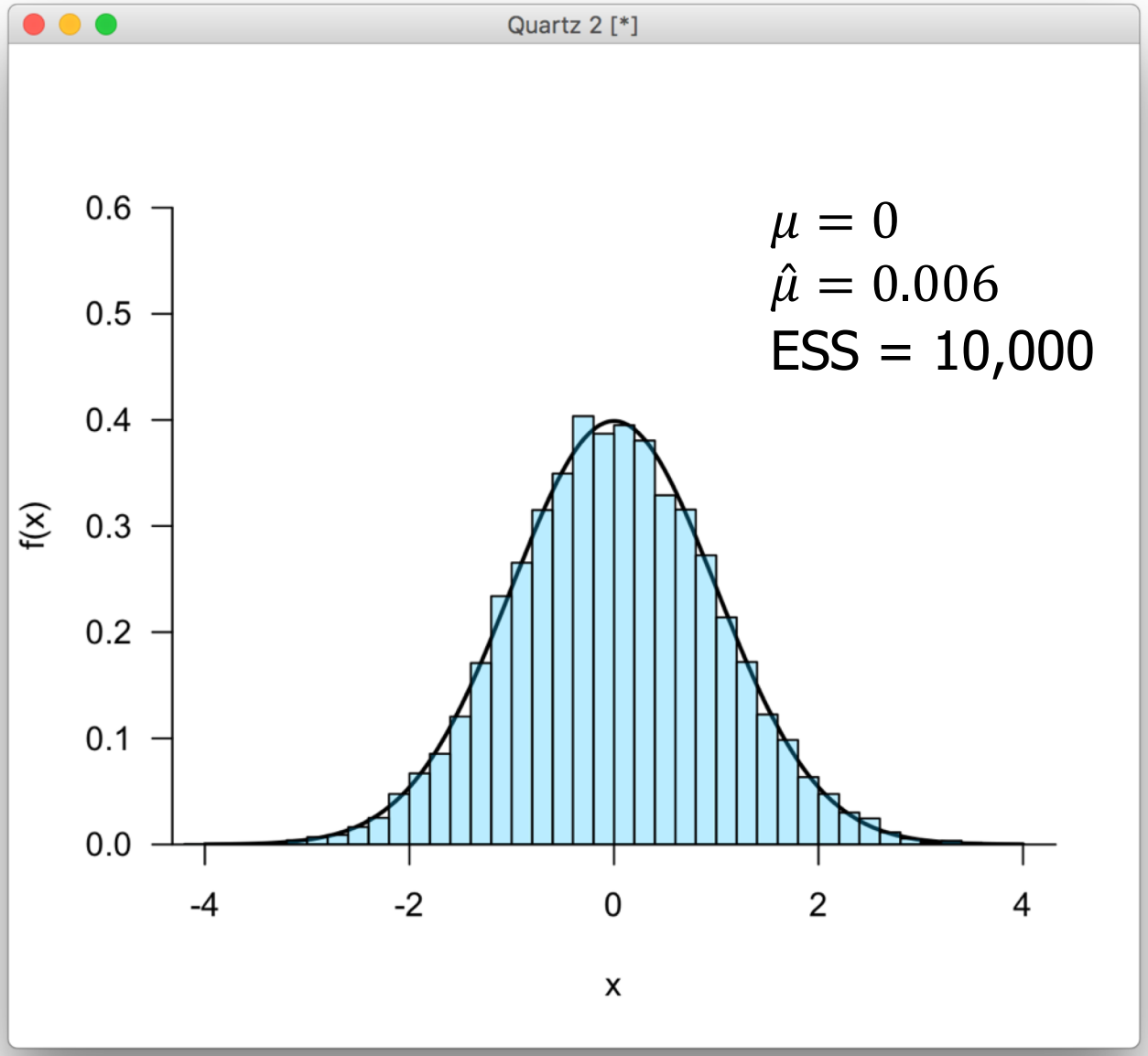
Convergence to Normal Dist



Convergence to Normal Dist



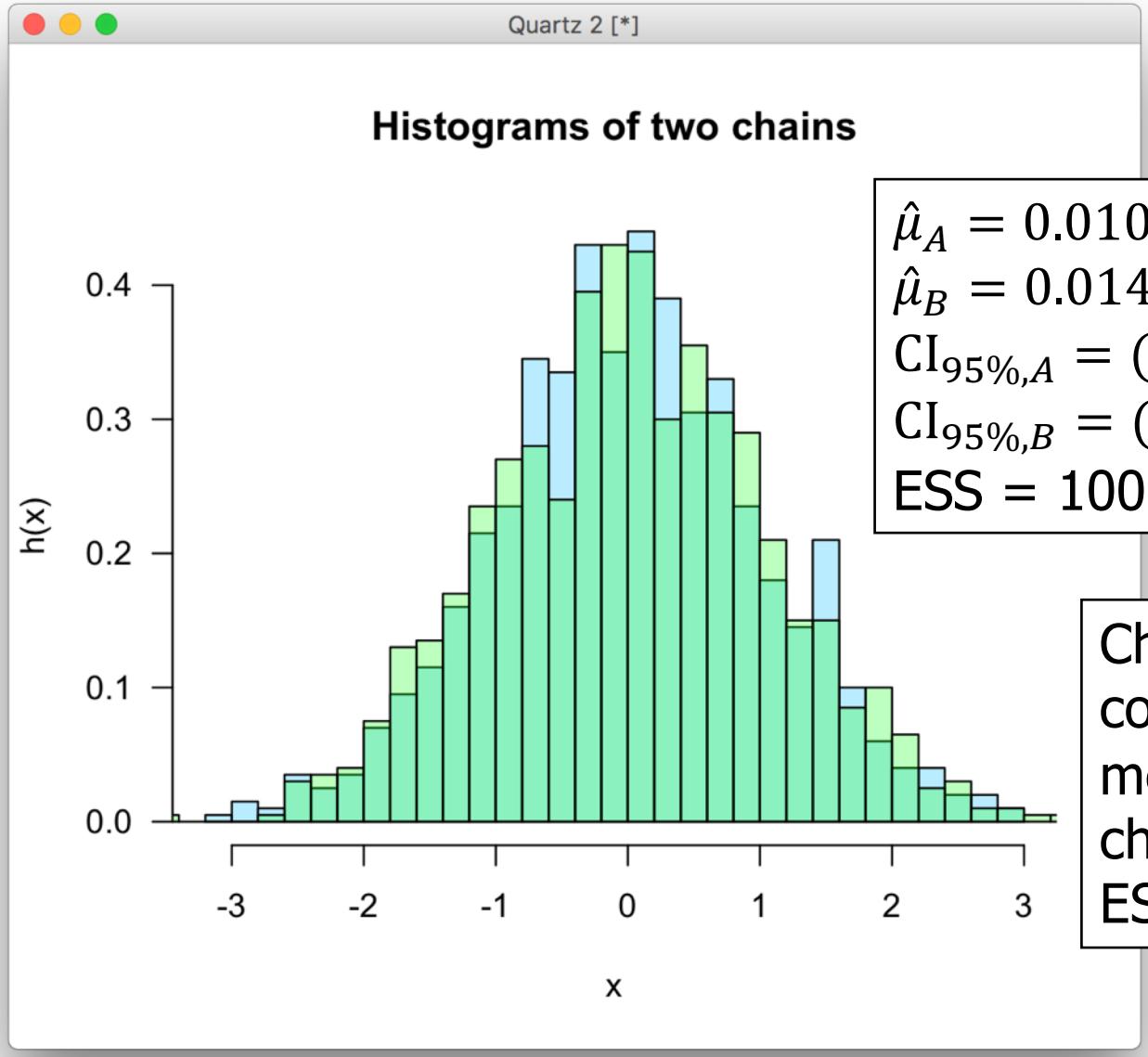
Convergence to Normal Dist



Convergence

- In practice the shape of the posterior density is not known
- Thus, you cannot compare your MCMC histogram to the true posterior
- The way around this is to **run two or more** MCMC chains and compare their histograms, traces, posterior means, and credibility intervals
- If they are similar it is likely you have converged
- **Important:**
  - The chains should start from different starting points
  - Starting points can be chosen randomly or
  - chosen so that they are over-dispersed

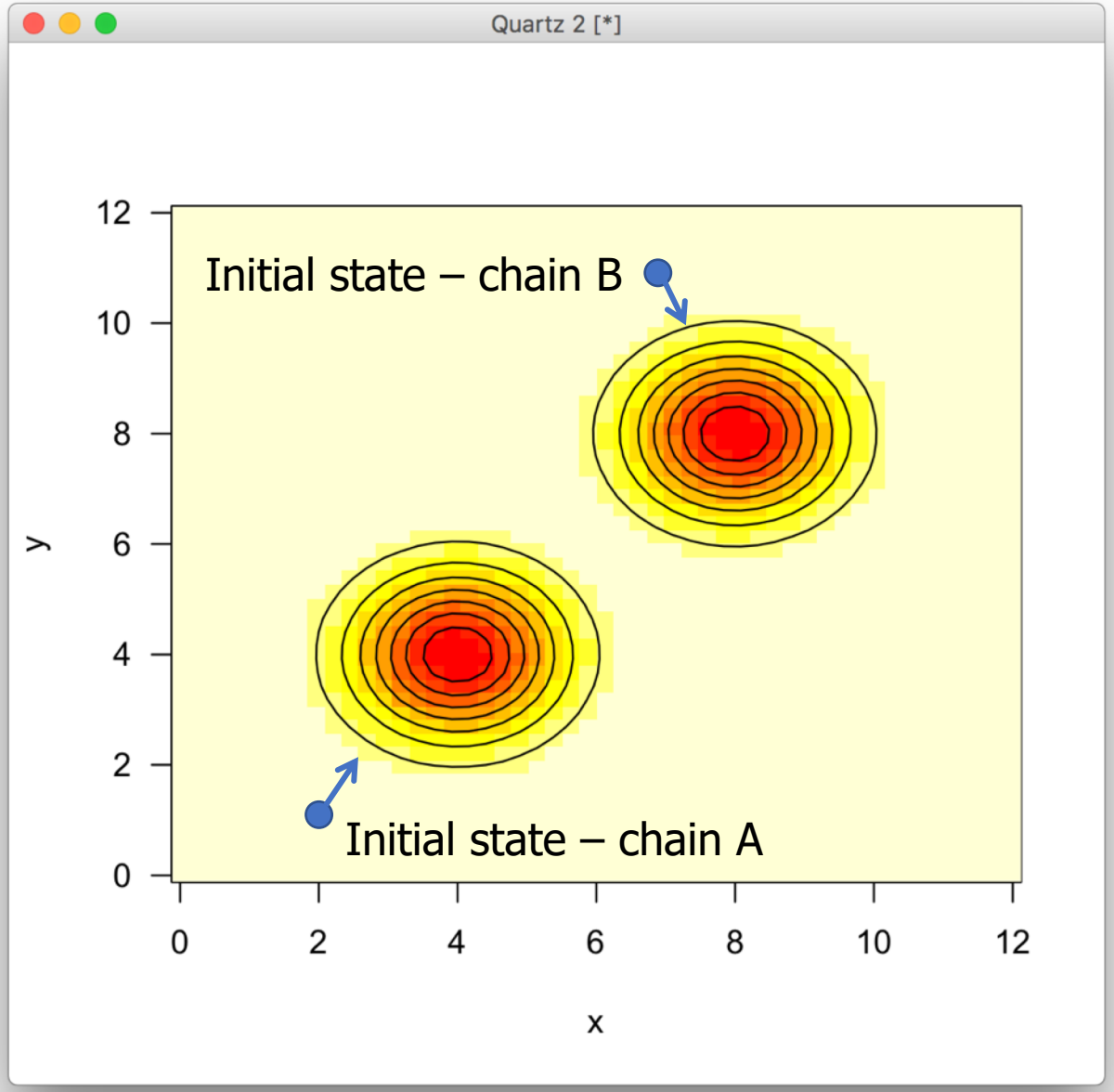
Convergence



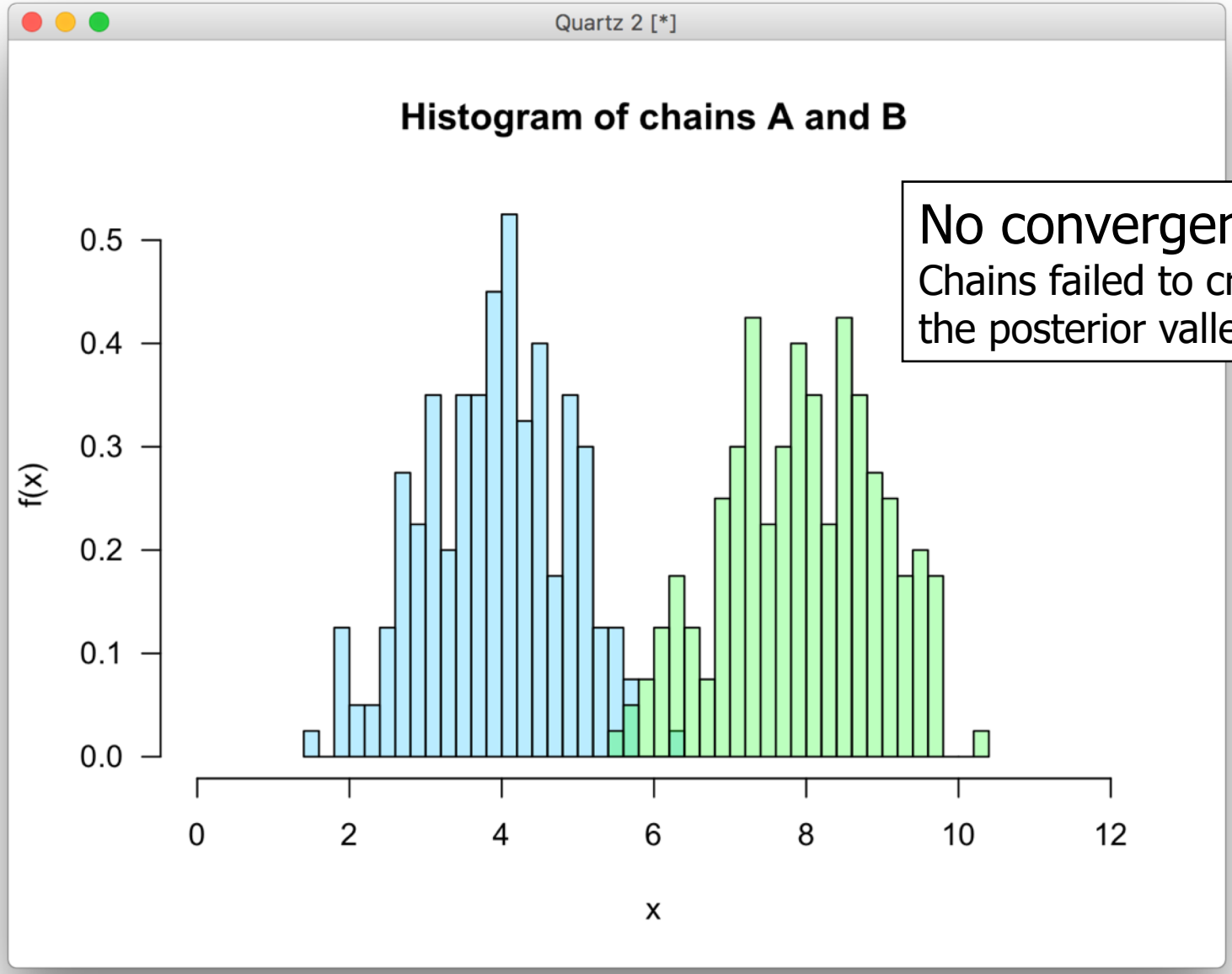
$\hat{\mu}_A = 0.010$
$\hat{\mu}_B = 0.014$
$\mathrm{CI}_{95\%,A} = (-1.99, 2.01)$
$\mathrm{CI}_{95\%,B} = (-1.98, 2.05)$
ESS = 1000

Chains that have converged can be merged into a larger chain
$\mathrm{ESS}_L = \mathrm{ESS}_A + \mathrm{ESS}_B$

Multi-modal Densities



Multi-modal Densities



No convergence!
Chains failed to cross the posterior valley

Multi-modal Densities

- Using over-dispersed or random starting points is a good way to detect multi-modal posteriors
- **If you detect a multi-modal posterior:**
  - Run the chains for a **very long time**
  - Eventually, the chains will cross the valley back and forth and the histograms will convergence
  - Note you **should not** merge short chains that are stuck at different modes
  - This is because the probability of the chain getting stuck is different from the probability mass under the mode
- Using fixed starting points is a bad idea
- ESS is not a measure of convergence

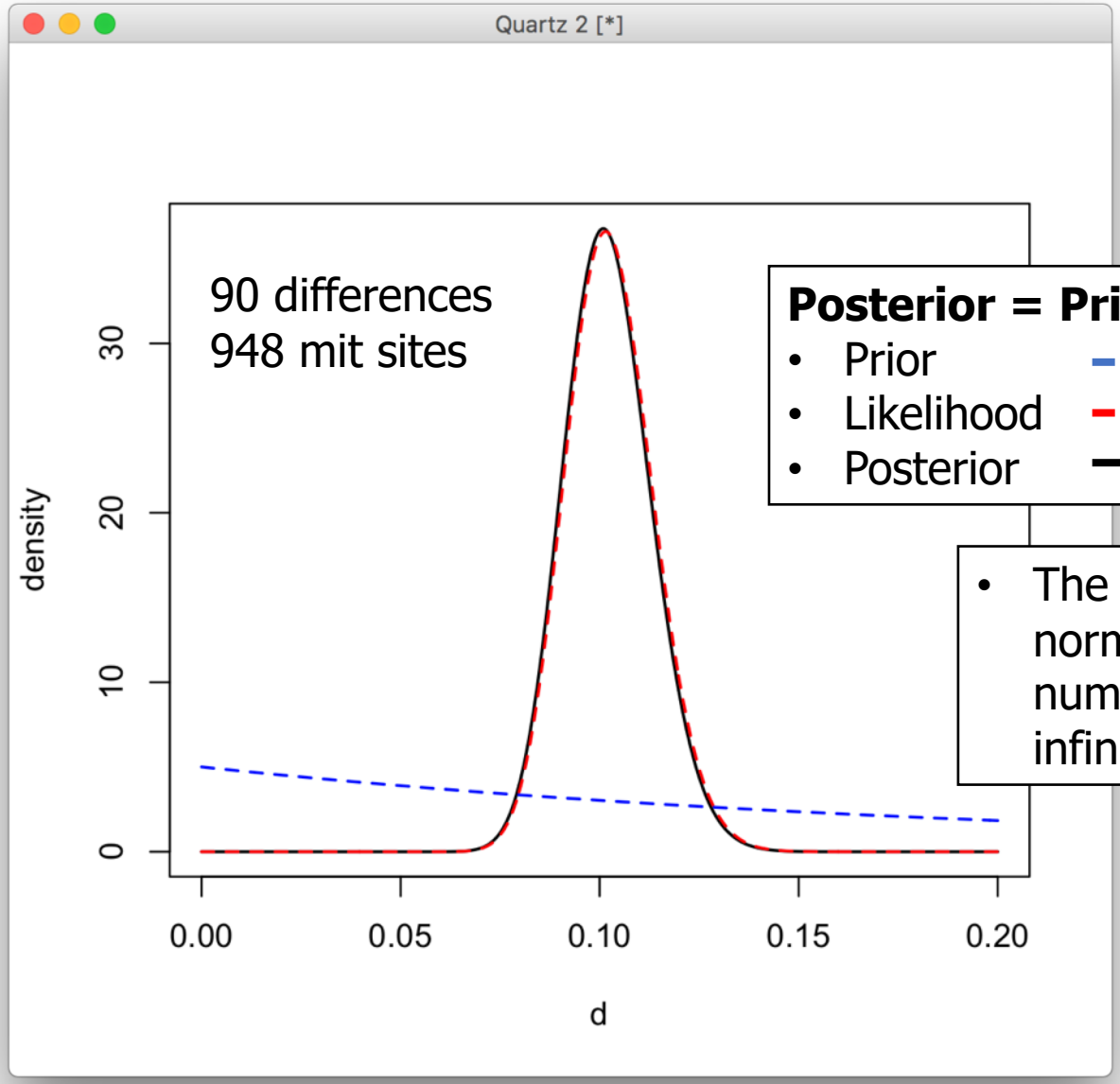
Thinning the Chain

- In phylogenomics it is difficult to construct efficient chains
- This happens because we usually have too many parameters
- Real-life phylogenomic MCMC chains are highly correlated
- To get good estimates, you must run the chain for a very long time
- If you write down every chain visit, you would run out of hard drive space very quickly
- **Thinning:** Writing down only a fraction of all chain visits (e.g. every 100^{th} or 1000^{th} visit)

Bayesian Asymptotics

- Asymptotics refers to how the posterior estimates behave as our sample size (the amount of data) increases
- In well-behaved problems, the posterior converges to the shape of the likelihood
- That is, the prior has little relevance when we have a lot of data

Example: 2s JC69



Posterior = Prior x Likelihood

- Prior
- Likelihood
- Posterior

- The likelihood tends to the normal distribution as the number of sites goes to infinity

Example: Phylogenomic Dating

- In phylogenomics, the likelihood of the alignment is the product of the likelihood of sites
- $L(\mathbf{G}|T, b) = \prod_{i=1}^{P} L(\mathbf{g}_i|T, b)$
- $\mathbf{g}_i$: i-th site pattern
- P: number of site patterns
- P: can be over one million in a phylogenomic alignment
- When analysing large genomes, the likelihood of the branch lengths **given the tree** is very close to the multivariate-normal (MVN) distribution
- Thus, we can approximate the likelihood using the MVN
- This is **much faster** than traditional likelihood

Bayesian Dating

- Thorne et al. (1998) MBE, 15: 1647 developed the approximation idea for Bayesian clock-dating

Estimating the Rate of Evolution of the Rate of Molecular Evolution

Jeffrey L. Thorne, Hirohisa Kishino,† and Ian S. Painter**

*Program in Statistical Genetics, Statistics Department, North Carolina State University; and †Department of Social and International Relations, University of Tokyo

A simple model for the evolution of the rate of molecular evolution is presented. With a Bayesian approach, this model can serve as the basis for estimating dates of important evolutionary events even in the absence of the

Approximate Likelihood Calculation on a Phylogeny for Bayesian Estimation of Divergence Times

Mario dos Reis[1] and Ziheng Yang*[,1,2]

[1]Department of Biology, University College London, Darwin Building, Gower Street, London, United Kingdom

[2]Center for Computational and Evolutionary Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

***Corresponding author:** z.yang@ucl.ac.uk

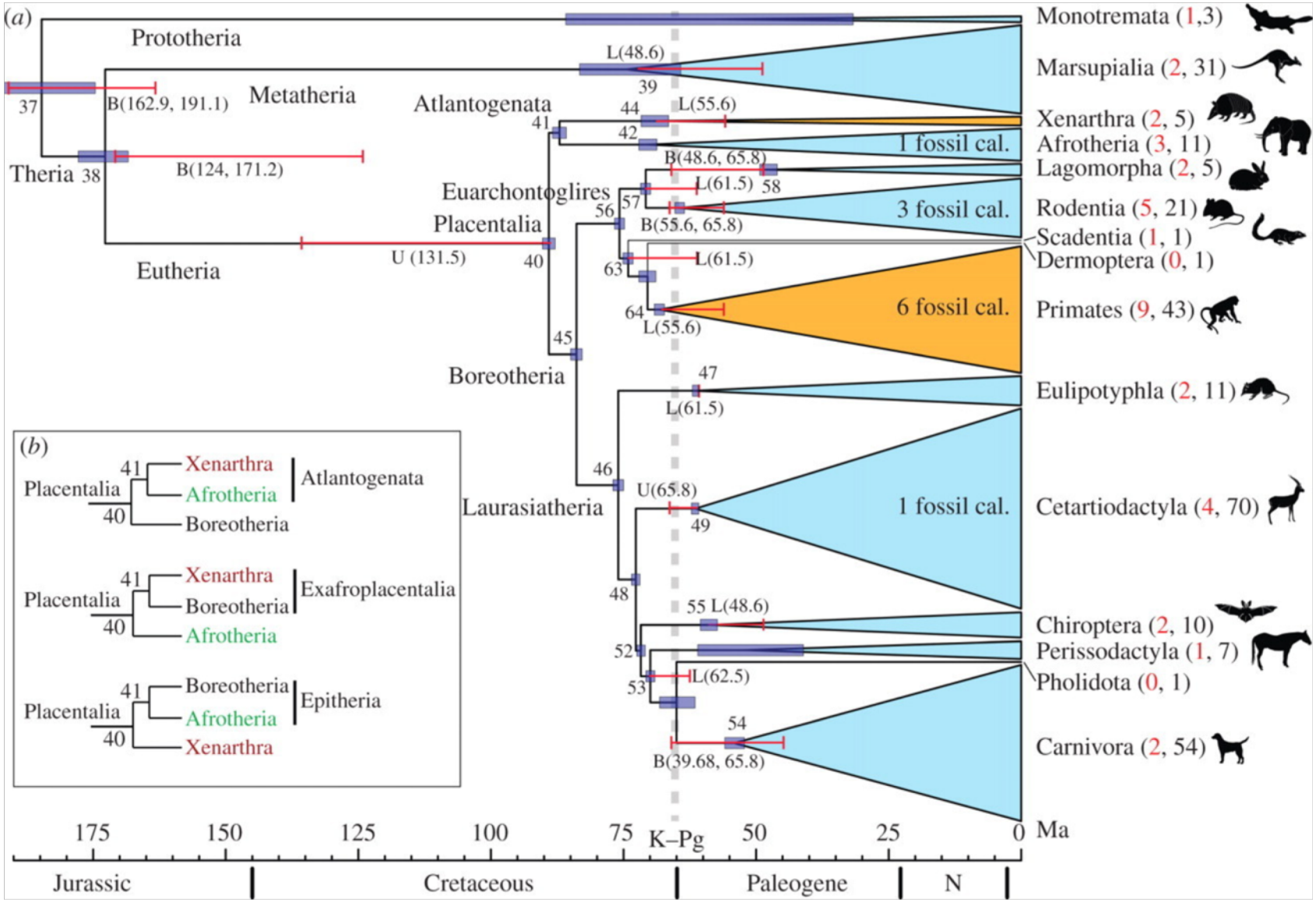
Associate editor: Oliver Pybus

Mol. Biol. Evol. 28(7):2161–2172. 2011

Example: Mammal Divergences

- dos Reis et al. (1998) Proc. R. Soc. B, 279: 3491
- Analysed 36 mammal genomes
- Alignment is 21 million sites
- MCMC approximate likelihood analysis time: 15 days
- Exact likelihood (not done): over one year
- We will use MCMCTree's approximate likelihood method in the practical with a Primates example

Example: Mammal Divergences



To Learn More …

- Holder & Lewis (2003) **Phylogeny estimation: Traditional and Bayesian approaches**. Nat. Rev. Genet., 4: 275

- Chen, Kuo & Lewis (2014) **Bayesian phylogenetics: Methods, algorithms, and applications**. CRC Press

- Yang (2014) **Molecular evolution: A statistical approach**. Oxford University Press

- Nascimento, dos Reis & Yang (2017) **A biologist's guide to Bayesian phylogenetic analysis**. Nat. Ecol. Evol., 1: 1446

- **THE END**