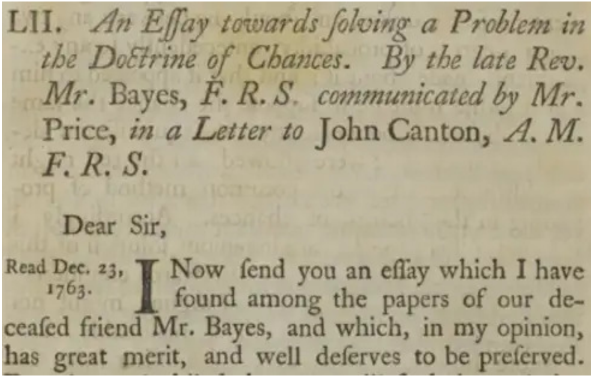
Introduction to Bayesian statistics and Markov Chain Monte Carlo

- Brief review of probability concepts
- Bayes theorem
- Bayesian inference (thought experiment)
- Introduction to Markov Chain Monte Carlo (without the theory)
- Understanding MCMC output

Rev. Mr. Thomas Bayes, FRS 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.*

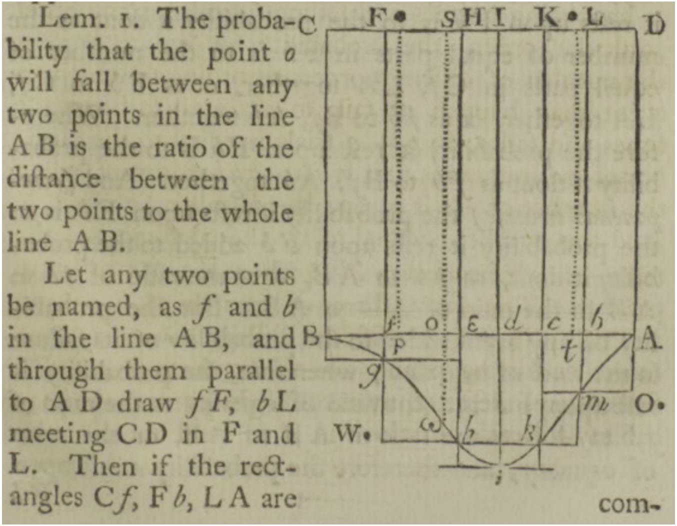
Dear Sir,

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, has great merit, and well deſerves to be preſerved.

Bayes, T. and Price, R. (1763). An Essay towards Solving a Problem in the Doctrine of Chances. Philosophical Transactions of the Royal Society of London, 53, 370-418.

Rev. Mr. Thomas Bayes, FRS paper



Lem. 1. The probability that the point *o* will fall between any two points in the line A B is the ratio of the diſtance between the two points to the whole line A B.

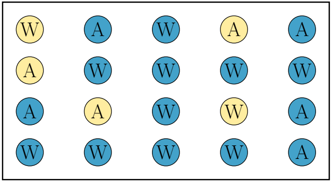
Let any two points be named, as *f* and *b* in the line A B, and through them parallel to A D draw *f* F, *b* L meeting C D in F and L. Then if the rectangles C *f*, F *b*, L A are com-

C
F
S
H
I
K
L
D
B
A
f
o
e
d
c
b
p
t
g
m
O
W
ω
h
k
i

Marbles in a box

- Suppose there are **twenty** marbles inside a box:

		Material		
		Wooden	**Plastic**	
Color	**Yellow**	2	3	
	Blue	10	5	
				20



Marbles in a box

- Suppose there are **twenty** marbles inside a box:

		Material		
		Wooden	**Plastic**	
Color	**Yellow**	2	3	
	Blue	10	5	
				20

- We take one marble randomly out of the box

Marbles in a box

- Suppose there are **twenty** marbles inside a box:

		Material		
		Wooden	**Plastic**	
Color	**Yellow**	2	3	
	Blue	10	5	
				20

- We take one marble randomly out of the box
- What is the probability that it is **yellow** and made of **wood**?
- $P(\mathbf{C} = Y, \mathbf{M} = W) = P(Y, W) = ?$

Marbles in a box

- Suppose there are **twenty** marbles inside a box:

		Material		
		Wooden	**Plastic**	
Color	**Yellow**	2	3	
	Blue	10	5	
				20

- $P(Y,W) = 2/20 = 0.1$ or 10%

Marbles in a box

- Suppose there are **twenty** marbles inside a box:

	Material			
		Wooden	**Plastic**	
Color	**Yellow**	2	3	
	Blue	10	5	
				20

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

Marbles in a box

- Suppose there are **twenty** marbles inside a box:

		Material		
		Wooden	**Plastic**	
Color	**Yellow**	2	3	
	Blue	10	5	
				20

- We place the marble back in the box, shuffle and take out another marble

Marbles in a box

- Suppose there are **twenty** marbles inside a box:

Color	Material:	**Wooden**	**Plastic**	
	Yellow	2	3	
	Blue	10	5	
				20

- We place the marble back in the box, shuffle and take out another marble
- What is the probability that it is blue?
- $P(B) =?$

Marbles in a box

- Suppose there are **twenty** marbles inside a box:

Material

Color

	Wooden	**Plastic**	table margin
Yellow	2	3	5
Blue	10	5	15
table margin	12	8	20

- $P(B) = 15/20 = 0.75$

Marbles in a box

- Suppose there are **twenty** marbles inside a box:

Material

table margin

Color		**Wooden**	**Plastic**	
	Yellow	2	3	5
	Blue	10	5	15
	table margin	12	8	20

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

Marbles in a box

- Suppose there are **twenty** marbles inside a box:

Color	Material:	**Wooden**	**Plastic**	
	Yellow	2	3	5
	Blue	10	5	15
		12	8	20

Note that:

- $P(B) = 10/20 + 5/20 = 0.75$

Marbles in a box

- Suppose there are **twenty** marbles inside a box:

		Material		
		Wooden	**Plastic**	
Color	**Yellow**	2	3	5
	Blue	10	5	15
		12	8	20

Note that:

- $P(B) = 10/20 + 5/20 = 0.75$ **or**
- $P(B) = P(B, W) + P(B, P)$
- The **marginal** is the sum over the **joints**

Marbles in a box

- Suppose there are **twenty** marbles inside a box:

		Material		
		Wooden	**Plastic**	
Color	**Yellow**	2	3	5
	Blue	10	5	15
		12	8	20

- Suppose we took out a **blue** marble, what is the probability that it is **wooden**?

- $P(W \mid B) = ?$

Marbles in a box

- Suppose there are **twenty** marbles inside a box:

Material

Color	Wooden	Plastic	
Yellow	2	3	5
Blue	10	5	15
	12	8	20

- $P(W \mid B) = 10/15 = 0.667$

Marbles in a box

- Suppose there are **twenty** marbles inside a box:

Material

Color	Wooden	Plastic	
Yellow	2	3	5
Blue	10	5	15
	12	8	20

- $P(W \mid B) = 10/15 = 0.667$
- $P(W \mid B)$ is the **conditional probability** of W given B

Marbles in a box

- Suppose there are **twenty** marbles inside a box:

		Material		
		Wooden	**Plastic**	
Color	**Yellow**	2	3	5
	Blue	10	5	15
		12	8	20

- $P(W \mid B)$ vs $P(W, B)$

Marbles in a box

- Suppose there are **twenty** marbles inside a box:

		Material		
		Wooden	**Plastic**	
Color	**Yellow**	2	3	5
	Blue	10	5	15
		12	8	20

  - $P(W \mid B)$ vs $P(W, B)$
  - **Conditional:** we have information. One variable is not random

Marbles in a box

- Suppose there are **twenty** marbles inside a box:

		Material		
		Wooden	**Plastic**	
Color	**Yellow**	2	3	5
	Blue	10	5	15
		12	8	20

  - $P(W \mid B)$ vs $P(W, B)$
  - **Conditional:** we have information. One variable is not random
  - **Joint:** both are random

Marbles in a box

- Suppose there are **twenty** marbles inside a box:

		Material		
		Wooden	**Plastic**	
Color	**Yellow**	2	3	5
	Blue	10	5	15
		12	8	20

- Note that:
- $P(W \mid B) = 10/15$

Marbles in a box

- Suppose there are **twenty** marbles inside a box:

Material

Color		Wooden	Plastic	
	Yellow	2	3	5
	Blue	10	5	15
		12	8	20

- Note that:
- $P(W \mid B) = 10/15$
- $P(W \mid B) = (10/20)/(15/20) = 0.667$ or

Marbles in a box

- Suppose there are **twenty** marbles inside a box:

Color \ Material	Wooden	Plastic	
Yellow	2	3	5
Blue	10	5	15
	12	8	20

- Note that:
- $P(W \mid B) = 10/15$
- $P(W \mid B) = (10/20)/(15/20) = 0.667$ or
- $P(W \mid B) = P(W, B)/P(B)$
- The **conditional** is the **joint** over the **marginal**

Marbles in a box

- Suppose there are **twenty** marbles inside a box:

		Material		
		Wooden	**Plastic**	
Color	**Yellow**	2	3	5
	Blue	10	5	15
		12	8	20

- Note that:
- $P(W \mid B) = 10/15$
- $P(W \mid B) = (10/20)/(15/20) = 0.667$ or
- $P(W \mid B) = P(W, B)/P(B)$
- The **conditional** is the **joint** over the **marginal**

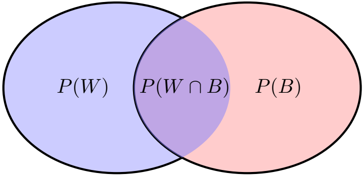
Marbles in a box

- Suppose there are **twenty** marbles inside a box:

Color	Wooden	Plastic	
Yellow	2	3	5
Blue	10	5	15
	12	8	20

Material

- Note that:
- $P(W \mid B) = 10/15$
- $P(W \mid B) = (10/20)/(15/20) = 0.667$ or
- $P(W \mid B) = P(W, B)/P(B)$
- The **conditional** is the **joint** over the **marginal**



Marbles in a box

- Suppose there are **twenty** marbles inside a box:

		Material		
		Wooden	**Plastic**	
Color	**Yellow**	2	3	5
	Blue	10	5	15
		12	8	20

- Note we can reverse the conditional:

Marbles in a box

- Suppose there are **twenty** marbles inside a box:

		Material		
		Wooden	**Plastic**	
Color	**Yellow**	2	3	5
	Blue	10	5	15
		12	8	20

- Note we can reverse the conditional:
- $P(B \mid W) = P(B, W)/P(W)$

Marbles in a box

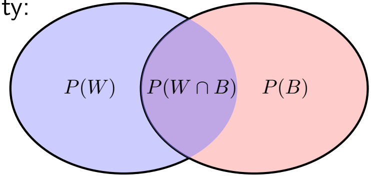
- Suppose there are **twenty** marbles inside a box:

		Material		
		Wooden	**Plastic**	
Color	**Yellow**	2	3	5
	Blue	10	5	15
		12	8	20

- Note we can reverse the conditional:
- $P(B \mid W) = P(B, W)/P(W)$
- $P(B \mid W) = (10/20)/(12/20) = 0.833$

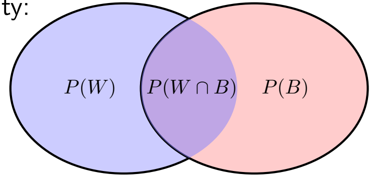
Bayes Theorem

- From the definition of conditional probability:
- $P(B \mid W) = P(B, W)/P(W)$
- $P(W \mid B) = P(B, W)/P(B)$



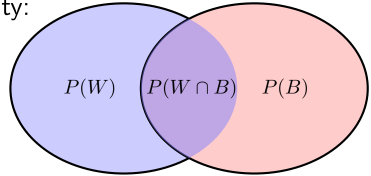
Bayes Theorem

- From the definition of conditional probability:
- $P(B \mid W) = P(B, W)/P(W)$
- $P(W \mid B) = P(B, W)/P(B)$



- We obtain:
- $P(B, W) = P(W) \times P(B \mid W)$
- $P(B, W) = P(B) \times P(W \mid B)$
- $P(W) \times P(B \mid W) = P(B) \times P(W \mid B)$

Bayes Theorem



- From the definition of conditional probability:
- $P(B \mid W) = P(B, W)/P(W)$
- $P(W \mid B) = P(B, W)/P(B)$

- We obtain:
- $P(B, W) = P(W) \times P(B \mid W)$
- $P(B, W) = P(B) \times P(W \mid B)$
- $P(W) \times P(B \mid W) = P(B) \times P(W \mid B)$

- Therefore:

$$P(B \mid W) = \frac{P(B) \times P(W \mid B)}{P(W)}$$

- This is known as the **Bayes theorem**

Marginal Probability

- $P(W) = P(Y, W) + P(B, W)$
- $P(W) = P(W \mid Y)P(Y) + P(W \mid B)P(B)$

Marginal Probability

- $P(W) = P(Y, W) + P(B, W)$
- $P(W) = P(W \mid Y)P(Y) + P(W \mid B)P(B)$

- Suppose there are marbles of n differnet colours in the box, then::
- $P(W) = P(W \mid C_1)P(C_1) + \ldots + P(W \mid C_n)$
- $P(W) = \sum_i^n P(W \mid C_i)P(C_i)$

Marginal Probability

- $P(W) = P(Y, W) + P(B, W)$
- $P(W) = P(W \mid Y)P(Y) + P(W \mid B)P(B)$

- Suppose there are marbles of n differnet colours in the box, then::
- $P(W) = P(W \mid C_1)P(C_1) + \ldots + P(W \mid C_n)$
- $P(W) = \sum_i^n P(W \mid C_i)P(C_i)$

- $P(W) = \int P(W \mid X)P(X)dX$ if X is continuous

Rev. Mr. Bayes thought experiment (modified)

- We have a **billiard table** (or flat plane)
- A **white** ball ○ is thrown onto the table at an **unknown position** (x, y)
- The position (x, y) is **unknown** to us (not revealed)
- A second **black** ball ● is thrown randomly onto the table and we are told if:
  - The ball lands to the left or right of the unknown position x
  - The ball lands to the front or behind of the unknown position y
- After n throws of the **black** ball, can we guess the position of the **white** ball?

Rev. Mr. Bayes thought experiment (modified)

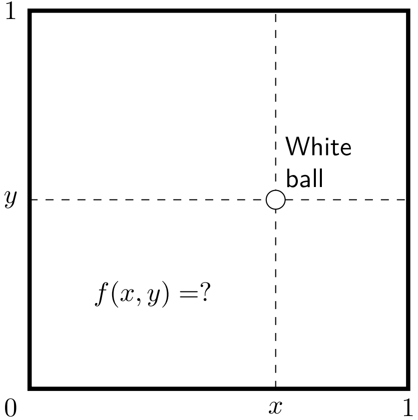
- Thomas Bayes showed how to **estimate the probability of the white ball's location based on observed data (inverse probability problem)**

- He further showed that with sufficient throws (data), we would eventually become **almost certain** of the white balls's position

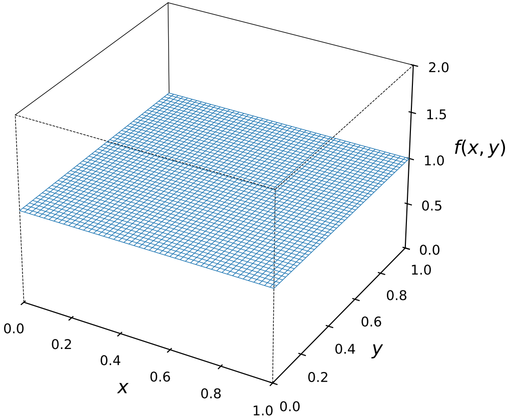
- We will go through the example

Rev. Mr. Bayes thought experiment (modified)

- We throw the white ball

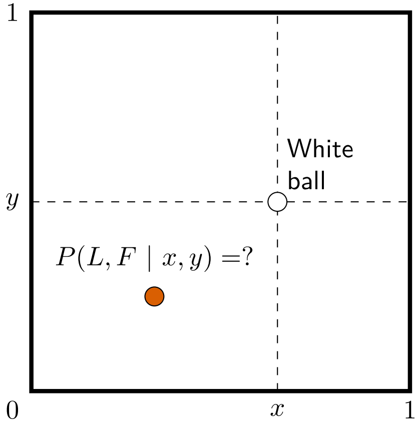


Rev. Mr. Bayes thought experiment (modified)



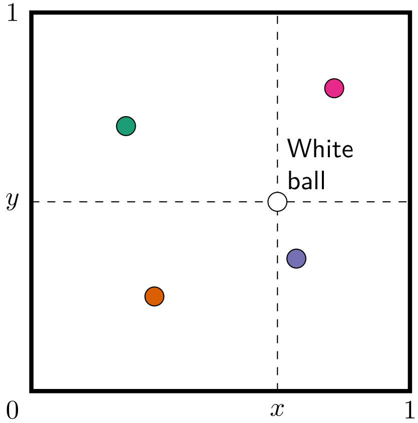
We assume a uniform distribution over x and y: $f(x,y) = 1$

Rev. Mr. Bayes thought experiment (modified)



- L: Left
- R: Right
- F: Front
- B: Back

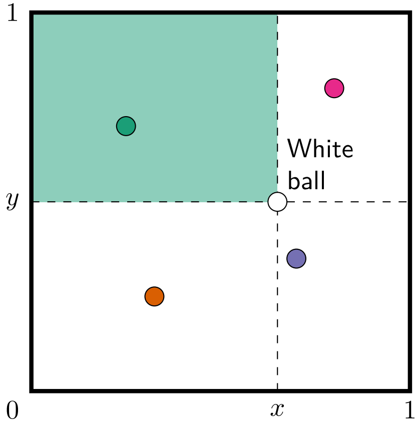
Rev. Mr. Bayes thought experiment (modified)



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

- $P(L, F \mid x, y) = xy$
- $P(L, B \mid x, y) = x(1 - y)$
- $P(R, F \mid x, y) = (1 - x)y$
- $P(R, B \mid x, y) = (1 - x)(1 - y)$

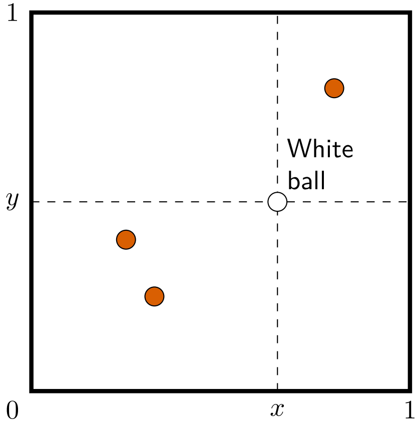
Rev. Mr. Bayes thought experiment (modified)



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

- $P(L, F \mid x, y) = xy$
- $P(L, B \mid x, y) = x(1 - y)$
- $P(R, F \mid x, y) = (1 - x)y$
- $P(R, B \mid x, y) = (1 - x)(1 - y)$

Rev. Mr. Bayes thought experiment (modified)



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

Data (ball throws):

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

Probability of observing data given (x, y):

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

Rev. Mr. Bayes thought experiment (modified)

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

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

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

- $P(D \mid x,y) = x^a(1-x)^{(n-a)}y^b(1-y)^{(n-b)}$
- $|D| = n$: number of throws (size of data)
- a: number of left landings
- b: number of front landings

Rev. Mr. Bayes thought experiment (modified)

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

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

Therefore, we can now define the **joint density** of D, x, y:

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

Rev. Mr. Bayes thought experiment (modified)

According to the **Bayes theorem:**

$$f(x, y \mid D) = \frac{f(x, y)P(D \mid x, y)}{P(D)}$$

The difficulty is in calculating the **marginal probability** $P(D)$

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:

$$\begin{aligned} P(D) &= \iint f(D, x, y)\, dx\, dy \\ &= \frac{a!(n-a)!b!(n-b)!}{((n+1)!)^2} \end{aligned}$$

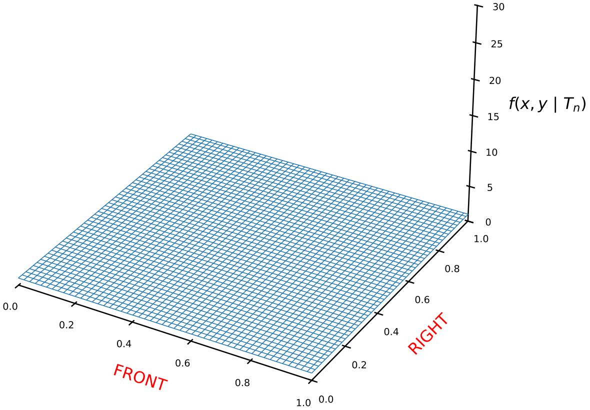
Rev. Mr. Bayes thought experiment (simulation)

Simulation for the modified Bayes thought experiment:

1. Sample x and y from the join uniform $f(x, y)$. This is the position of the **white ball**
2. Initilize $a = b = n = 0$
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 to 5
7. Calculate $f(x, y \mid D)$. **Note:** Our data is $D = \{a, b, n\}$

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



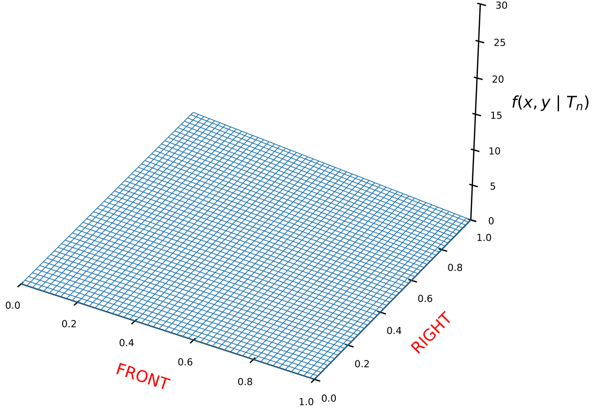
a:	0	# left
b:	0	# fronts
n:	0	# throws

No data

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



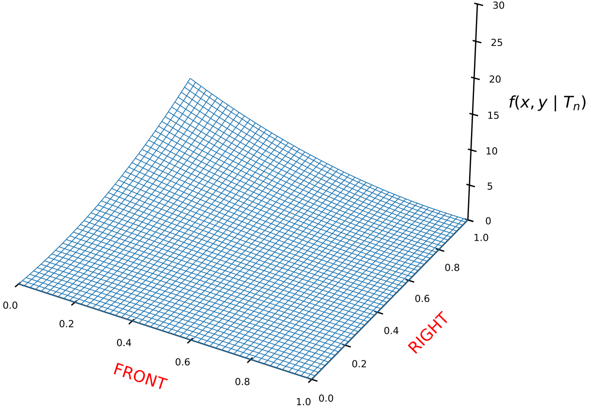
a:	0	# left
b:	1	# fronts
n:	1	# throws

front,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



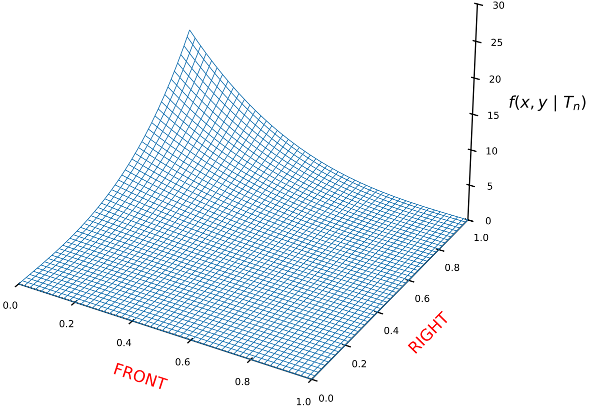
a:	0	# left
b:	2	# fronts
n:	2	# throws

front,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



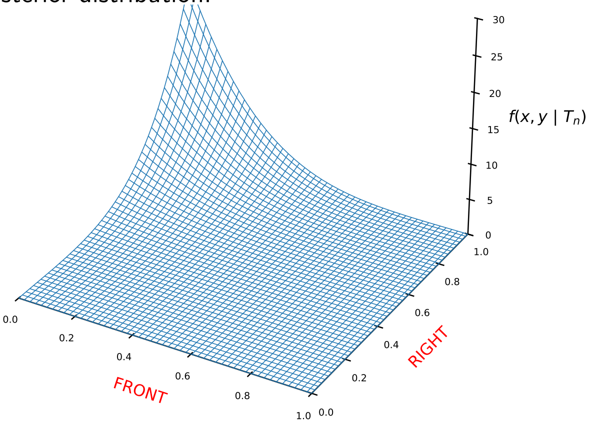
a:	0	# left
b:	3	# fronts
n:	3	# throws

front,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



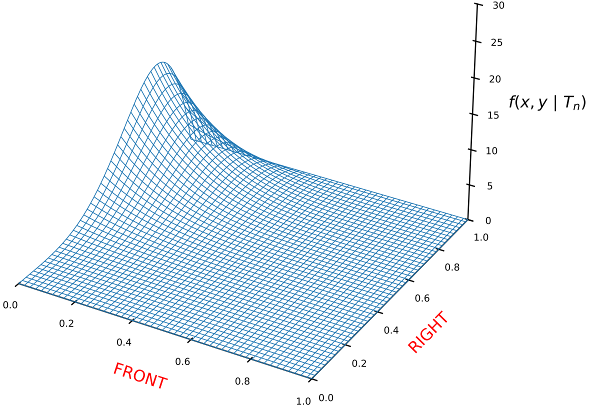
a:	0	# left
b:	4	# fronts
n:	4	# throws

front,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



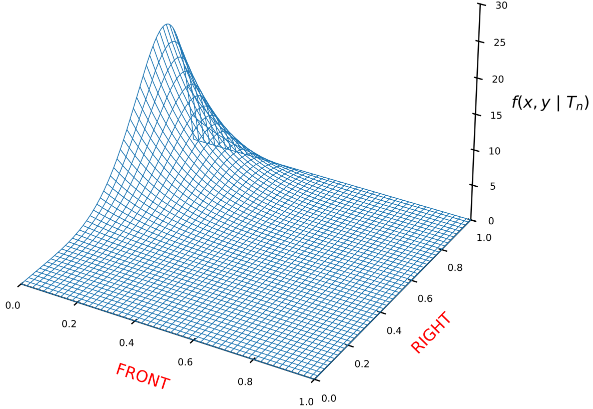
a:	0	# left
b:	4	# fronts
n:	5	# throws

back,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



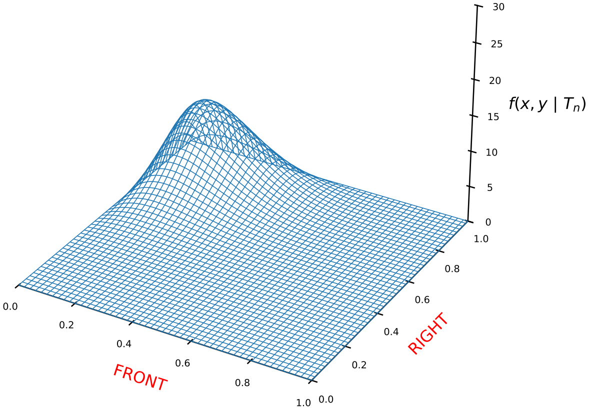
a:	0	# left
b:	5	# fronts
n:	6	# throws

front,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



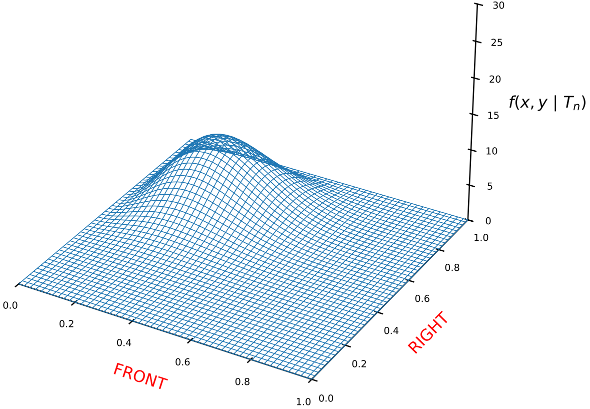
a:	1	# left
b:	6	# fronts
n:	7	# throws

front,left

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



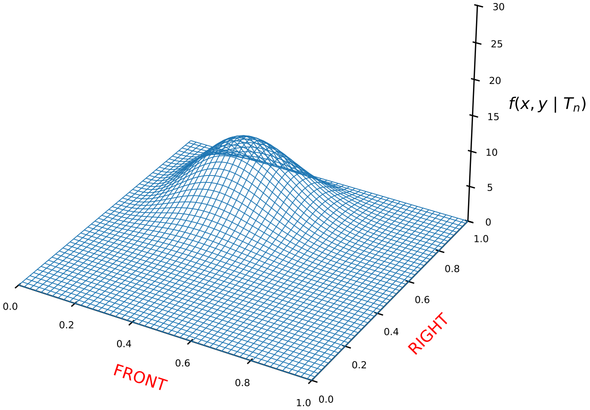
a:	2	# left
b:	6	# fronts
n:	8	# throws

back,left

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



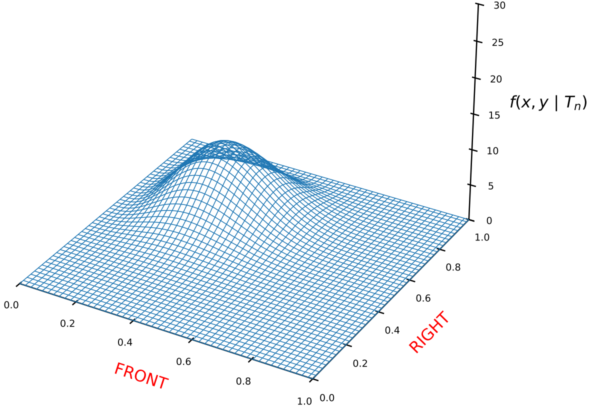
a:	3	# left
b:	7	# fronts
n:	9	# throws

front,left

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



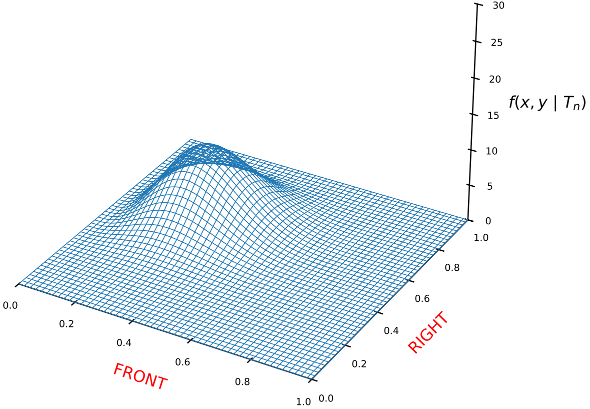
a:	3	# left
b:	7	# fronts
n:	10	# throws

back,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



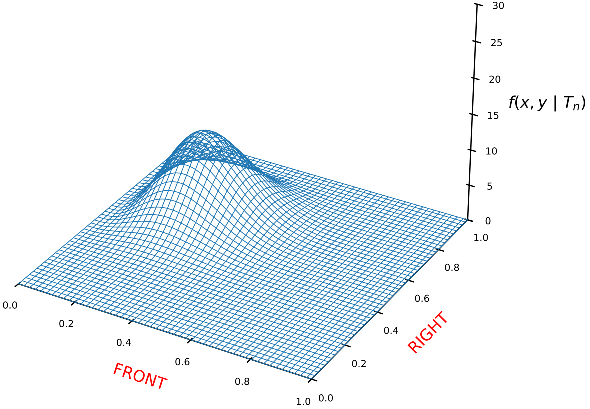
a:	3	# left
b:	7	# fronts
n:	11	# throws

back,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



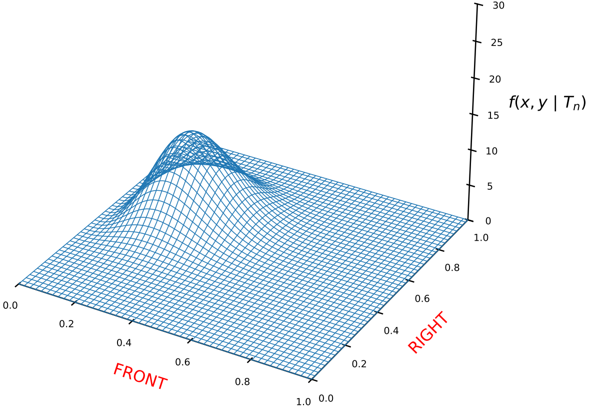
a:	3	# left
b:	8	# fronts
n:	12	# throws

front,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



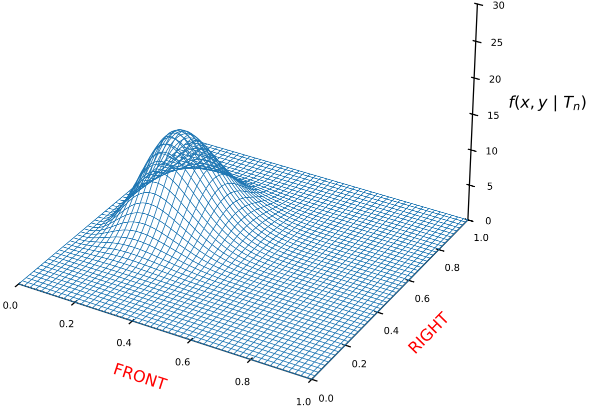
a:	3	# left
b:	8	# fronts
n:	13	# throws

back,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



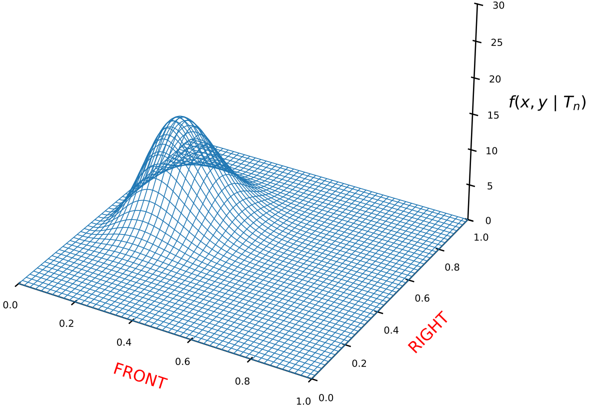
a:	3	# left
b:	8	# fronts
n:	14	# throws

back,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



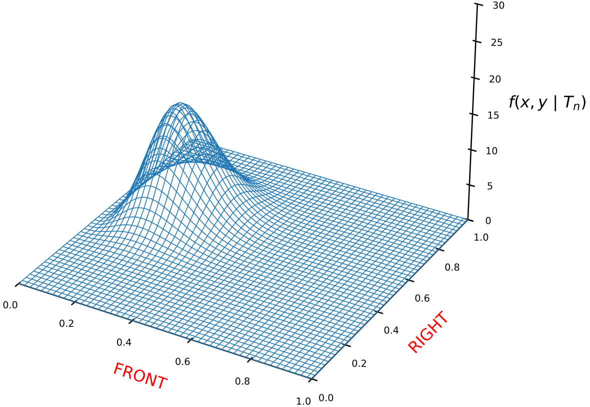
a:	3	# left
b:	9	# fronts
n:	15	# throws

front,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



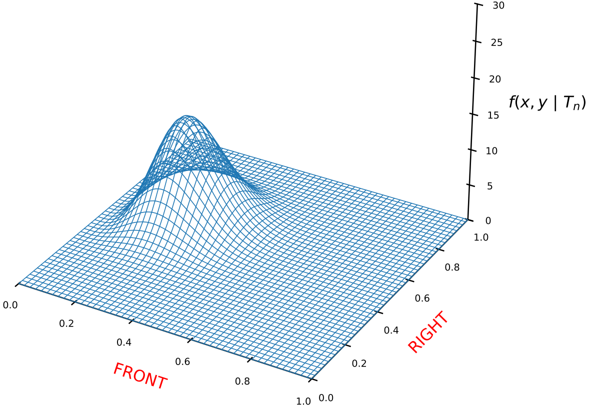
a:	3	# left
b:	10	# fronts
n:	16	# throws

front,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



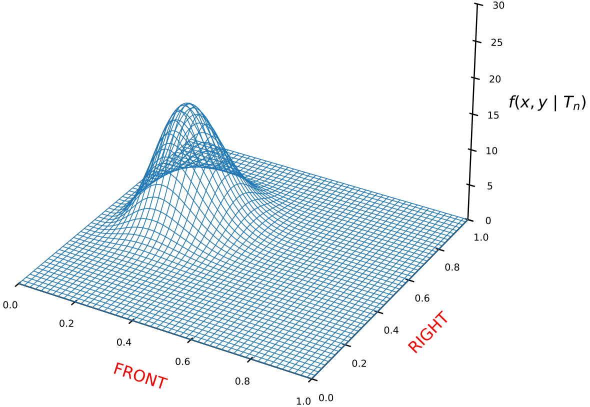
a:	4	# left
b:	10	# fronts
n:	17	# throws

back,left

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



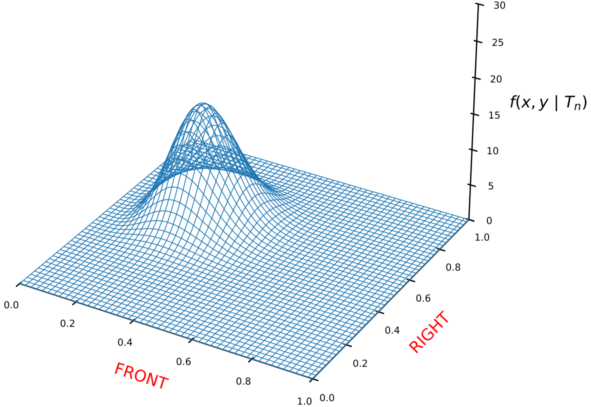
a:	4	# left
b:	11	# fronts
n:	18	# throws

front,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



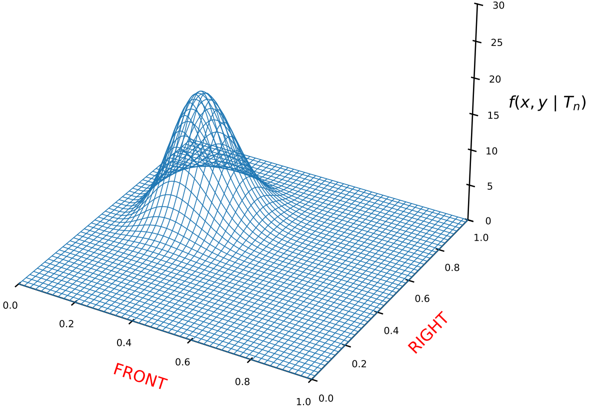
a:	5	# left
b:	12	# fronts
n:	19	# throws

front,left

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



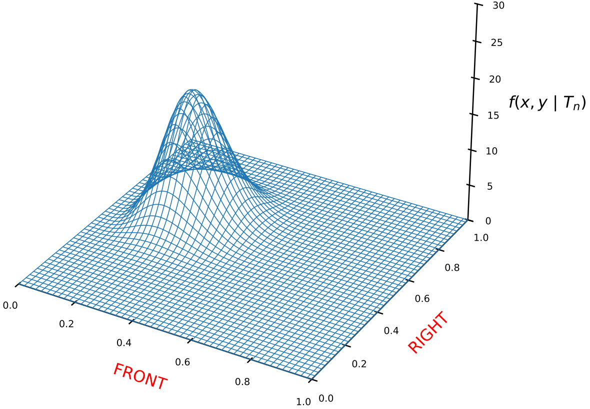
a:	5	# left
b:	13	# fronts
n:	20	# throws

front,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



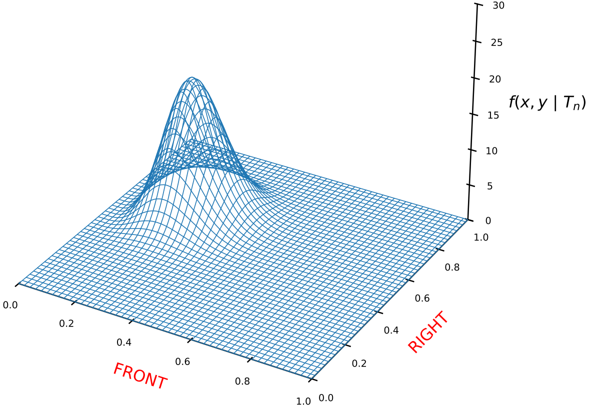
a:	5	# left
b:	13	# fronts
n:	21	# throws

back,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



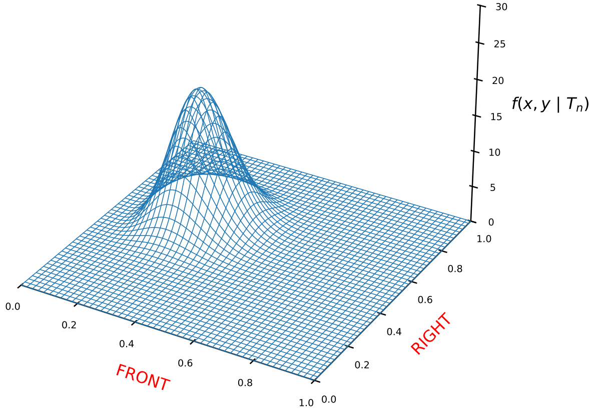
a:	5	# left
b:	14	# fronts
n:	22	# throws

front,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



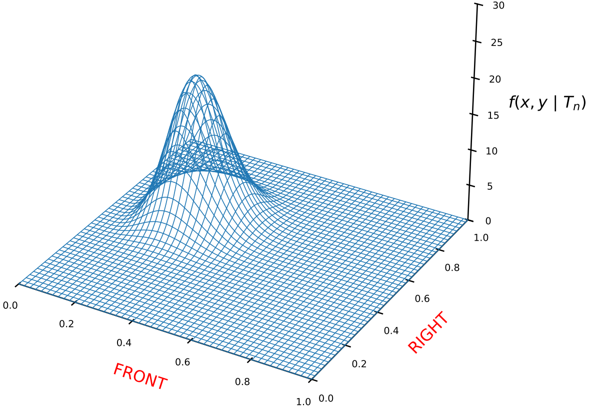
a:	6	# left
b:	14	# fronts
n:	23	# throws

back,left

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



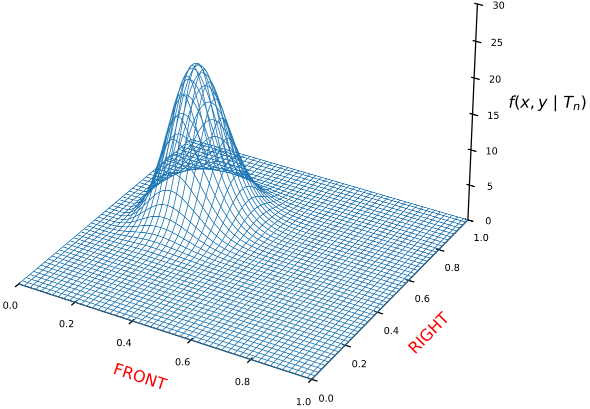
a:	6	# left
b:	15	# fronts
n:	24	# throws

front,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



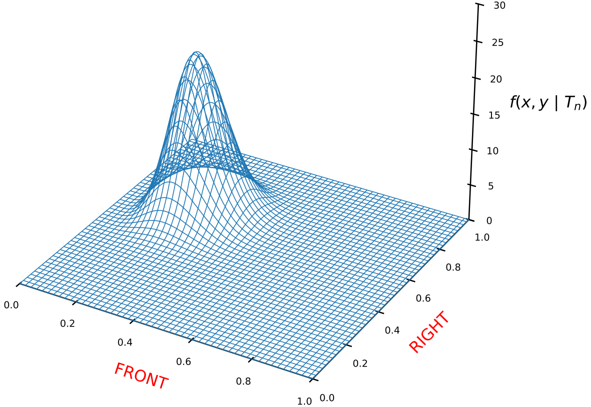
a:	6	# left
b:	16	# fronts
n:	25	# throws

front,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



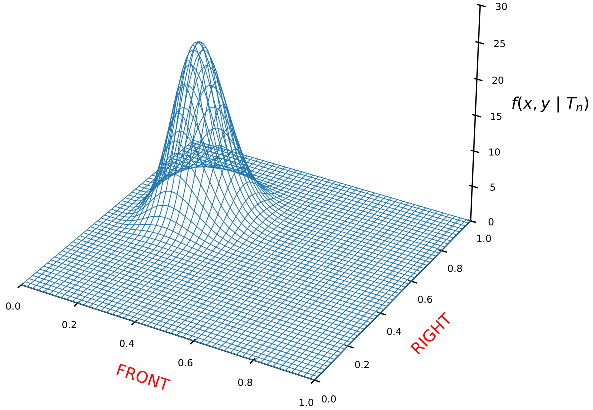
a:	6	# left
b:	17	# fronts
n:	26	# throws

front,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



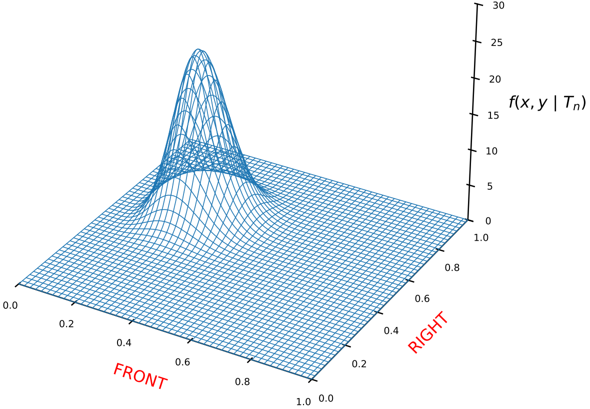
a:	6	# left
b:	18	# fronts
n:	27	# throws

front,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



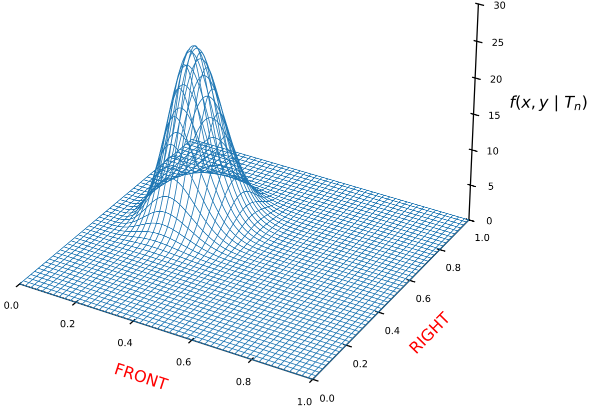
a:	7	# left
b:	18	# fronts
n:	28	# throws

back,left

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:



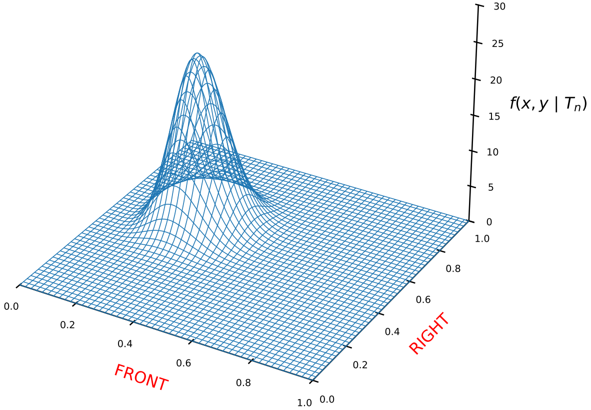
a:	7	# left
b:	18	# fronts
n:	29	# throws

back,right

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

Rev. Mr. Bayes thought experiment (simulation)

Posterior distribution:

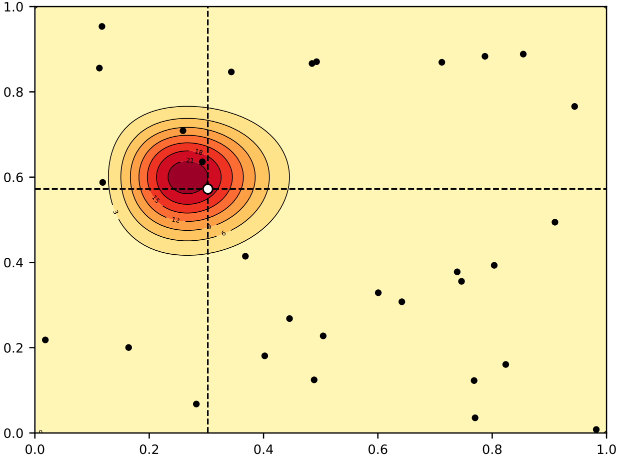


a:	8	# left
b:	18	# fronts
n:	30	# throws

back,left

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

Rev. Mr. Bayes thought experiment (simulation)



Bayesian Terminology

$$f(x, y \mid D) = \frac{f(x, y) f(D \mid x, y)}{P(D)}$$

- The marginal of x and y, $f(x, y)$, is known as the **prior distribution** of x and y
- The prior $f(x, y)$ reflects our **prior knowledge** about x and y before any data has been observed

Bayesian Terminology

$$f(x, y \mid D) = \frac{f(x, y) f(D \mid x, y)}{P(D)}$$

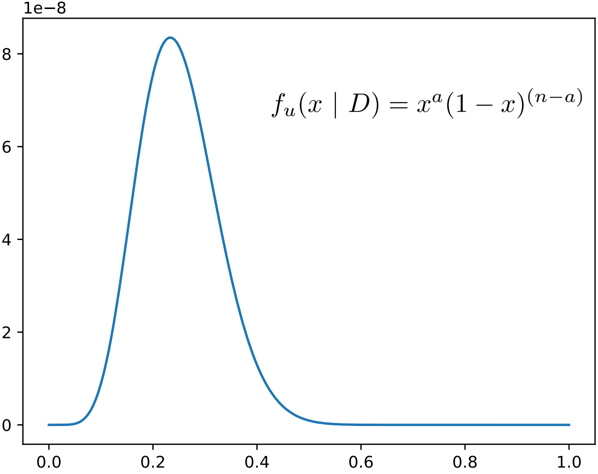
- The marginal of x and y, $f(x, y)$, is known as the **prior distribution** of x and y
- The prior $f(x, y)$ reflects our **prior knowledge** about x and y before any data has been observed
- The conditional $f(D \mid x, y)$ is known as the **likelihood** of the data D
- $P(D)$ is known as the **marginal likelihood**

Bayesian Terminology

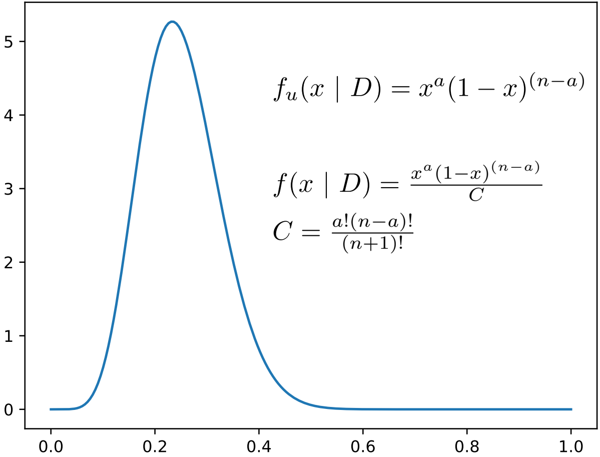
$$f(x, y \mid D) = \frac{f(x, y) f(D \mid x, y)}{P(D)}$$

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

Marginal likelihood



Marginal likelihood



$$f_u(x \mid D) = x^a(1-x)^{(n-a)}$$

$$f(x \mid D) = \frac{x^a(1-x)^{(n-a)}}{C}$$

$$C = \frac{a!(n-a)!}{(n+1)!}$$

f_u: unnormalised density — has the same shape as the normalised density f

Marginal likelihood

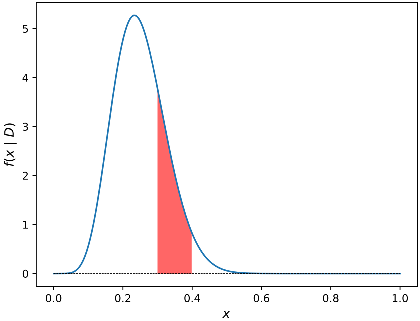
- Can we ignore the marginal likelihood P(T)?

Marginal likelihood

- Can we ignore the marginal likelihood P(T)?
- No.

Marginal likelihood

- Can we ignore the marginal likelihood P(T)?
- No.
- The density must be normalised because the probability is the area under the curve:
- $P(a < x < b) = \int_a^b f(x \mid D)\, dx$



Note:

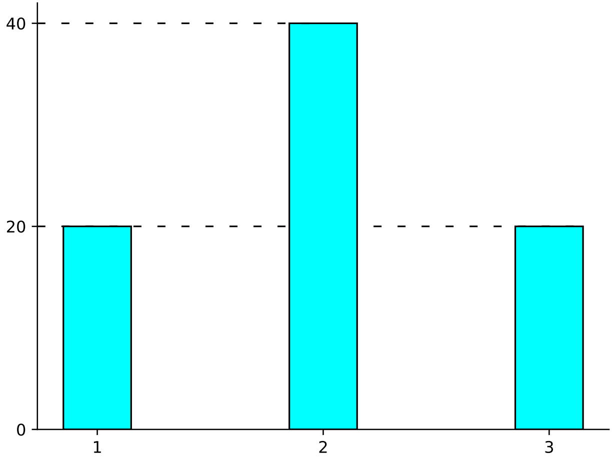
- $P(0 \leq x \leq 1) = 1$
- For multi-dimensional densities, the probability is the volume under the surface

General Bayesian Model

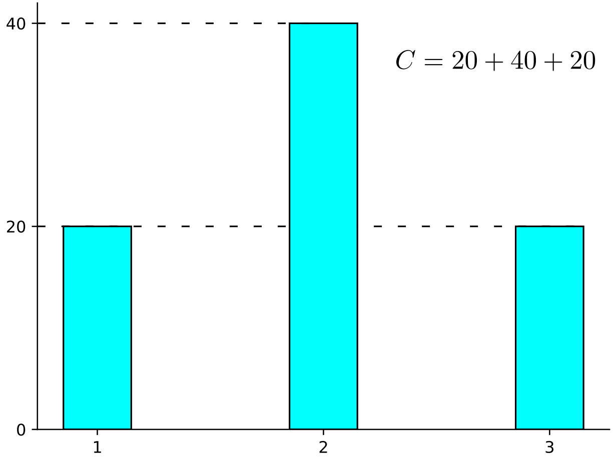
$$\boxed{\overbrace{f(\boldsymbol{\theta} \mid \boldsymbol{D})}^{\text{Posterior}}} = \boxed{\overbrace{f(\boldsymbol{\theta})}^{\text{Prior}}}\,\boxed{\overbrace{f(\boldsymbol{D} \mid \boldsymbol{\theta})}^{\text{Likelihood}}} \Big/ \boxed{\overbrace{f(\boldsymbol{D})}^{\text{Marginal}}}$$

- D is the **data**
- $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ is the set of **model parameters**
- $f(D) = \int f(\boldsymbol{\theta}) f(D \mid \boldsymbol{\theta})\, d\boldsymbol{\theta}$ is the **marginal likelihood**
- $f(D)$ is an n-dimensional integral
- Usually, this integral **does not have an analytical solution** or is hard to calculate
- What do we do?

Sampling from histograms

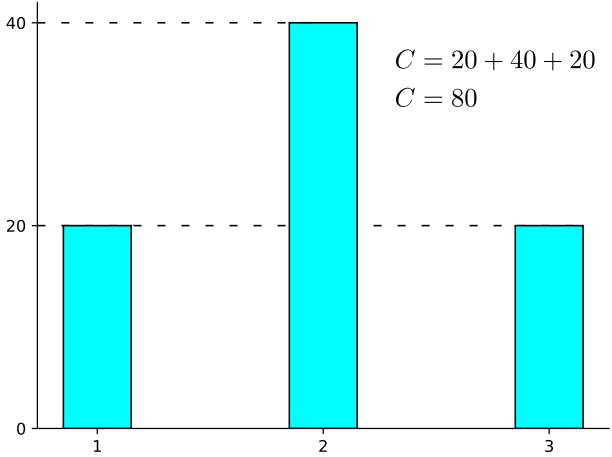


Sampling from histograms



$C = 20 + 40 + 20$

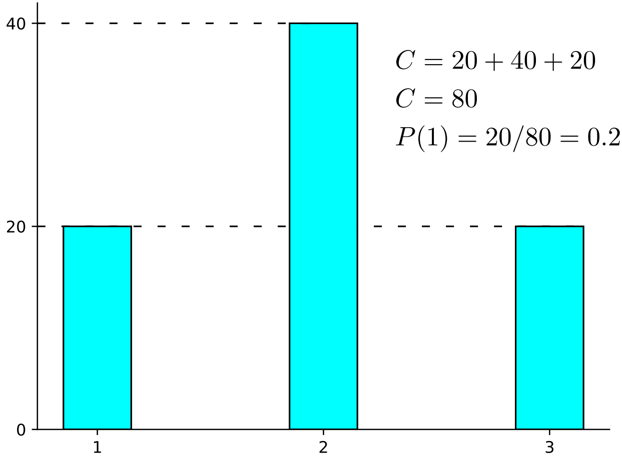
Sampling from histograms



$$C = 20 + 40 + 20$$

$$C = 80$$

Sampling from histograms

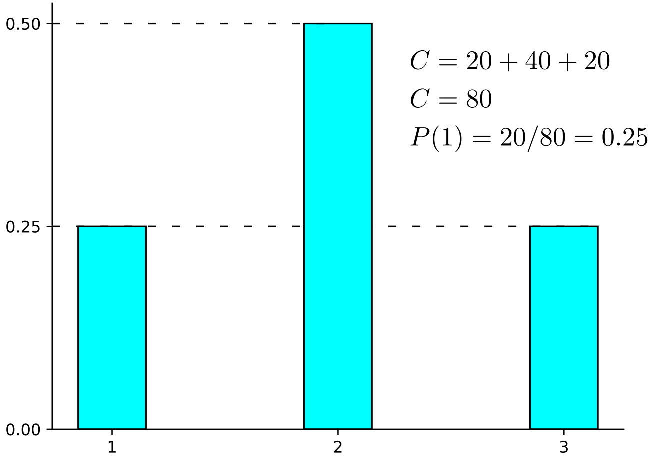


$$C = 20 + 40 + 20$$

$$C = 80$$

$$P(1) = 20/80 = 0.25$$

Sampling from histograms

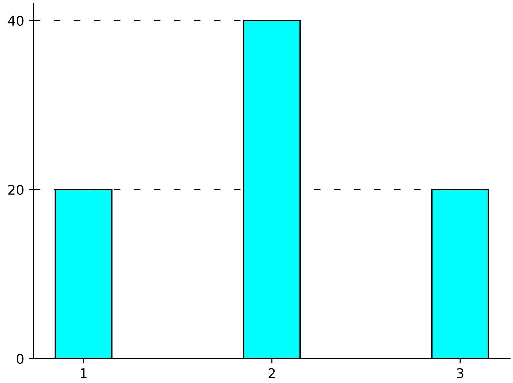


$$C = 20 + 40 + 20$$

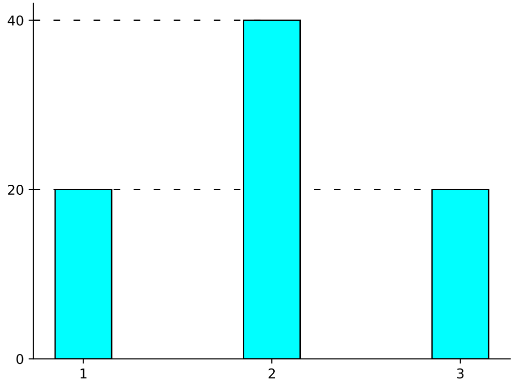
$$C = 80$$

$$P(1) = 20/80 = 0.25$$

Sampling from histograms

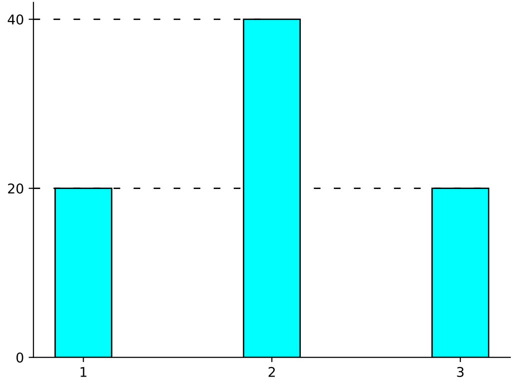


Sampling from histograms



Algorithm:

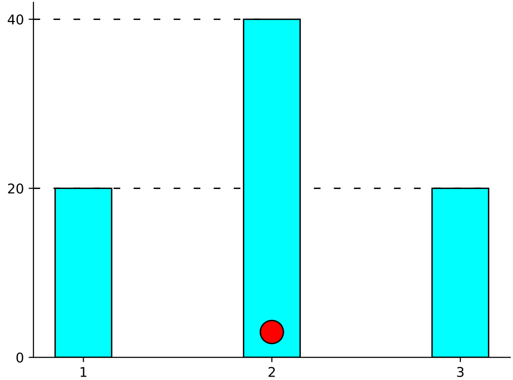
Sampling from histograms



Algorithm:

1. Select a starting point (x)

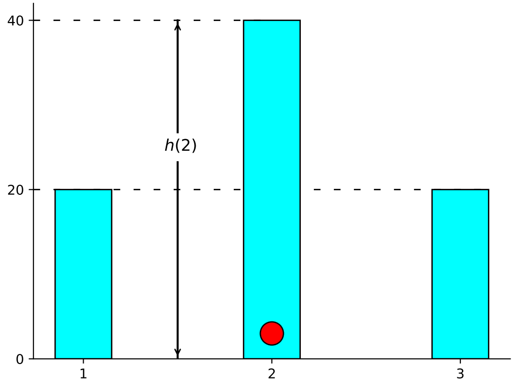
Sampling from histograms



Algorithm:

1. Select a starting point (x)

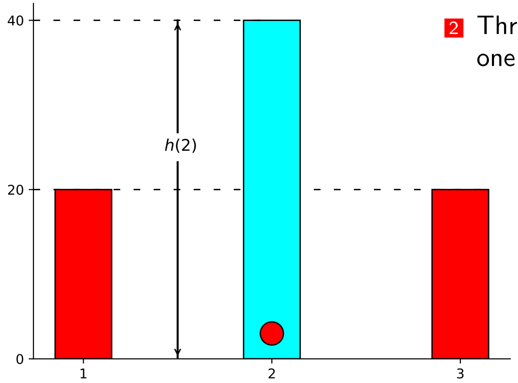
Sampling from histograms



Algorithm:

1. Select a starting point (x)

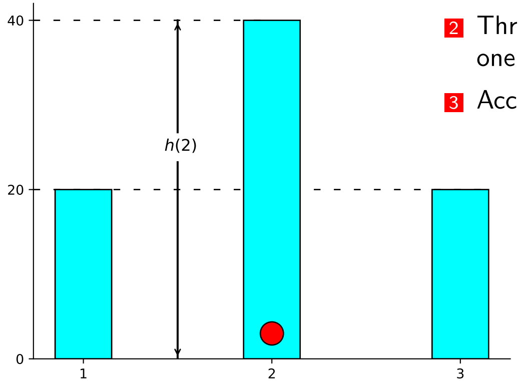
Sampling from histograms



Algorithm:

1. Select a starting point (x)
2. Throw a coin to propose a visit to one of the adjacent bars (x^*)

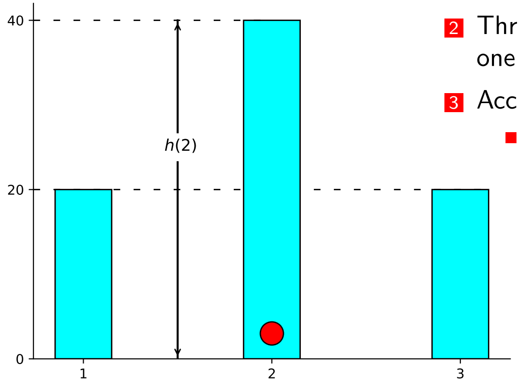
Sampling from histograms



Algorithm:

1. Select a starting point (x)
2. Throw a coin to propose a visit to one of the adjacent bars (x^*)
3. Accept or reject the visit:

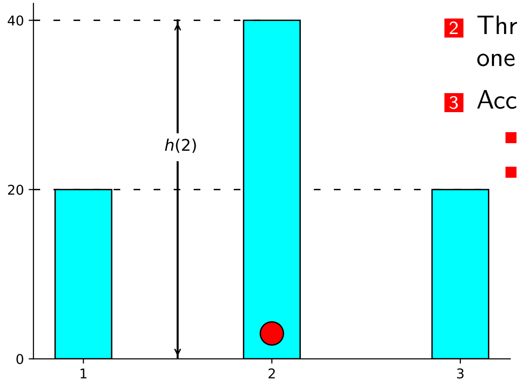
Sampling from histograms



Algorithm:

1. Select a starting point (x)
2. Throw a coin to propose a visit to one of the adjacent bars (x^*)
3. Accept or reject the visit:
    - If $h(n) > h(c)$ then **accept**

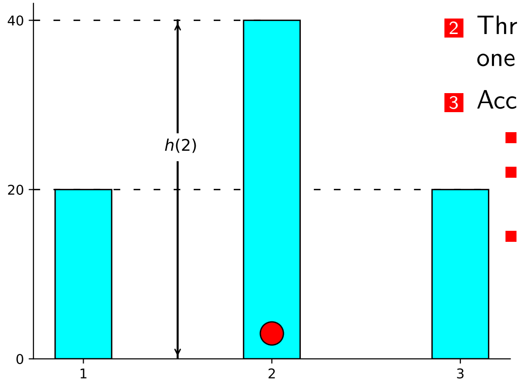
Sampling from histograms



Algorithm:

1. Select a starting point (x)
2. Throw a coin to propose a visit to one of the adjacent bars (x^*)
3. Accept or reject the visit:
    - If $h(n) > h(c)$ then **accept**
    - Otherwise, accept with $\mathcal{A} = h(x^*)/h(x)$

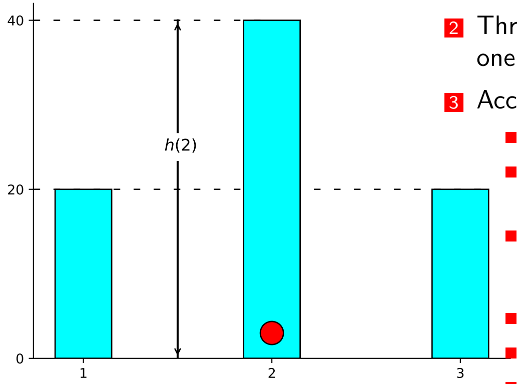
Sampling from histograms



Algorithm:

1. Select a starting point (x)
2. Throw a coin to propose a visit to one of the adjacent bars (x^*)
3. Accept or reject the visit:
   - If $h(n) > h(c)$ then **accept**
   - Otherwise, accept with $\mathcal{A} = h(x^*)/h(x)$
   - If visit accepted set $x = x^*$

Sampling from histograms



Algorithm:

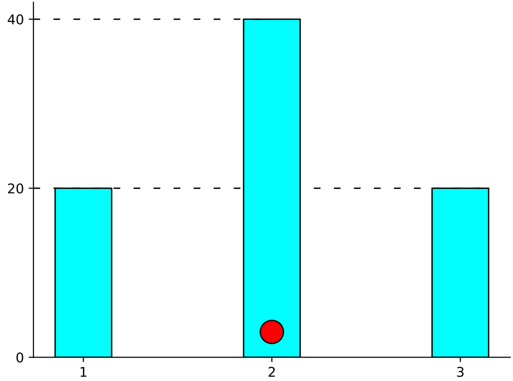
1. Select a starting point (x)
2. Throw a coin to propose a visit to one of the adjacent bars (x^*)
3. Accept or reject the visit:
    - If $h(n) > h(c)$ then **accept**
    - Otherwise, accept with $\mathcal{A} = h(x^*)/h(x)$
    - If visit accepted set $x = x^*$

Note:

- $P(x) = h(x)/C$
- $h(x^*)/h(x) = P(x^*)/P(x)$
- We do not need to know C
- Repeat steps 2-3 many times

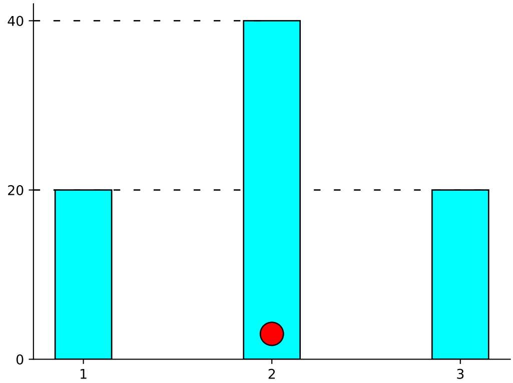
Sampling from histograms

Given we are current at bar 2 (●):



Sampling from histograms

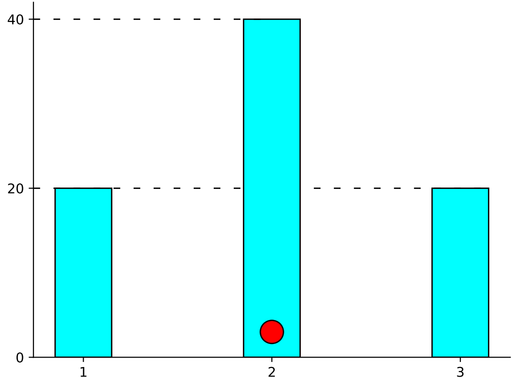
Given we are current at bar 2 (●):



Visit 1

Sampling from histograms

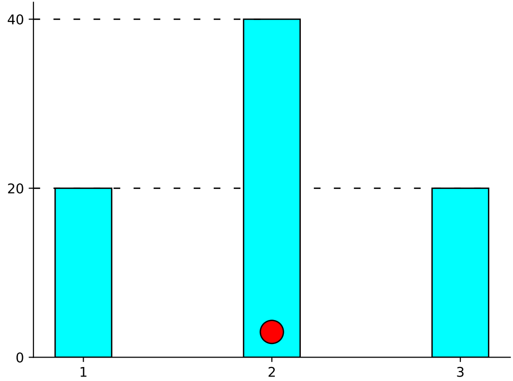
Given we are current at bar 2 (●):



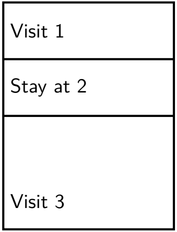
Visit 1

Visit 3

Sampling from histograms



Given we are current at bar 2 (●):



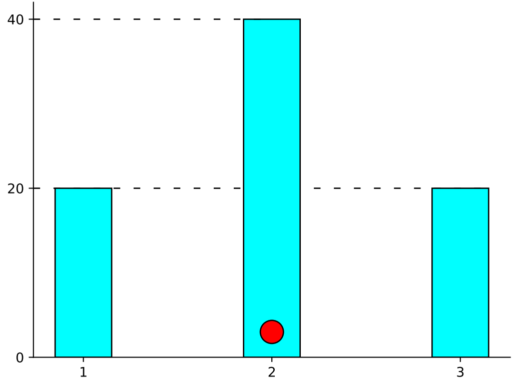
Visit 1

Stay at 2

Visit 3

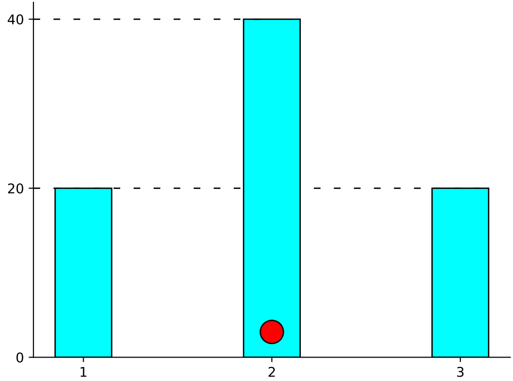
Sampling from histograms

Given we are current at bar 2 (●):



Visit 1
Stay at 2
Stay at 2
Visit 3

Sampling from histograms



Given we are current at bar 2 (●):

Visit 1

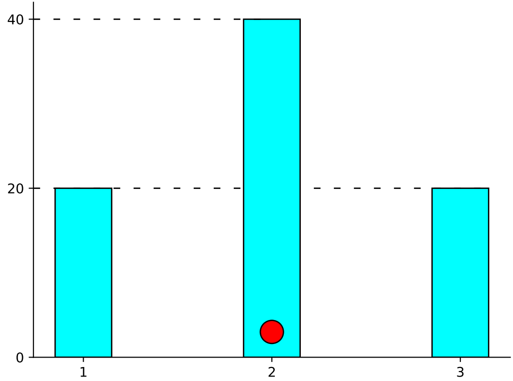
Stay at 2

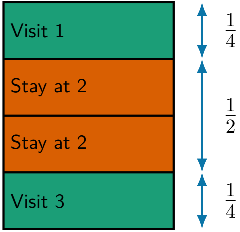
Stay at 2

Visit 3

Sampling from histograms

Given we are current at bar 2 (●):



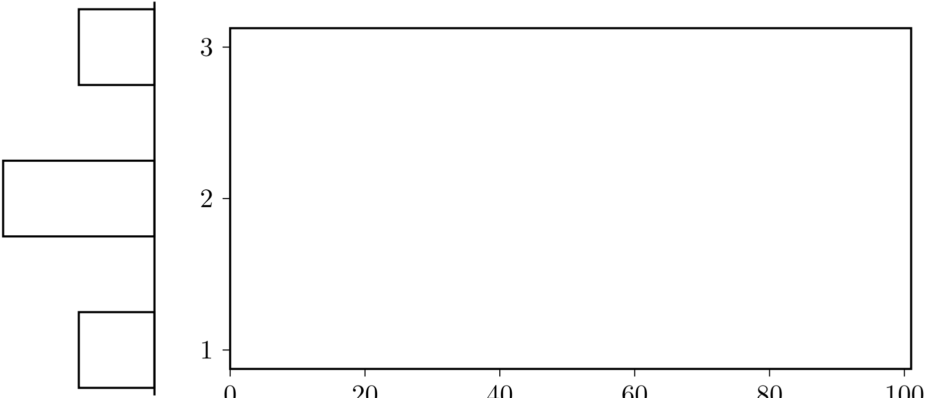

Visit 1

Stay at 2

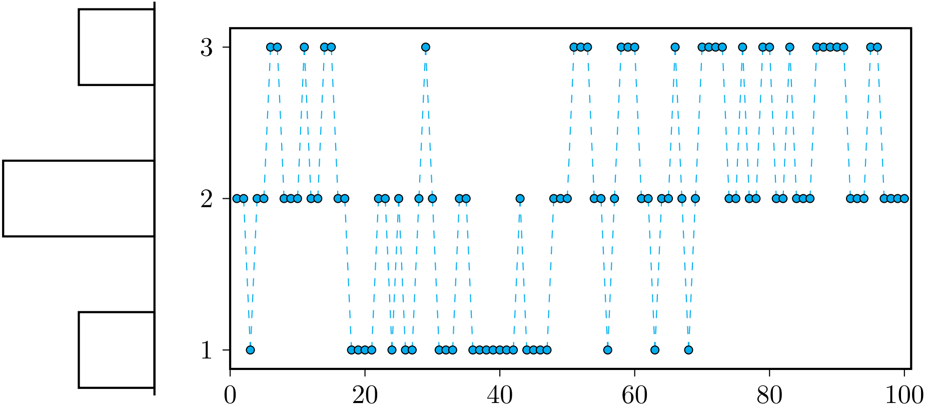
Stay at 2

Visit 3

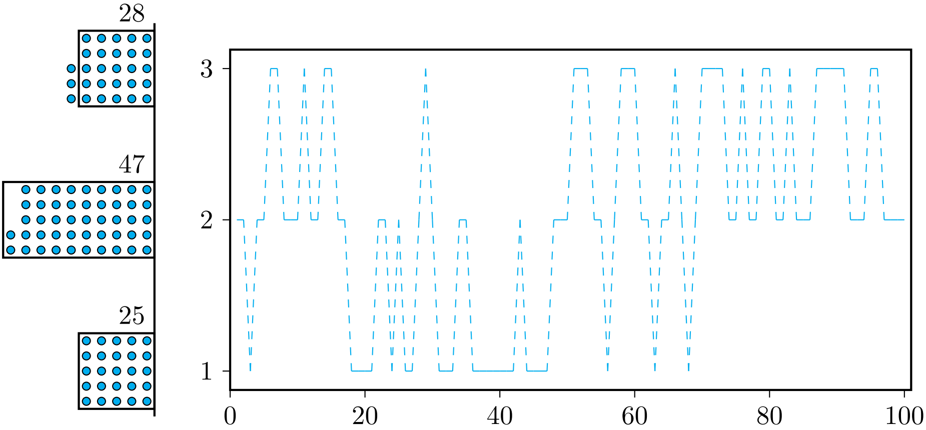
Sampling from histograms



Sampling from histograms

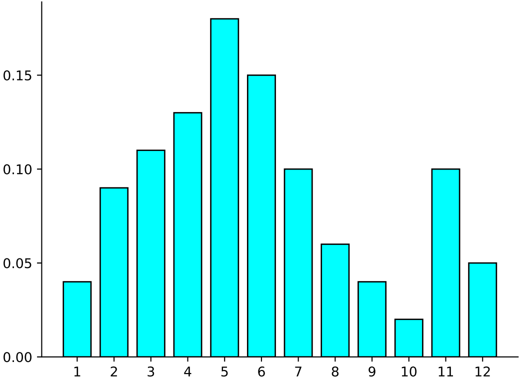


Sampling from histograms



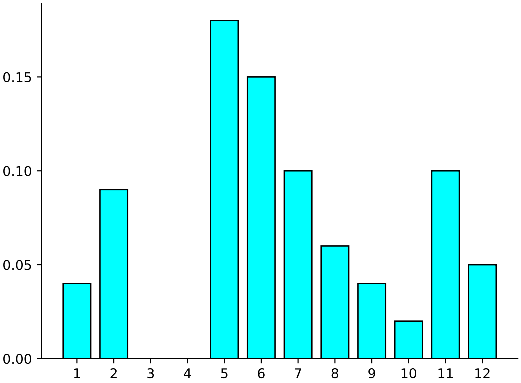
Expectation: 25:50:25

Sampling from histograms



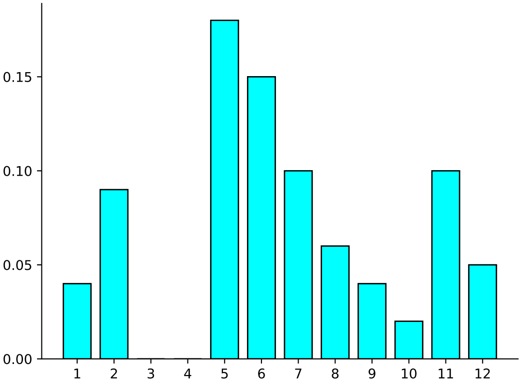
- Works for any histogram

Sampling from histograms



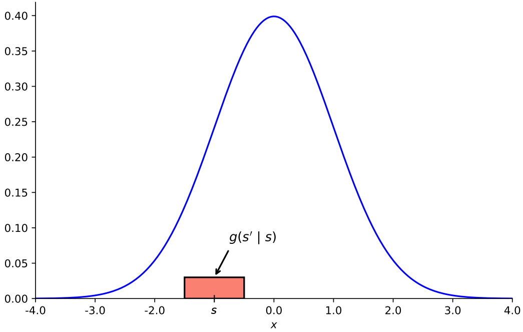
- Works for any histogram
- We can overcome gaps (areas with $h = 0$) by using proposals of different lengths

Sampling from histograms



- Works for any histogram
- We can overcome gaps (areas with $h = 0$) by using proposals of different lengths
- This class of algorithms is known as Markov Chain Monte Carlo or **MCMC**

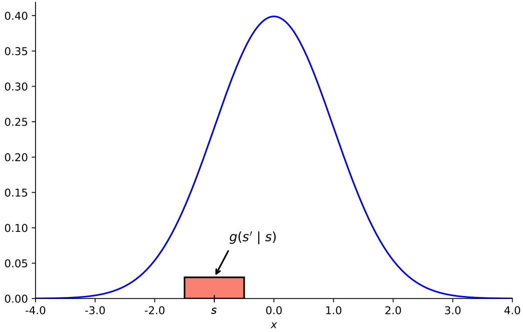
Sampling from histograms



MCMC

- Also works for continuous densities
- Start at some point s
- Use a density $g(s' \mid s)$ to propose the next point s'
- Accept or reject with $P = \min\{1, f(s')/f(s)\}$

Sampling from histograms



MCMC

- Also works for continuous densities
- Start at some point s
- Use a density $g(s' \mid s)$ to propose the next point s'
- Accept or reject with $P = \min\{1, f(s')/f(s)\}$

Make sure that:

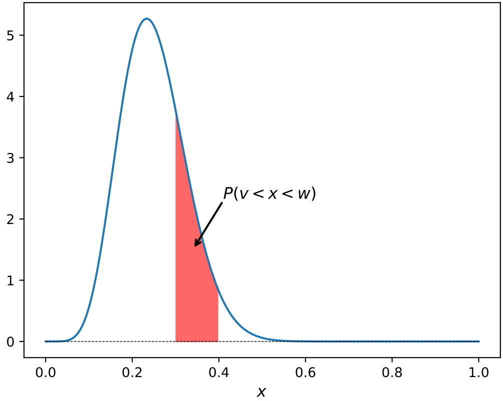
- $g(s' \mid s) = g(s \mid s')$
- This is known as the **Metropolis algorithm**[1]
- Asymmetric proposals[2]

[1] Metropolis *et al*, J. Chem. Phys., (1953) 21:1087-1092

[2] Hastings, Biometrika, (1970) 57:97-109

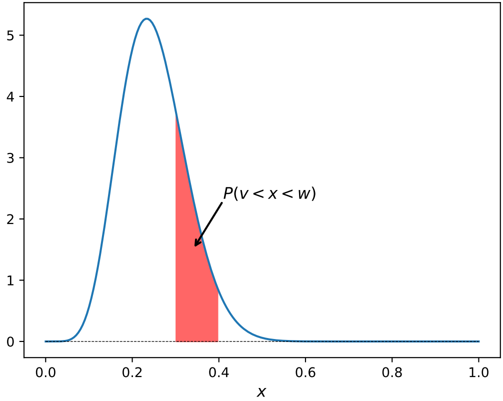
Markov Chain Monte Carlo

How do we calculate $P(v < x < w) = \int_v^w f(x)\,dx$?



Markov Chain Monte Carlo

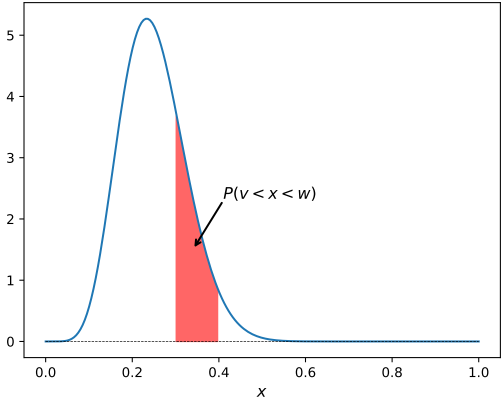
How do we calculate $P(v < x < w) = \int_v^w f(x)\,dx$?



- $P(v < x < w) \approx \frac{n_a}{N}$
- n_a: times red area was visited
- N: total number of MCMC iterations

Markov Chain Monte Carlo

How do we calculate $P(v < x < w) = \int_v^w f(x)\,dx$?

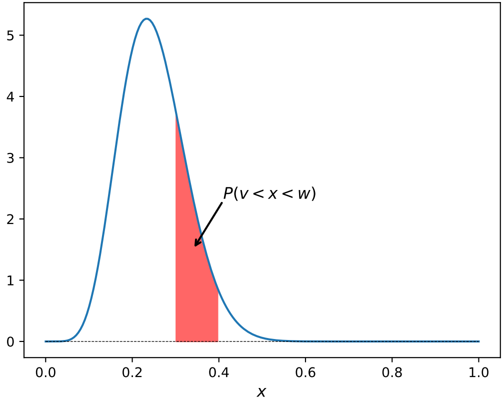


- $P(v < x < w) \approx \frac{n_a}{N}$
- n_a: times red area was visited
- N: total number of MCMC iterations

- $\overline{x} = \int_0^\infty x f(x)\,dx$
- $\overline{x} \approx \frac{1}{N}\sum_{i=1}^N x_i$
- x_i: MCMC sample (visited values)

Markov Chain Monte Carlo

How do we calculate $P(v < x < w) = \int_v^w f(x)\,dx$?



- $P(v < x < w) \approx \frac{n_a}{N}$
- n_a: times red area was visited
- N: total number of MCMC iterations

- $\overline{x} = \int_0^\infty x f(x)\,dx$
- $\overline{x} \approx \frac{1}{N}\sum_{i=1}^{N} x_i$
- x_i: MCMC sample (visited values)

- MCMC gives an approximate answer
- Answer improves with large N

Bayesian Phylogenomics

When analysing phylogenomic data, we are typically interested in estimating:

- A tree topology T
- The branch lengths $\boldsymbol{b}$ given the topology T
- Other model parameters $\boldsymbol{\theta}$

Our data is typically in the form of an alignment matrix D.

Bayesian Phylogenomics

When analysing phylogenomic data, we are typically interested in estimating:

- A tree topology T
- The branch lengths $\boldsymbol{b}$ given the topology T
- Other model parameters $\boldsymbol{\theta}$

Our data is typically in the form of an alignment matrix D.

In a Bayesian framework, we are inferring the **posterior distribution** of $T, \boldsymbol{b}, \boldsymbol{\theta}$ given the data D:

$$f(T, \boldsymbol{b}, \boldsymbol{\theta} \mid D) = \frac{f(\boldsymbol{\theta}) P(T) f(\boldsymbol{b} \mid T) \times P(D \mid \boldsymbol{\theta}, T, \boldsymbol{b})}{P(D)}$$

Bayesian Phylogenomics

When analysing phylogenomic data, we are typically interested in estimating:

- A tree topology T
- The branch lengths $\boldsymbol{b}$ given the topology T
- Other model parameters $\boldsymbol{\theta}$

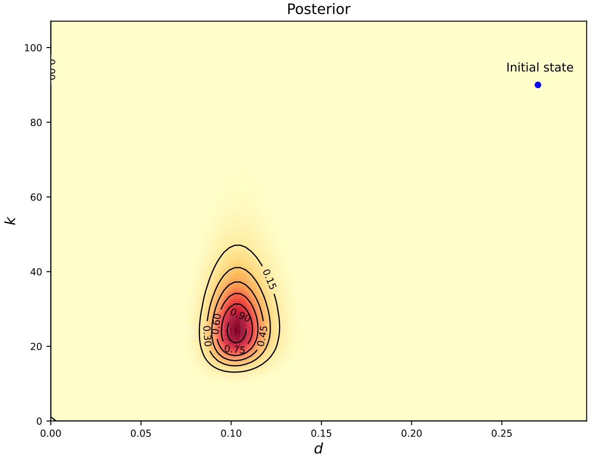
Our data is typically in the form of an alignment matrix D.

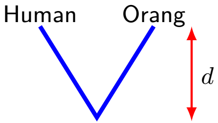
In a Bayesian framework, we are inferring the **posterior distribution** of $T, \boldsymbol{b}, \boldsymbol{\theta}$ given the data D:

$$f(T, \boldsymbol{b}, \boldsymbol{\theta} \mid D) = \frac{f(\boldsymbol{\theta}) P(T) f(\boldsymbol{b} \mid T) \times P(D \mid \boldsymbol{\theta}, T, \boldsymbol{b})}{P(D)}$$

- $P(D) = \sum_i \iint f(\boldsymbol{\theta}, T_i, \boldsymbol{b}) P(D \mid \boldsymbol{\theta}, T_i, \boldsymbol{b})\, d\boldsymbol{\theta} d\boldsymbol{b}$
- $P(D)$ is impossible to calculate, and so we need MCMC
- For example, $P(T \mid D) \approx n_T / N$

Example: K80 model

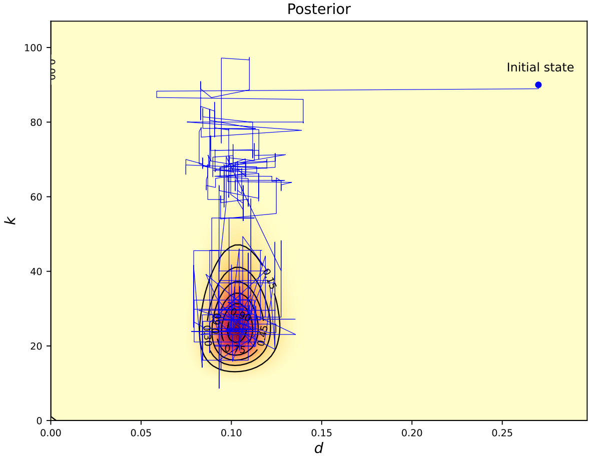


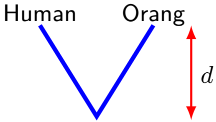


- Two sequences
- d: genetic distance
- k: trans/transv ratio
- Kimura (1980) substitution model
- Data: 948 mit. sites, 84 trans, 6 trasnv

Data from: Yang 2014, p.7, Table 1.3

Example: K80 model

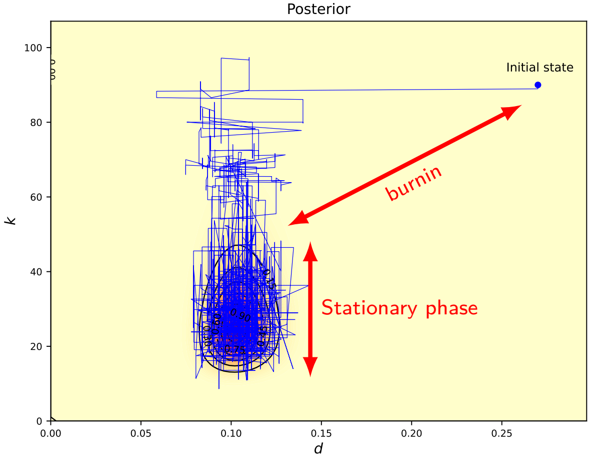


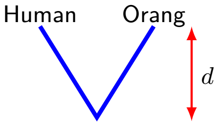


- Two sequences
- d: genetic distance
- k: trans/transv ratio
- Kimura (1980) substitution model
- Data: 948 mit. sites, 84 trans, 6 trasnv

Data from: Yang 2014, p.7, Table 1.3

Example: K80 model

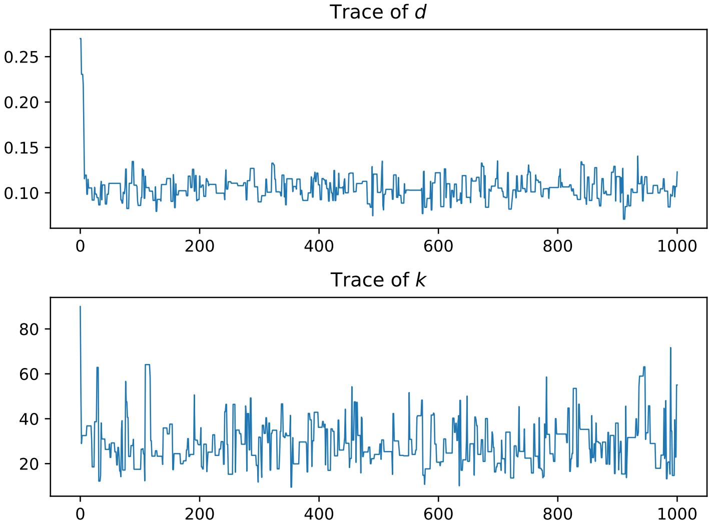




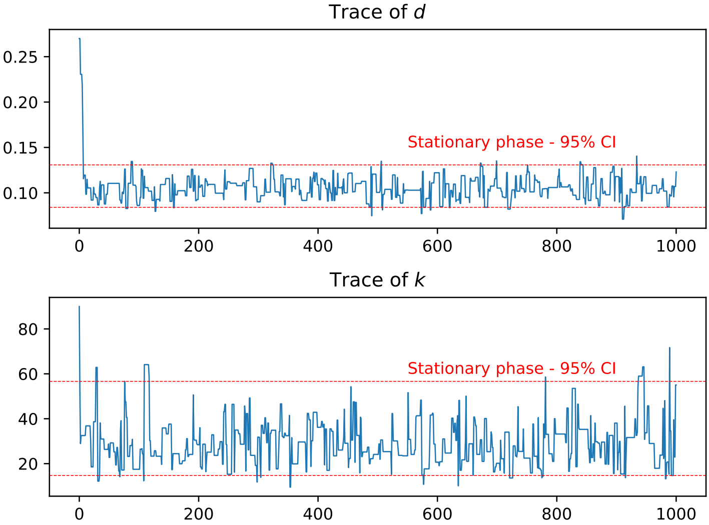
- Two sequences
- d: genetic distance
- k: trans/transv ratio
- Kimura (1980) substitution model
- Data: 948 mit. sites, 84 trans, 6 trasnv

Data from: Yang 2014, p.7, Table 1.3

Example: K80 model

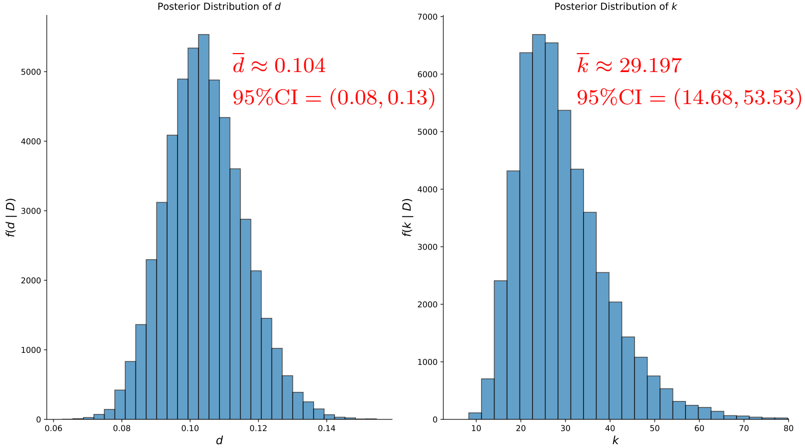


Example: K80 model



Example: K80 model

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



$\overline{d} \approx 0.104$

$95\%\mathrm{CI} = (0.08, 0.13)$

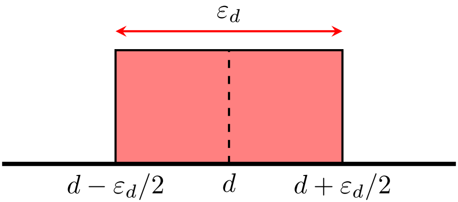
$\overline{k} \approx 29.197$

$95\%\mathrm{CI} = (14.68, 53.53)$

Proposal step size

In this example, we use uniform distributions to propose new values:

- $d' \sim U(d - \varepsilon_d/2, d + \varepsilon_d/2)$
- $k' \sim U(k - \varepsilon_k/2, k + \varepsilon_k/2)$
- $\varepsilon_d, \varepsilon_k$ are known as the proposal step sizes



Mixing and convergence rate

Mixing: refers to how quickly a chain explores the state space.

- Rejecting too many proposals means we stay in the same place too long
- If we accept too many proposals usually means we are moving slowly, remaining in the same region too long.

Mixing and convergence rate

Mixing: refers to how quickly a chain explores the state space.

- Rejecting too many proposals means we stay in the same place too long
- If we accept too many proposals usually means we are moving slowly, remaining in the same region too long.

Proposal step size **affects** mixing:

- Step is too big: we reject most proposals
- Step is too small: we accept most proposals (baby steps)

Mixing and convergence rate

Mixing: refers to how quickly a chain explores the state space.

- Rejecting too many proposals means we stay in the same place too long
- If we accept too many proposals usually means we are moving slowly, remaining in the same region too long.

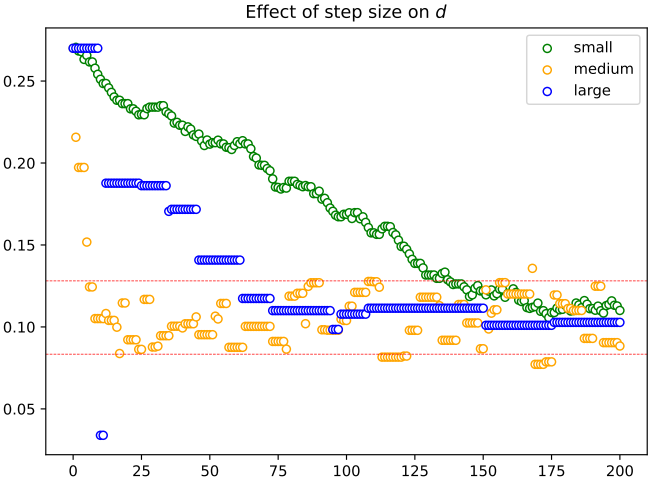
Proposal step size **affects** mixing:

- Step is too big: we reject most proposals
- Step is too small: we accept most proposals (baby steps)

Convergence rate: refers to how quickly the chain moves into the stationary phase

- Proposal step size also affects convergence rate
- Small sizes lead to low convergence rate

Example: K80 model



Mixing (acceptance %): Small: 82%, Medium: 34%, Large: 7%

Mixing and fine tuning

Fine-tuning: The process of adjusting step sizes to achieve optimal mixing

- Analysis of normal distribution indicates that mixing is best at $\sim$ 23.4% (20% – 40%)^{1,2}
- Most MCMC software will do this automatically, but sometimes it is useful to do it manually:
    - is too high: increase step size
    - is too low: decrease step size

[1] Gelman et al, Ann. Appl. Probab. 7(1):110-120, 1997

[2] Roberts and Rosenthal, Statist. Sci 16(4):351-367, 2001

Mixing and fine tuning

Fine-tuning: The process of adjusting step sizes to achieve optimal mixing

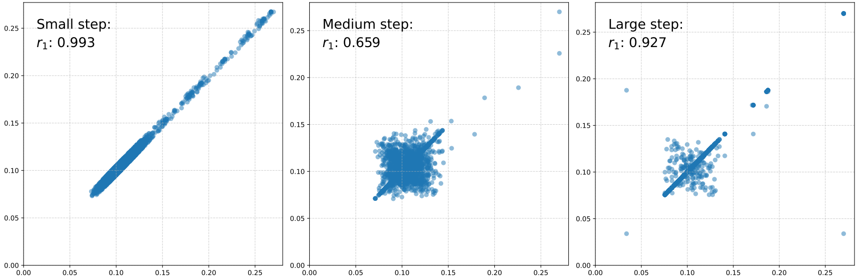
- Analysis of normal distribution indicates that mixing is best at $\sim$ 23.4% (20% – 40%)[1,2]
- Most MCMC software will do this automatically, but sometimes it is useful to do it manually:
  - is too high: increase step size
  - is too low: decrease step size
- Recall MCMC estimates are approximate, e.g. $\overline{d} \approx \sum_i d_i/N$
- For two chains of the same length, the errors in the estimates are larger for the chain with poorest mixing
- Note: calculations are done after removing burn-in samples

[1] Gelman et al, Ann. Appl. Probab. 7(1):110-120, 1997

[2] Roberts and Rosenthal, Statist. Sci 16(4):351-367, 2001

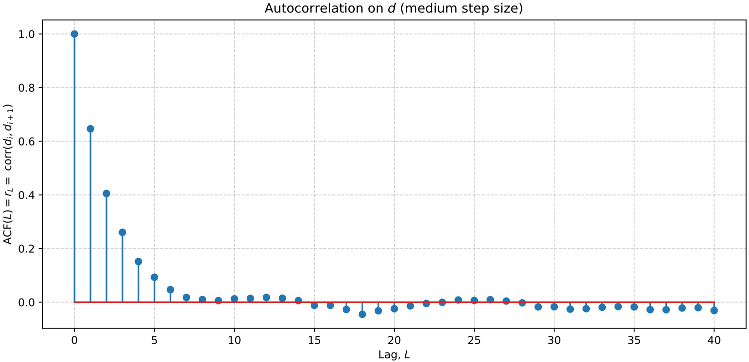
Autocorrelation

- MCMC samples are autocorrelated because accepted values are modifications of the previous values
- K80 example, $r_L = \mathrm{corr}(d_i, d_{i+L})$
- L indicates the lag. Plots below are for $L = 1$



Mixing (acceptance %): Small: 82%, Medium: 34%, Large: 7%

Autocorrelation Function



Chains that mix well have ACF that decays fast!

Efficiency

Chains that lead to estimates with small errors with respect to the chain's size are said to be **efficient**

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:

$$\mathrm{Eff} = \frac{1}{1 + 2(r_1 + r_2 + r_3 + \ldots)}$$

- **High (+) autocorrelation:** Low efficiency
- **Moderate (+) autocorrelation:** Efficient chain
- **No autocorrelation:** Independent sampling (very efficient)
- **(-) autocorrelation:** Super-efficient chain

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:

$$\mathrm{Eff} = \frac{1}{1 + 2(r_1 + r_2 + r_3 + \ldots)}$$

- **High (+) autocorrelation:** Low efficiency
- **Moderate (+) autocorrelation:** Efficient chain
- **No autocorrelation:** Independent sampling (very efficient)
- **(-) autocorrelation:** Super-efficient chain

- $\mathrm{Eff} = 1$: as efficient as independent sampling
- $\mathrm{Eff} = 0.2$: 20% as efficient as independent sampling

Effective Sample Size

Effective Sample Size is the chain size × efficiency

$$\mathrm{ESS} = N \times \mathrm{Eff}$$

Example:

- We have an MCMC chain with $N = 1000$ samples and $\mathrm{Eff} = 20\%$
- Then, $\mathrm{ESS} = 200$, meaning the chain has the same estimate error as an equivalent, independent chain of size 200

Effective Sample Size

Effective Sample Size is the chain size × efficiency

$$\mathrm{ESS} = N \times \mathrm{Eff}$$

Example:

- We have an MCMC chain with $N = 1000$ samples and $\mathrm{Eff} = 20\%$
- Then, $\mathrm{ESS} = 200$, meaning the chain has the same estimate error as an equivalent, independent chain of size 200

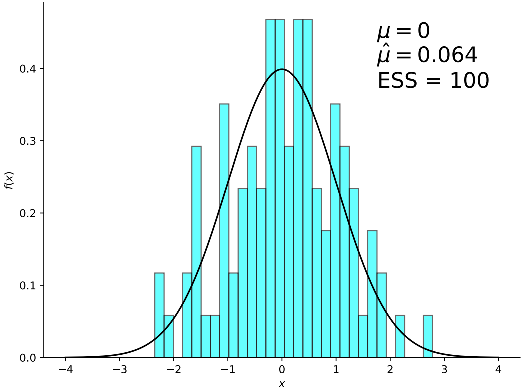
Stachastic simulation theory recommendation:

- N should be between $1\,000$ to $10\,000$ for independent sampling
- Thus, ESS should be between $1\,000$ to $10\,000$
- This is typically hard to achieve in Bayesian phylogenomics
- We aim to have at least $\mathrm{ESS} > 200$

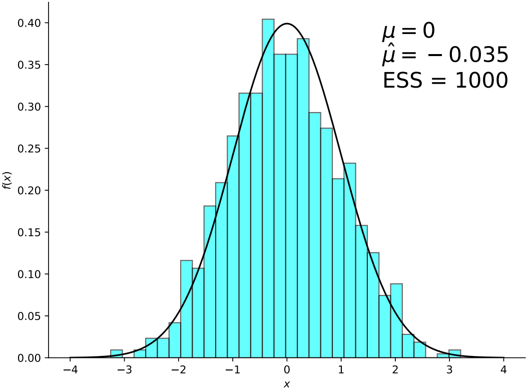
Convergence

- MCMC is a class of stochastic algorithms
- An MCMC histogram is just an approximation of the posterior density
- This approximation improves as $N \to \infty$
- We must use **convergence diagnostics** to assess whether the MCMC sample has converged to the posterior

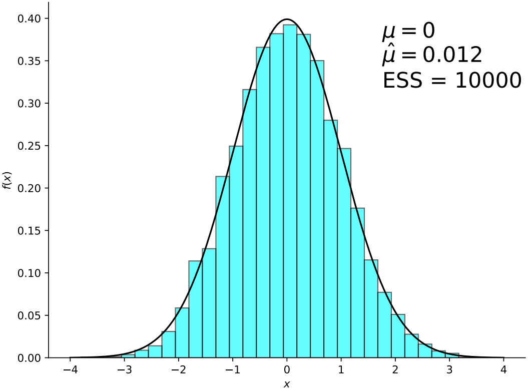
Convergence to Normal Distribution



Convergence to Normal Distribution



Convergence to Normal Distribution



Convergence

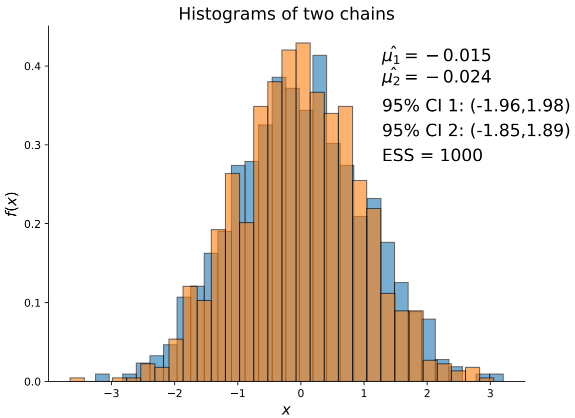
- In practice the shape of the posterior density is not known
- Thus, we cannot compare the MCMC histogram to the true posterior

Convergence

- In practice the shape of the posterior density is not known
- Thus, we cannot compare the 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 we have converged

Important: The chains must start from different starting points

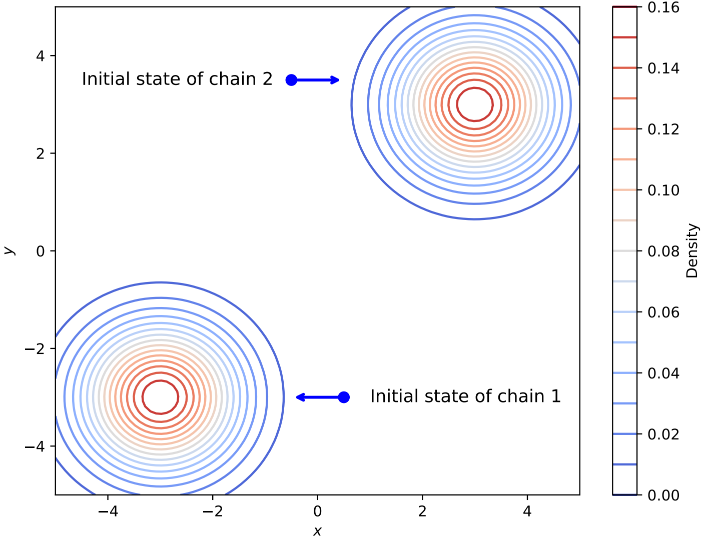
Convergence to Normal Distribution



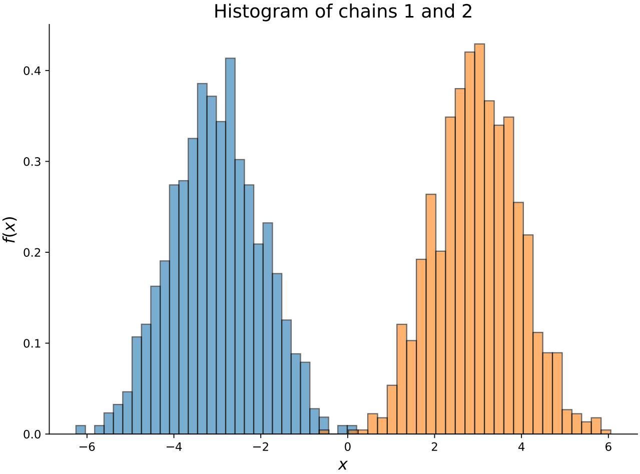
Chains that have converged can be merged into a larger chain

$$\mathrm{ESS}_L = \mathrm{ESS}_1 + \mathrm{ESS}_2$$

Multi-modal densities



Multi-modal densities



No convergence! Chains failed to cross the posterior valley

Convergence

- Running **many chains** with **random starting points** is a good way to detect multi-modal posteriors

Convergence

- Running **many chains** with **random starting points** is a good way to detect multi-modal posteriors

- **If we 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 converge
  - **Do not merge** short chains that are stuck at different modes

Convergence

- Running **many chains** with **random starting points** is a good way to detect multi-modal posteriors

- **If we 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 converge
  - **Do not merge** short chains that are stuck at different modes

- **Important:**
  - Avoid using fixed starting points (or seeds)
  - ESS is not a measure of convergence

Thinning the chain

- In phylogenomics it is difficult to construct efficient chains
- That is because we usually have too many parameters in our models
- Phylogenomic MCMC chains are thus highly correlated
- To get good estimates, we need to run the chains for a very long time
- If we store every chain visit, we would run out of hard disk space very quickly.
- **Thinning:** Writing down only a fraction of all chain visits (e.g. every 100^{th} or 1000^{th} visit)

Bayesian Phylogenomics

In phylogenomics, we are interested in estimating the topology T, branch lengths $\boldsymbol{b}$ and model parameters $\boldsymbol{\theta}$, given the alignment D

- $f(T, \boldsymbol{b}, \boldsymbol{\theta} \mid D) = \frac{f(T,\boldsymbol{b},\boldsymbol{\theta}) f(D \mid T,\boldsymbol{b},\boldsymbol{\theta})}{f(D)}$
- $f(D) = \sum \iint f(T, \boldsymbol{b}, \boldsymbol{\theta}) f(D \mid T, \boldsymbol{b}, \boldsymbol{\theta}) \, d\boldsymbol{b} \, d\boldsymbol{\theta}$

$f(D)$ is typically not available analytically and thus we use MCMC

Bayesian Phylogenomics

The likelihood of the data $D[1,\ldots,n]$ (alignment) is the product of the likelihood of sites

$$f(D \mid T, \boldsymbol{b}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(D[i] \mid T, \boldsymbol{b}, \boldsymbol{\theta})$$

- $D[i]$ is the i-th site pattern
- n is the number of site patterns

Bayesian Phylogenomics

The likelihood of the data $D[1,\ldots,n]$ (alignment) is the product of the likelihood of sites

$$f(D \mid T, \boldsymbol{b}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(D[i] \mid T, \boldsymbol{b}, \boldsymbol{\theta})$$

- $D[i]$ is the i-th site pattern
- n is the number of site patterns

MCMC algorithm:

1. Choose random initial state for $T, \boldsymbol{b}, \boldsymbol{\theta}$
2. Propose topology T and accept/reject
3. Propose branch lengths $\boldsymbol{b}$ and accept/reject
4. Propose model parameters $\boldsymbol{\theta}$ and accept/reject
5. Store the current values of parameters into a sample file
6. Repeat steps 2-5 **many many** times

Additional resources

- Holder & Lewis (2003) **Phylogeny estimation: Traditional and Bayesian approaches**. *Nat. Rev. Genet.*, 4:275
- Yang (2014) **Molecular evolution: A statistical approach**. *Oxford University Press*
- Chen, Kuo, Lewis (2014) **Bayesian phylogenetics: Methods, algorithms, and applications**. *CRC Press*
- Kapli *et al* (2020) **Phylogenetic tree building in the genomic age**. *Nat. Rev. Genet.*, 21(7):428-444
- **THE END**

Thanks to Mario dos Reis for several lecture materials