

# Bayes' Theorem

ESS 575 Models for Ecological Data

N. Thompson Hobbs

February 14, 2019





## THÉORIE

ANALYTIQUE

### DES PROBABILITÉS;

PAR M. LE COMTE LAPLACE,

Chancelier du Sénat-Conservateur, Grand-Officier de la Légion d'Honneur;  
Membre de l'Institut impérial et du Bureau des Longitudes de France;  
des Sociétés royales de Londres et de Göttingen; des Académies des  
Sciences de Russie, de Danemark, de Suède, de Prusse, de Hollande,  
d'Italie, etc.

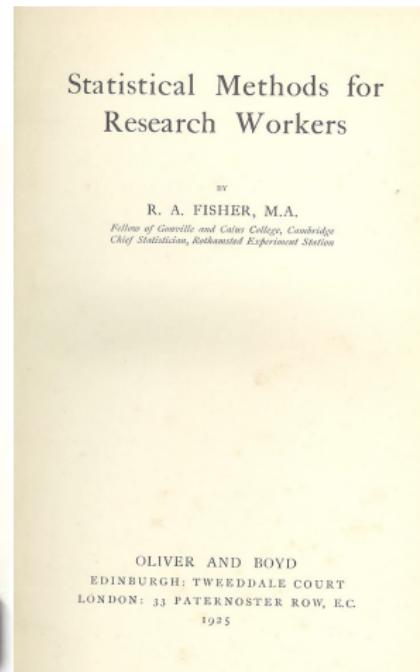
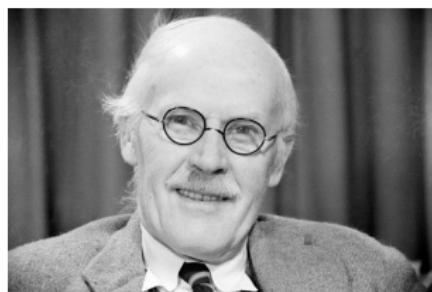
---

PARIS,

M<sup>e</sup> V<sup>e</sup> COURCIER, Imprimeur-Libraire pour les Mathématiques,  
quai des Augustins, n<sup>o</sup> 57.  
1812.

"Laplace's principle being dead, it should be decently buried out of sight, and not embalmed in text-books and examination papers... The indiscretions of great men should be quietly allowed to be forgotten."

George Chrystal 1891



“My personal conviction is that the theory of inverse probability is founded upon an error and must be wholly rejected”

R. A. Fischer

“There has not been a single date in the history of the law of gravitation when a modern significance test would not have rejected the law outright.”

Harold Jeffreys

“The p-value is almost nothing sensible you can think of. I tell students to give up trying.”

Stephen Goodman

What is the collective noun for a group of statisticians?

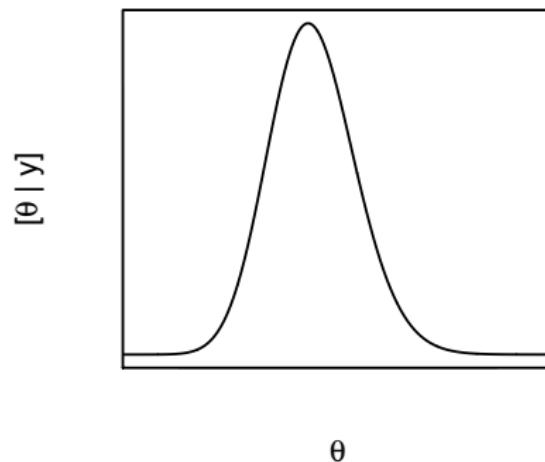


Gelfand, A. E., and A. F. M. Smith. 1990. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85:398-409.



# Bayesian inference

All unobserved quantities are treated as *random variables*.



# Random variables

All unobserved quantities are treated in exactly the same way.

- ▶ model parameters
- ▶ latent states
- ▶ missing data
- ▶ predictions and forecasts
- ▶ observations (before they are observed)

## Exercise

- ▶ Assume we have two, jointly distributed random variables,  $\theta$  and  $y$ . The random variable  $\theta$  represents unobserved quantities of interest. The random variable  $y$  represents observations, which become fixed *after* they are observed.
- ▶ Derive Bayes' Theorem

$$[\theta|y] = \frac{[y|\theta][\theta]}{[y]} \quad (1)$$

using your knowledge of the laws of probability, particularly the definition of conditional probability.

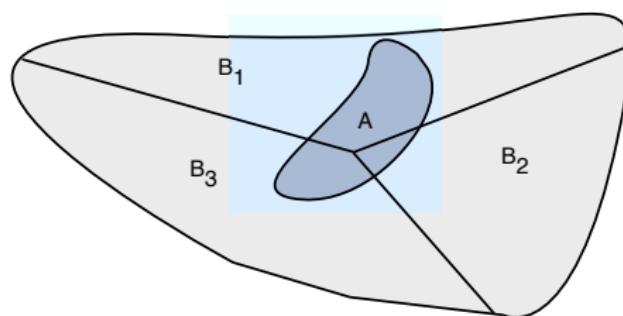
We will often make use of the equivalent equation

$$[\theta|y] = \frac{[y, \theta]}{[y]} \quad (2)$$

as a starting point for developing hierarchical models by factoring  $[y, \theta]$  into ecologically sensible components that can be treated in MCMC as univariate distributions. More about that soon.

# What is $[y]$ ?

Recall the law of total probability for discrete random variables



$$[A] = \sum_i [A | B_i] [B_i]. \quad (3)$$

and for continuous random variables

$$[A] = \int_B [A|B] [B] dB. \quad (4)$$

# What is $[y]$ ?

It follows that

$$[y] = \sum_{\theta_i \in \{\Theta\}} [y|\theta_i][\theta_i] \text{ for discrete parameters} \quad (5)$$

$$[y] = \int_{\theta} [y|\theta][\theta] d\theta \text{ for continuous parameters.} \quad (6)$$

Thus, Bayes theorem for discrete valued parameters is

$$[\theta|y] = \frac{[y|\theta_i][\theta_i]}{\sum_{\theta_i \in \{\Theta\}} [y|\theta_i][\theta_i]} \quad (7)$$

and for parameters that are continuous,

$$[\theta|y] = \frac{[y|\theta][\theta]}{\int_{\theta} [y|\theta][\theta] d\theta}. \quad (8)$$

## More about $[y]$

- ▶  $[y]$  is the marginal distribution of the data, a *distribution* before the data are observed and a *normalizing constant* after the data are observed.
- ▶ It is also called the *prior predictive distribution*. Why?
- ▶ Because  $[y]$  is a constant after the data are observed,

$$[\theta|y] \propto [y, \theta] \tag{9}$$

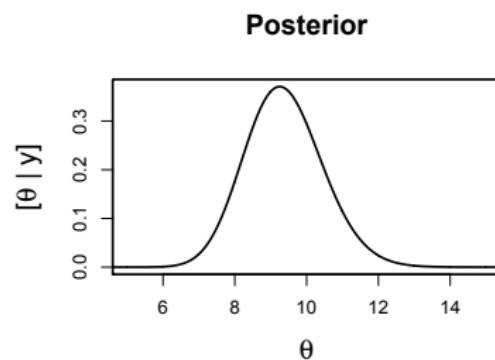
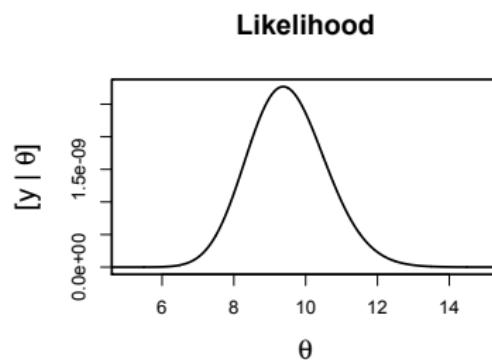
$$\propto [y|\theta] [\theta] \tag{10}$$

## [y] is critical to Bayes

$$\mathbf{y} = (5, 10, 11, 12, 14, 9, 8, 6)'$$

$$\text{likelihood} = \prod_{i=1}^8 \text{Poisson}(y_i | \theta)$$

$$\text{posterior} = \frac{\prod_{i=1}^8 \text{Poisson}(y_i | \theta) \text{gamma}(\theta | .0001, .0001)}{[y]}$$



Cut to example

## Probability mass function $[y|\theta]$

$\theta$  is known to be  $\frac{1}{2}$ . Probability of number of whites conditional on three draws and  $\theta = \frac{1}{2}$ :

$y = \text{Number of whites}$	$[y \theta]$
0	.125
1	.375
2	.375
3	.125
$\sum_{i=1}^4 [y \theta_i] =$	1

# Likelihood $[y|\theta]$

Probability of two whites on three draws conditional on  $\theta_i$

Parameter	Likelihood $[y \theta_i]$
$\theta_1=5/6$	.347
$\theta_2=1/2$	.375
$\theta_3=1/6$	.069
$\sum_{i=1}^3 [y \theta_i] =$	.791

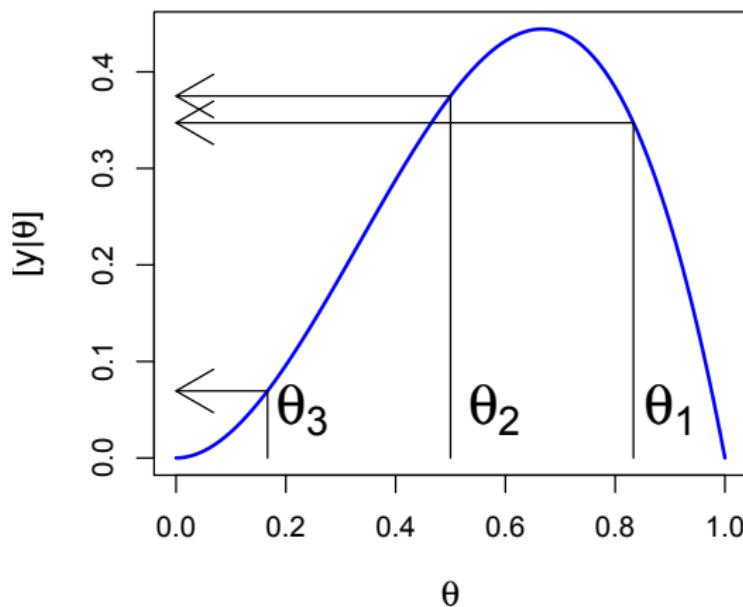
# Posterior distribution $[\theta|y]$

Probability of  $\theta_i$  conditional on two whites on three draws

Parameter	Prior $[\theta_i]$	Likelihood $[y \theta_i]$	Joint $[y \theta_i][\theta_i]$	Posterior $\frac{[y \theta_i][\theta_i]}{[y]} = [\theta_i y]$
$\theta_1$	0.333	0.347	0.115	0.439
$\theta_2$	0.333	0.375	0.125	0.474
$\theta_3$	0.333	0.069	0.023	0.087
$[y] = \sum_{i=1}^3 [y \theta_i][\theta_i] =$			0.261	$\sum_{i=1}^3 [\theta_i y] = 1$

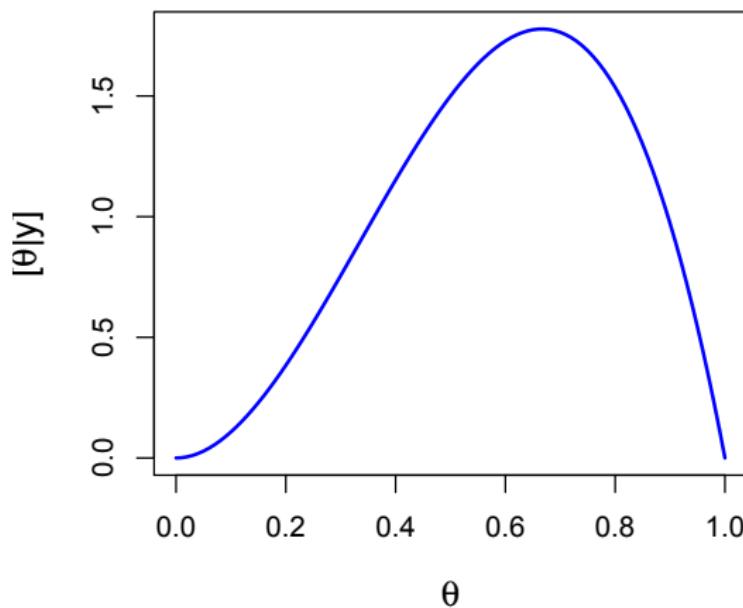
# Likelihood profile $[y|\theta]$

[2 white on 3 draws| $\theta$ ]



# Posterior distribution $[\theta|y]$

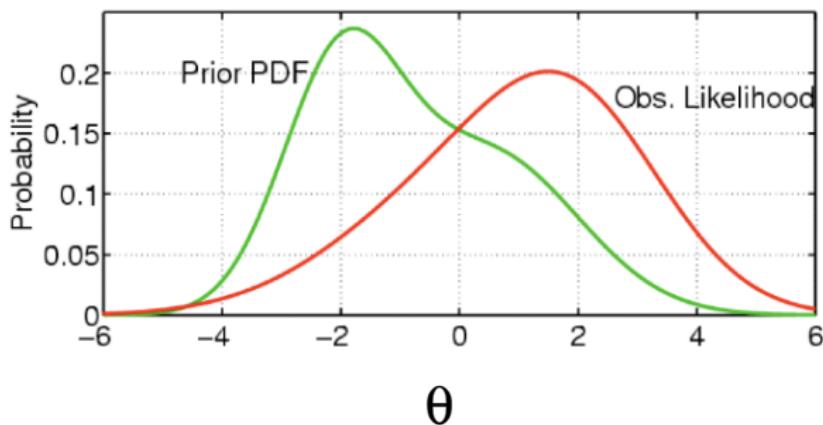
$[\theta|2 \text{ white on 3 draws}]$



# The components of Bayes theorem

$$\widehat{[\theta|y]} = \frac{\widehat{[y|\theta]} \widehat{[\theta]}}{\underbrace{\int_{\theta} [y|\theta][\theta] d\theta}_{\text{marginal distribution of data}}} \quad (11)$$

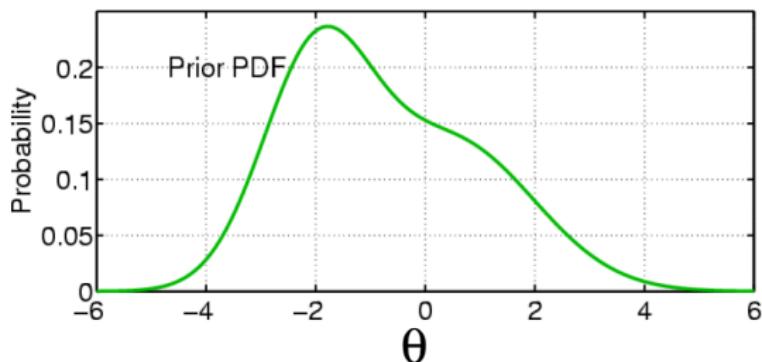
## The components of Bayes Theorem



Courtesy of Chris Wikle, University of Missouri

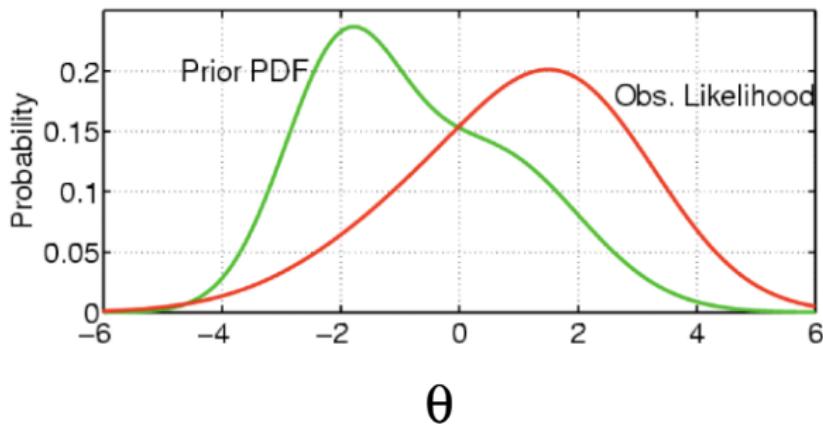
$$[\theta|y] = \frac{[y|\theta][\theta]}{\int_{\theta} [y|\theta][\theta] d\theta} \quad (12)$$

The prior,  $[\theta]$ , can be informative or vague.



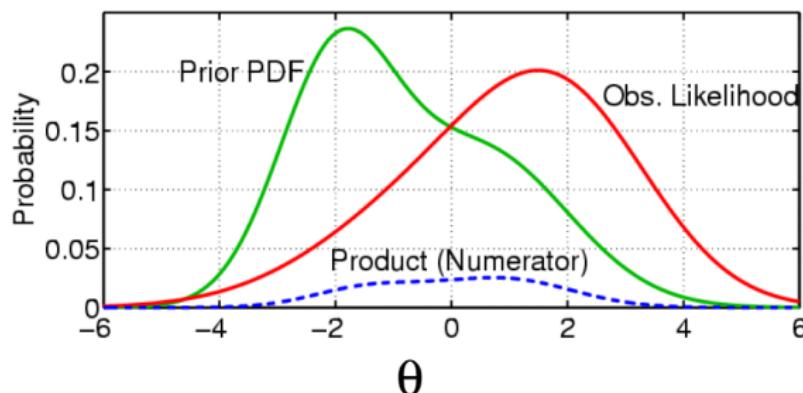
$$[\theta|y] = \frac{[y|\theta][\theta]}{\int_{\theta} [y|\theta][\theta] d\theta} \quad (13)$$

The likelihood (a.k.a. data distribution,  $[y|\theta]$ )



$$[\theta|y] = \frac{[y|\theta][\theta]}{[y]} = \frac{[y|\theta][\theta]}{\int_{\theta} [y|\theta][\theta] d\theta} \quad (14)$$

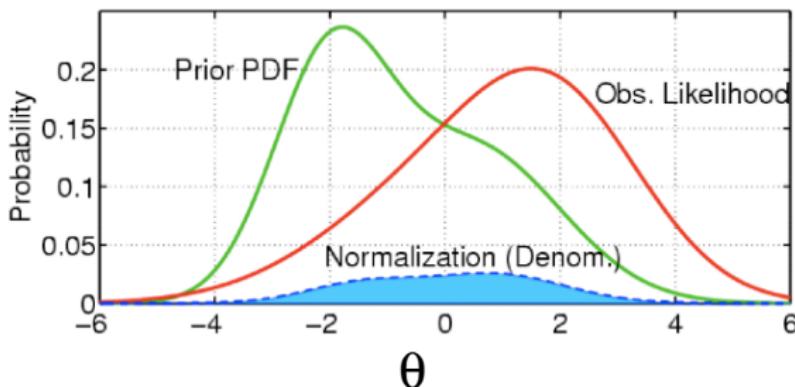
The product of the prior and the likelihood,  $[y|\theta][\theta]$ , the joint distribution of the parameters and the data,  $[y,\theta]$ .



What is the maximum likelihood estimate of  $\theta$ ?

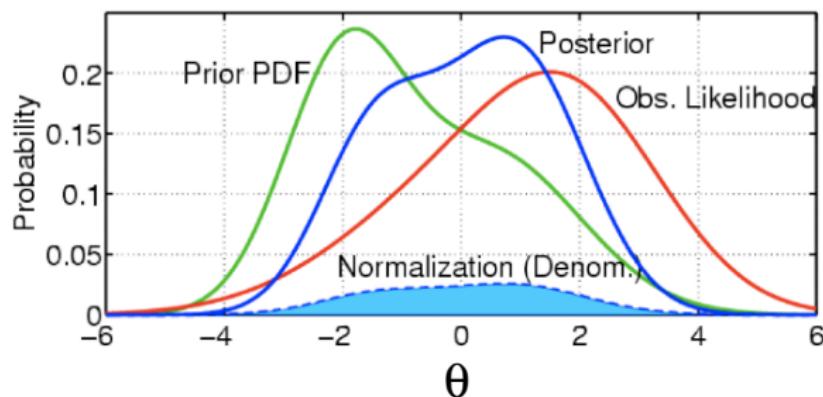
$$[\theta|y] = \frac{[y|\theta][\theta]}{\int_{\theta} [y|\theta][\theta] d\theta} \quad (15)$$

The marginal distribution of the data (the denominator) is the area under the joint distribution.



What we are seeking: The posterior distribution,  $[\theta|y]$ .

$$[\theta|y] = \frac{[y|\theta][\theta]}{\int_{\theta} [y|\theta][\theta] d\theta} \quad (16)$$



Note that we are dividing each point on the dashed line by the area under the dashed line to obtain a probability density function reflecting our prior and current knowledge about  $\theta$ .

## So what?

What does this enable you to do? Review factoring joint distributions:

Remember from the basic laws of probability that

$$p(z_1, z_2) = p(z_1 | z_2)p(z_2) = p(z_2 | z_1)p(z)_1$$

This generalizes to:

$$\mathbf{z} = (z_1, z_2, \dots, z_n)$$

$$p(z_1, z_2, \dots, z_n) = p(z_n | z_{n-1}, \dots, z_1) \dots p(z_3 | z_2, z_1)p(z_2 | z_1)p(z_1)$$

where the components  $z_i$  may be scalars or subvectors of  $\mathbf{z}$  and the sequence of their conditioning is arbitrary. This equation can be simplified using knowledge of independence.

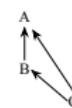
# So what?

I



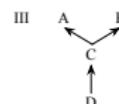
$$\Pr(A, B) = \Pr(A|B) \Pr(B)$$

II



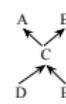
$$\Pr(A, B, C) = \Pr(A|B, C) \times \Pr(B|C) \Pr(C)$$

III



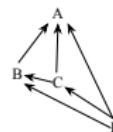
$$\Pr(A, B, C, D) = \Pr(A|C) \times \Pr(B|C) \Pr(C|D) \Pr(D)$$

IV



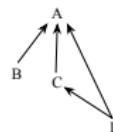
$$\Pr(A, B, C, D, E) = \Pr(A|C) \times \Pr(B|C) \Pr(C|D, E) \times \Pr(D) \Pr(E)$$

V



$$\Pr(A, B, C, D) = \Pr(A|B, C, D) \times \Pr(B|C, D) \times \Pr(C|D) \Pr(D)$$

VI



$$\Pr(A, B, C, D) = \Pr(A|B, C, D) \times \Pr(C|D) \times \Pr(B) \Pr(D)$$

# So what?

$$\widehat{[\theta|y]} = \frac{[y, \theta]}{[y]} = \frac{\overbrace{[y|\theta]}^{\text{likelihood}} \overbrace{[\theta]}^{\text{prior}}}{\underbrace{\int_{\theta} [y|\theta][\theta] d\theta}_{\text{marginal}}} \quad (17)$$

Useful models will be more complex:

$$\underbrace{[\theta_1, \theta_2, \theta_3, \dots, \theta_n, z_1, z_2 \dots z_n | y_1, y_2]}_{\text{multiple parameters, latent states, data sets}} \propto \underbrace{[\theta_1, \theta_2, \theta_3, \dots, \theta_n, z_1, z_2 \dots z_n, y_1, y_2]}_{\text{factor into conditional distributions}}$$

We use the rules of probability to factor complex joint distributions into a series of conditional distributions. We can then use the Markov chain Monte Carlo algorithm to escape the need for integrating the marginal data distribution, allowing us to find the marginal posterior distributions of all of the unobserved quantities. Which, of course, is where we started out. And where we are headed.