Bayesian statistics – Nanjing Forestry University

Lecture 2

Wei Wu, The University of Southern Mississippi

December, 2016

Acknowledgement: SESYNC Bayesian modeling for socio-environmental data workshop

THE UNIVERSITY OF
SOUTHERN MISSISSIPPI.
GULF COAST RESEARCH LABORATORY

Bayesian inference

Bayesian inference

- Bayes' theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(\theta | data) = \frac{P(data | \theta)P(\theta)}{P(data)} \propto P(data | \theta)P(\theta)$$

Posterior = likelihood*prior/normalizing constant

Posterior $\propto$ likelihood*prior

- Advantage of Bayesian inference

-- simplicity of interpretation of the results

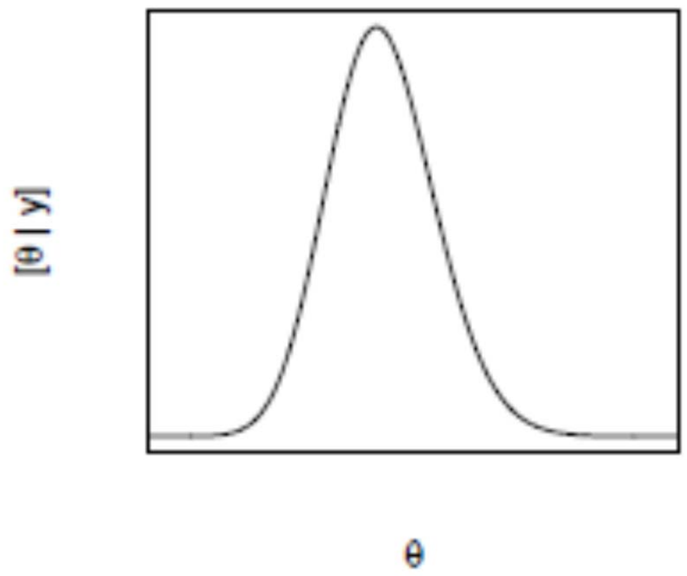
-- being consistent and ability of leaning from previous knowledge

-- estimating different sources of errors

-- data assimilation is automatic and the data need not be at the same scale and time

Bayesian inference

All unobserved quantities are treated as random variables



Exercise

Assume we have two jointly distributed random variables: θ and y. The random variable θ represents unobserved quantities of interest. The random variable y represents observations, which become fixed after they are observed.

Derive Bayes' theorem

$$[\theta \mid y] = \frac{[y \mid \theta][\theta]}{[y]}$$

Using your knowledge of laws of probability, particular the definition of conditional probability.

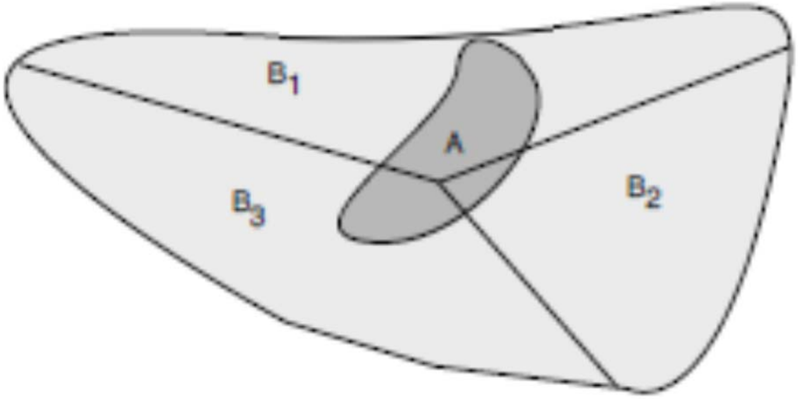
We will often make use of the equivalent equation

$$[\theta \mid y] = \frac{[y,\theta]}{[y]}$$

As starting point for developing hierarchical models by factoring [y,θ] into ecologically sensible components that can be treated in Markov Chain Monte Carlo simulation as univariate distributions.

What is [y]

Recall the law of total probability for discrete random variables



$$[A] = \sum_i [A|B_i][B_i]$$

and for continuous random variables

$$[\mathrm{A}] = \int_B [A \mid B][B]dB$$

What is [y]

It follows that

$$[y] = \sum_{\theta_i \in \{\Theta\}} [y \mid \theta_i][\theta_i] \text{ for discrete parameters}$$

$$[\mathrm{y}] = \int_{\theta} [y \mid \theta][\theta] d\theta \text{ for continuous parameters}$$

$$[\theta \mid \mathrm{y}] = \frac{[\mathrm{y} \mid \theta_{\mathrm{i}}][\theta_i]}{\sum_{\theta_i \in \{\Theta\}} [y \mid \theta_i][\theta_i]} \text{ for discrete valued parameters}$$

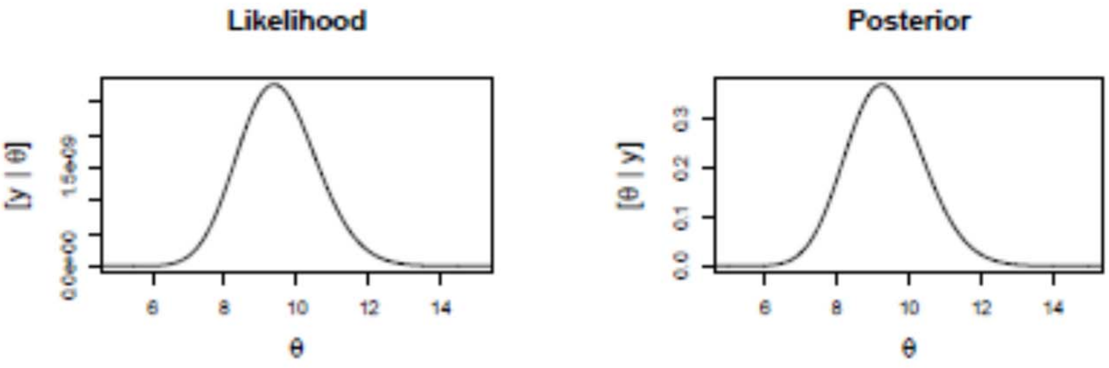
$$[\theta \mid \mathrm{y}] = \frac{[\mathrm{y} \mid \theta_{\mathrm{i}}][\theta_i]}{\int_{\theta} [y \mid \theta][\theta] d\theta} \text{ for parameters that are continous}$$

[y] is critical to Bayes

Y=(5,10,11,12,14,9,8,6)

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

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

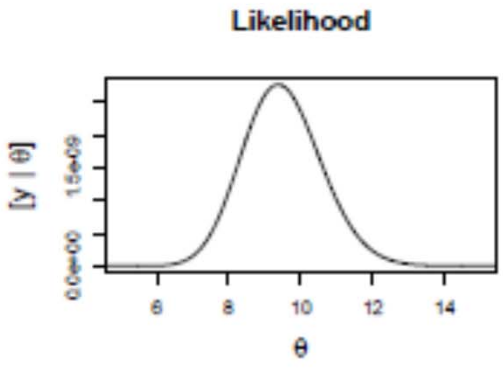


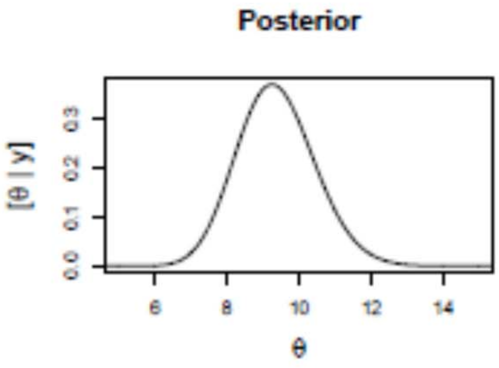
[y] is critical to Bayes

Y=(5,10,11,12,14,9,8,6)

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

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

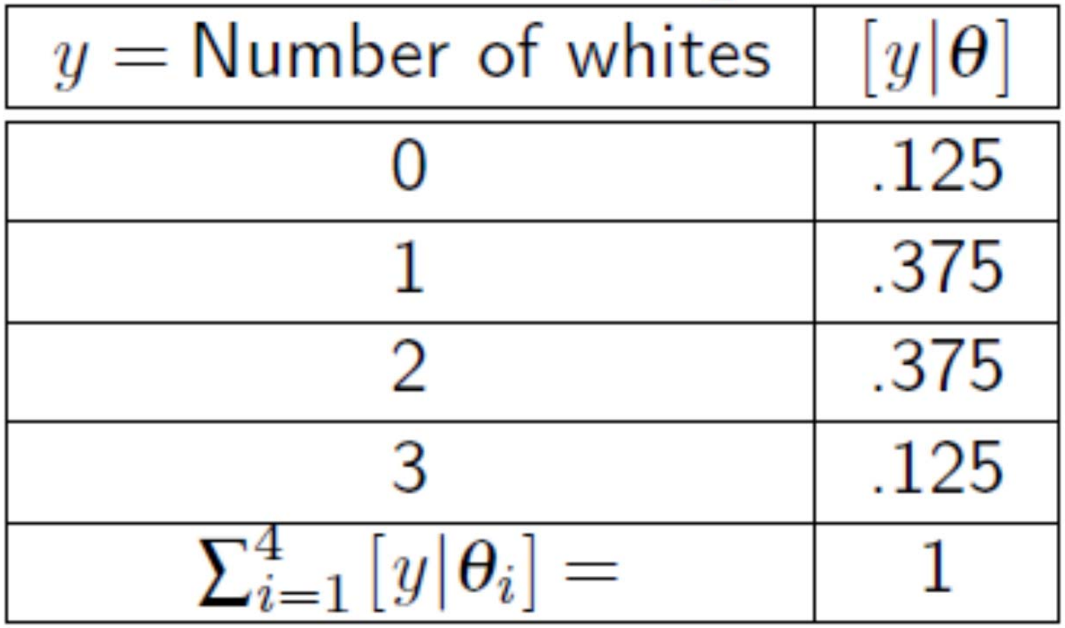




Probability [y|θ]

Θ is known as to be ½. Probability of number of whites conditional on three draws

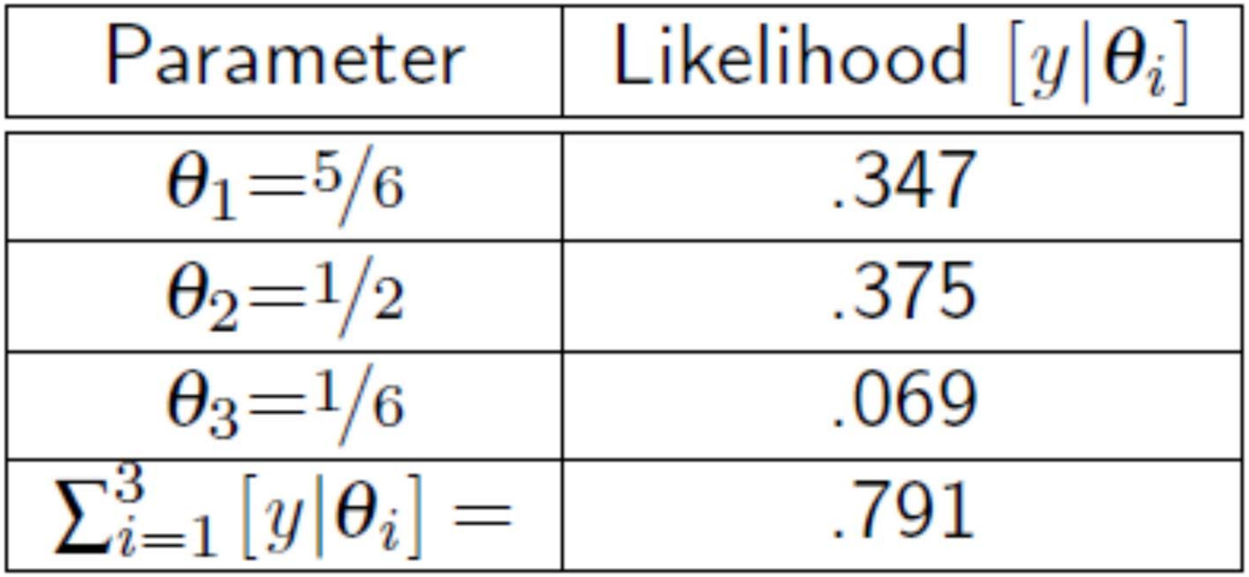
$$p(y \mid n, p) = \binom{n}{y} p^{y} (1 - p)^{n-y}$$



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

Likelihood [y|θ]

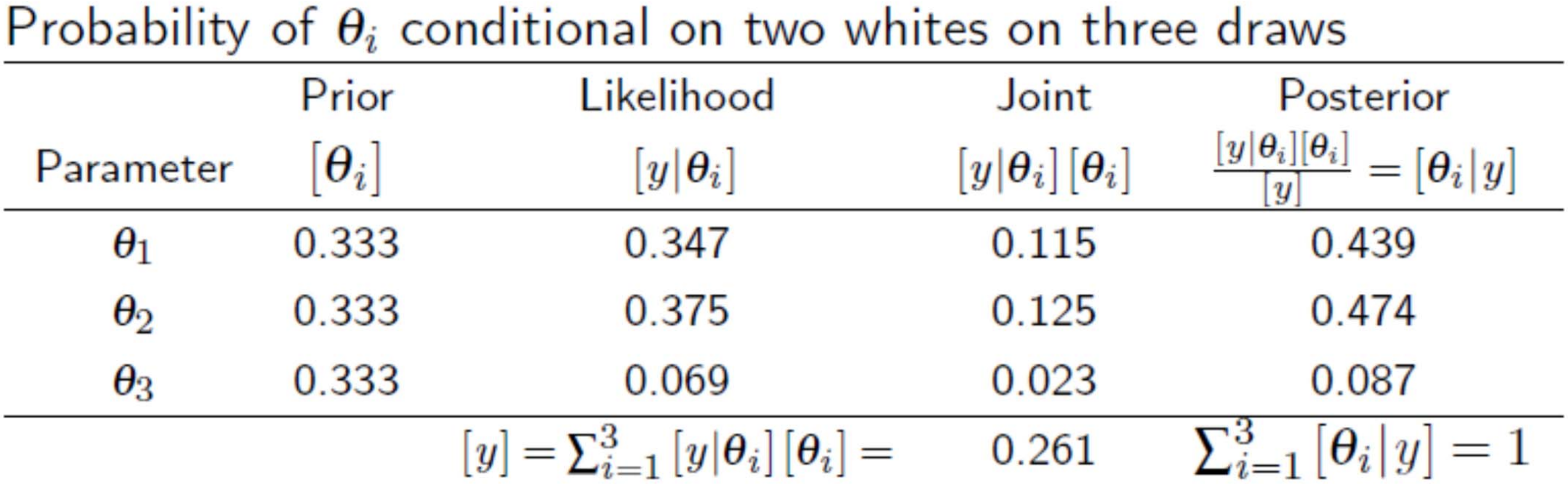
Probability of two whites on three draws conditional on θ_i



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 [θ|y]

Probability of two whites on three draws conditional on θ_i

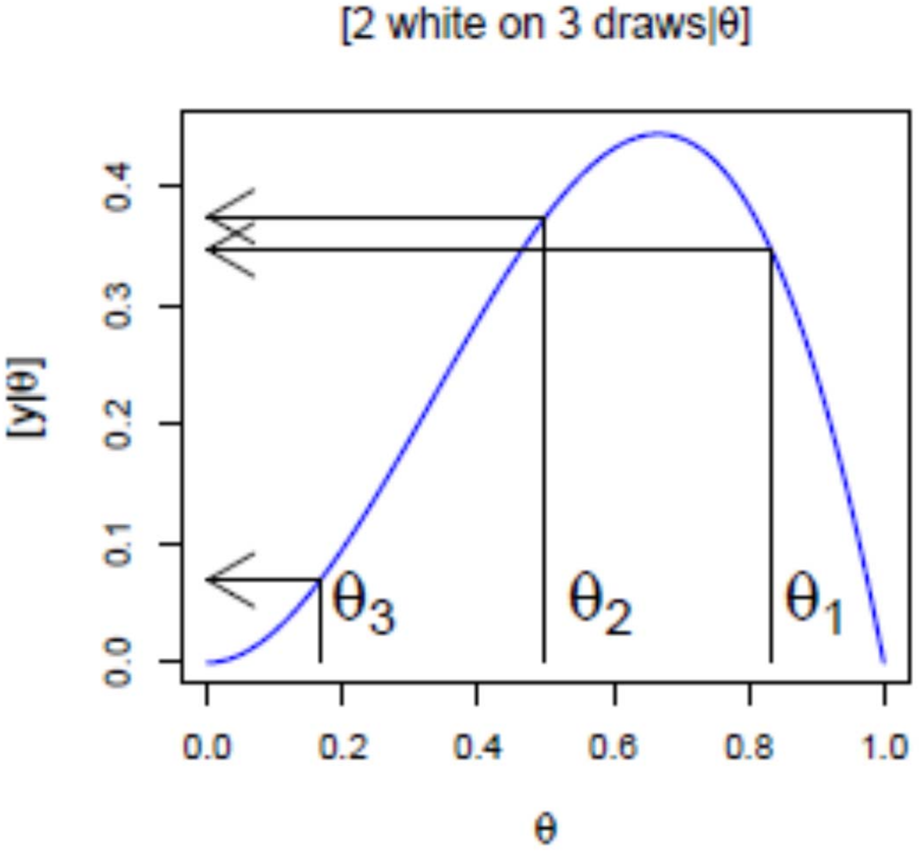


Probability of θ_i conditional on two whites on three draws

	Prior	Likelihood	Joint	Posterior
Parameter	$[\theta_i]$	$[y\|\theta_i]$	$[y\|\theta_i]\,[\theta_i]$	$\frac{[y\|\theta_i][\theta_i]}{[y]} = [\theta_i\|y]$
θ_1	0.333	0.347	0.115	0.439
θ_2	0.333	0.375	0.125	0.474
θ_3	0.333	0.069	0.023	0.087
		$[y] = \Sigma_{i=1}^{3}\,[y\|\theta_i]\,[\theta_i] =$	0.261	$\Sigma_{i=1}^{3}\,[\theta_i\|y] = 1$

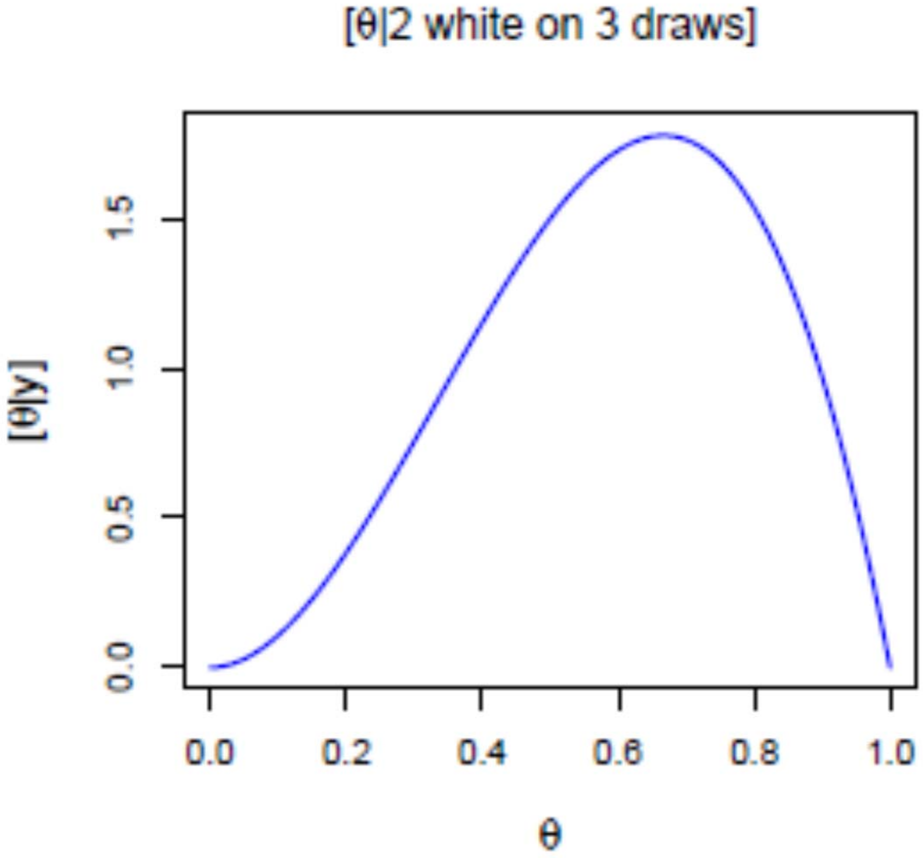
Likelihood profile [y|θ]

[2 white on 3 draws|θ]

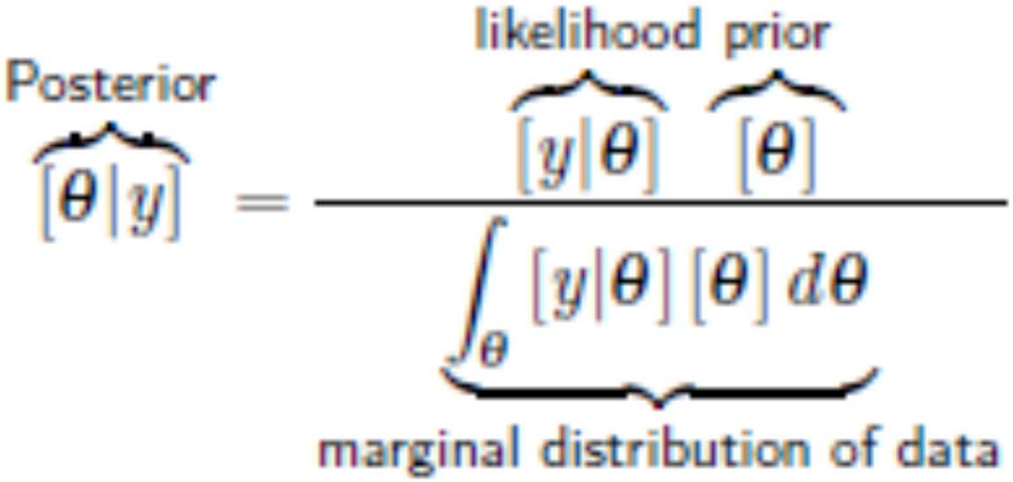


Posterior distribution [θ|y]

[θ| 2 white on 3 draws]

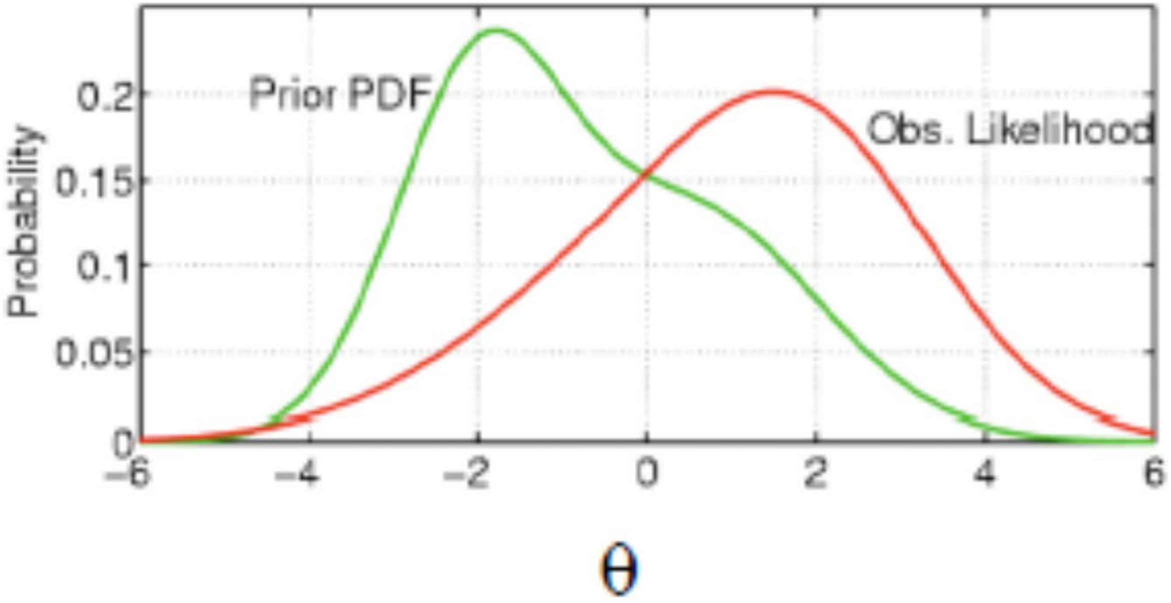


The components of Bayes theorem



$$\overbrace{[\boldsymbol{\theta}|y]}^{\text{Posterior}} = \frac{\overbrace{[y|\boldsymbol{\theta}]}^{\text{likelihood}}\ \overbrace{[\boldsymbol{\theta}]}^{\text{prior}}}{\underbrace{\int_{\boldsymbol{\theta}} [y|\boldsymbol{\theta}]\,[\boldsymbol{\theta}]\,d\boldsymbol{\theta}}_{\text{marginal distribution of data}}}$$

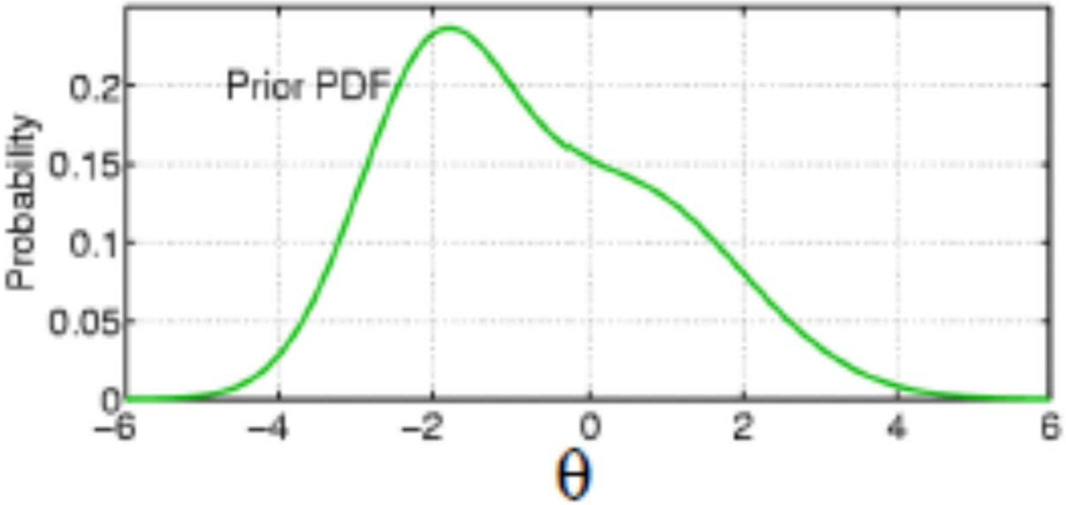
The components of Bayes theorem



Courtesy of Chris Wikle, University of Missouri

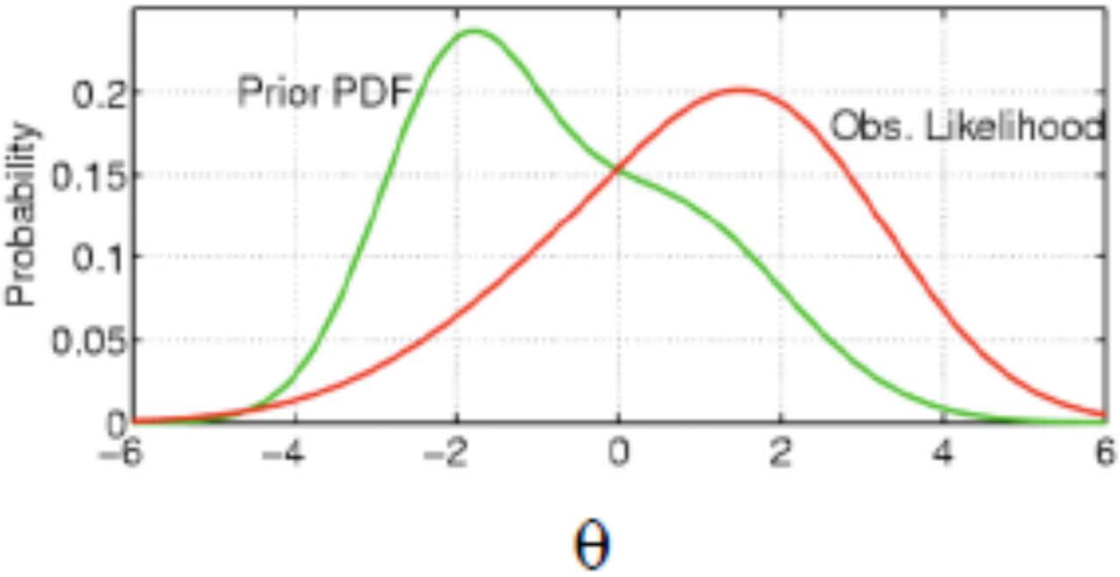
$$[\theta \mid y] = \frac{[y \mid \theta][\theta]}{\int_{\theta} [y \mid \theta][\theta] d\theta}$$

The prior, [θ], can be informative or vague.



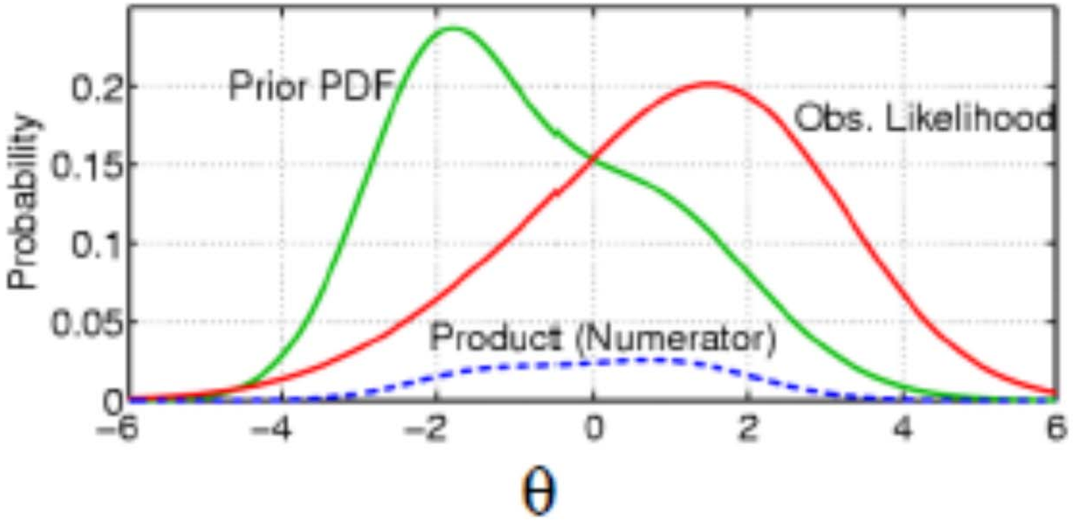
$$[\theta \mid y] = \frac{[y \mid \theta][\theta]}{\int\limits_{\theta} [y \mid \theta][\theta] d\theta}$$

The likelihood (data distribution [y|θ])



$$[\theta \mid y] = \frac{[y,\theta]}{[y]} = \frac{[y \mid \theta][\theta]}{\int_{\theta} [y \mid \theta][\theta] d\theta}$$

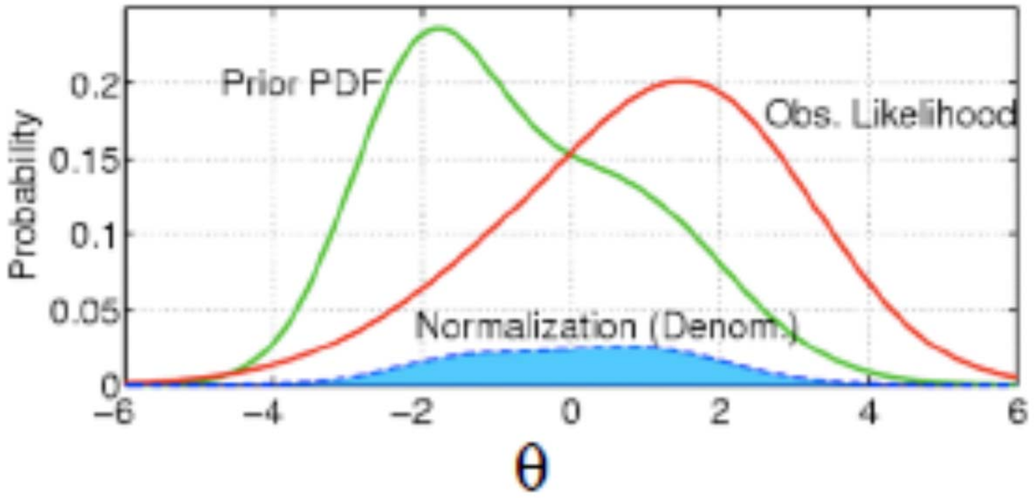
The product of the prior and the likelihood, [y|θ][θ], is the joint distribution of the parameters and the data, [y,θ]



What is the maximum likelihood estimate of θ?

$$[\theta \mid y] = \frac{[y \mid \theta][\theta]}{\int_{\theta} [y \mid \theta][\theta] d\theta}$$

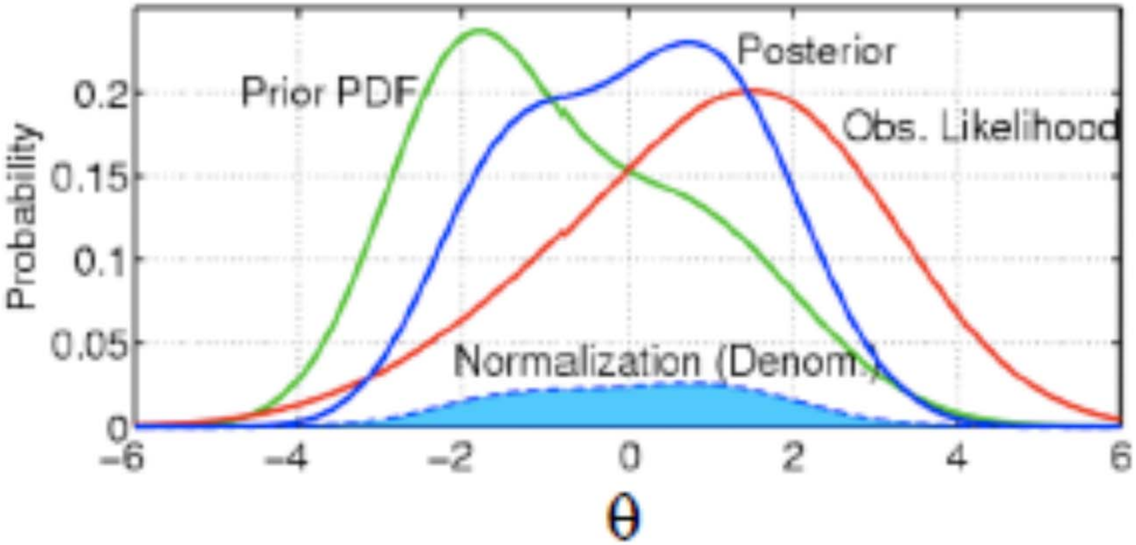
The marginal distribution of the data (the denominator) is the area under the joint distribution.



What are we seeking: The Posterior distribution [θ|y]

$$[\theta \mid y] = \frac{[y \mid \theta][\theta]}{\int_{\theta} [y \mid \theta][\theta] d\theta}$$

- Posterior distribution represents a balance between the information contained in the likelihood and the information contained in the prior distribution.
- An informative prior influences the posterior distribution. A vague prior exerts minimal influence on posterior distribution.



So what

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

  Remember from the basic laws of probability that

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

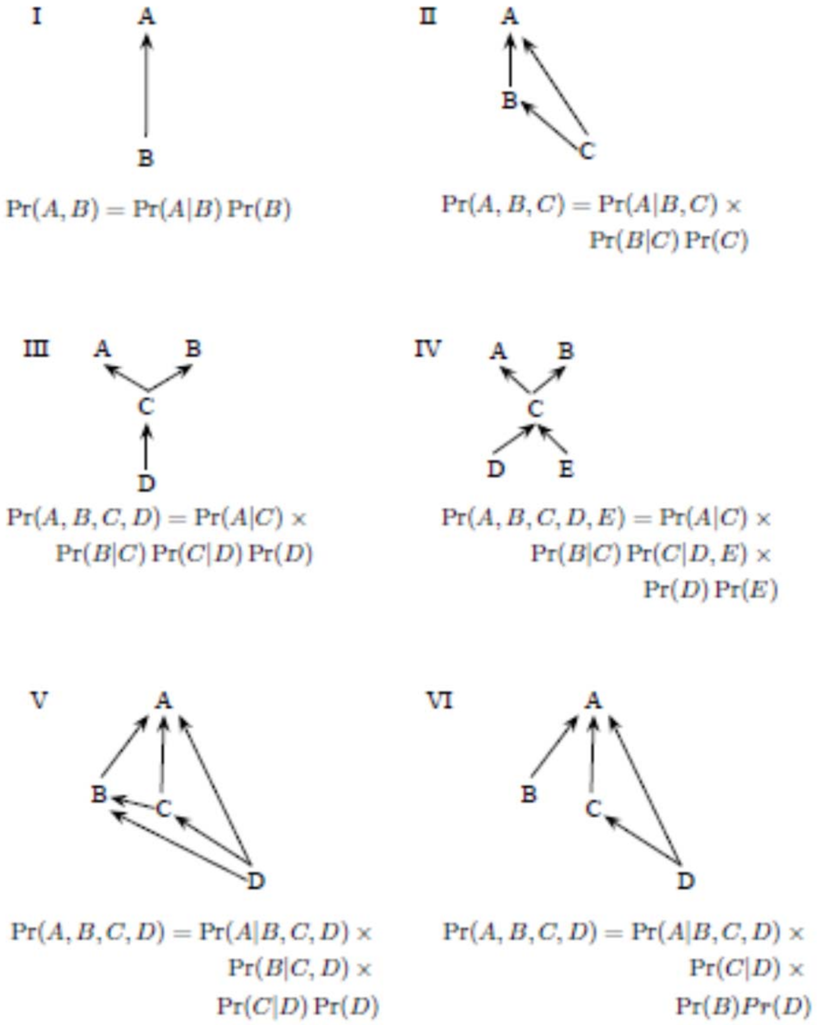
  This generalizes to:

  $z=(z_1, z_2, \ldots, z_n)$

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

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

So what



$$\mathrm{Pr}(A,B) = \mathrm{Pr}(A|B)\,\mathrm{Pr}(B)$$

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

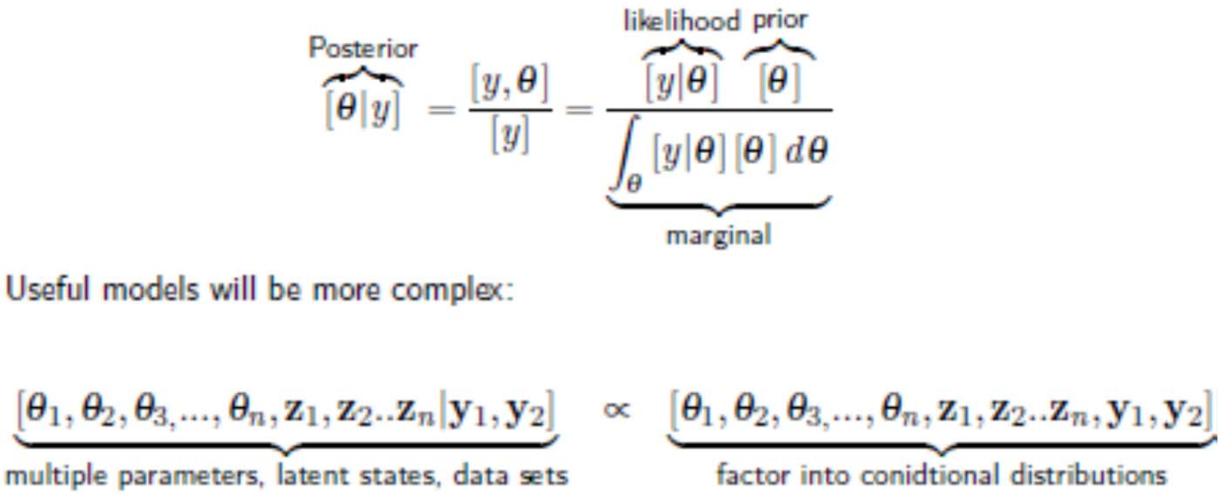
$$\mathrm{Pr}(A,B,C,D) = \mathrm{Pr}(A|C) \times \mathrm{Pr}(B|C)\,\mathrm{Pr}(C|D)\,\mathrm{Pr}(D)$$

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

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

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

So what



$$\overbrace{[\boldsymbol{\theta}|y]}^{\text{Posterior}} = \frac{[y,\boldsymbol{\theta}]}{[y]} = \frac{\overbrace{[y|\boldsymbol{\theta}]}^{\text{likelihood}}\ \overbrace{[\boldsymbol{\theta}]}^{\text{prior}}}{\underbrace{\int_{\theta}[y|\boldsymbol{\theta}]\,[\boldsymbol{\theta}]\,d\boldsymbol{\theta}}_{\text{marginal}}}$$

Useful models will be more complex:

$$\underbrace{[\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_{3,\ldots},\boldsymbol{\theta}_n,\mathbf{z}_1,\mathbf{z}_2..\mathbf{z}_n|\mathbf{y}_1,\mathbf{y}_2]}_{\text{multiple parameters, latent states, data sets}} \propto \underbrace{[\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_{3,\ldots},\boldsymbol{\theta}_n,\mathbf{z}_1,\mathbf{z}_2..\mathbf{z}_n,\mathbf{y}_1,\mathbf{y}_2]}_{\text{factor into conidtional distributions}}$$

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.

Probability distributions

Why do we need to understand the concepts

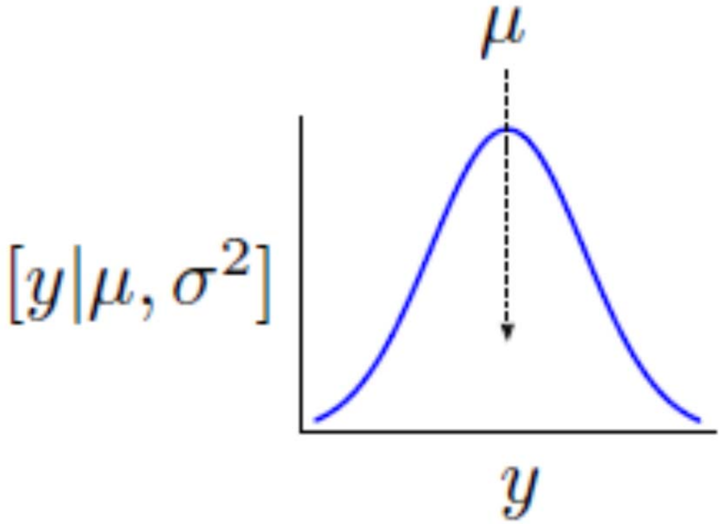
Concept	Why do you need to understand the concept
Conditional probability	It is the foundation for Bayes' theorem and all inferences we will make.
The law of total probability	Basis for the denominator of Bayes' theorem [y].
Factoring joint distributions	This is the procedure we will use to build models.
Independence	Allow us to simplify fully factored joint distributions.
Marginal distributions	Bayesian inference is based on marginal distributions of unobserved quantities.
Statistical distributions	Toolbox representing uncertainty and for understanding unobserved quantities based on observed ones.
Moments	Basis for inference from MCMC.
Moment matching	Allow us to embed the predictions of models into any statistical distribution. Use scientific literature to inform priors.

The essence of Bayes

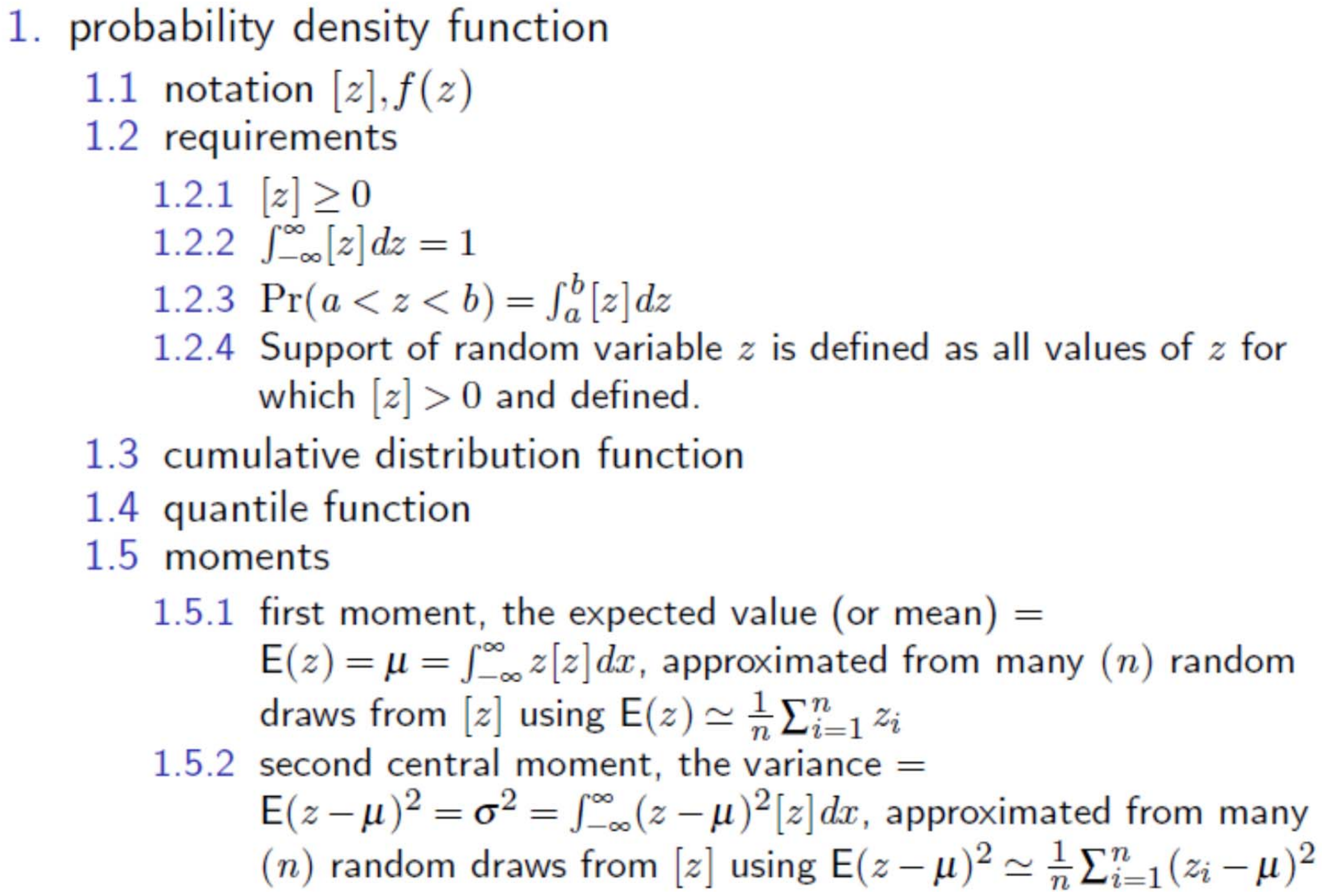
Bayesian analysis is the only branch of statistics that treats all unobserved quantities as random variables. We seek to understand the characteristics of the probability distributions governing the behavior of these random variables.

Linking models to data

$$\mu = g(\theta, x)$$



Probability distributions



1. probability density function
   - 1.1 notation $[z], f(z)$
   - 1.2 requirements
     - 1.2.1 $[z] \geq 0$
     - 1.2.2 $\int_{-\infty}^{\infty} [z] dz = 1$
     - 1.2.3 $\Pr(a < z < b) = \int_a^b [z] dz$
     - 1.2.4 Support of random variable z is defined as all values of z for which $[z] > 0$ and defined.
   - 1.3 cumulative distribution function
   - 1.4 quantile function
   - 1.5 moments
     - 1.5.1 first moment, the expected value (or mean) = $\mathsf{E}(z) = \mu = \int_{-\infty}^{\infty} z[z] dx$, approximated from many (n) random draws from $[z]$ using $\mathsf{E}(z) \simeq \frac{1}{n} \sum_{i=1}^{n} z_i$
     - 1.5.2 second central moment, the variance = $\mathsf{E}(z - \mu)^2 = \sigma^2 = \int_{-\infty}^{\infty} (z - \mu)^2 [z] dx$, approximated from many (n) random draws from $[z]$ using $\mathsf{E}(z - \mu)^2 \simeq \frac{1}{n} \sum_{i=1}^{n} (z_i - \mu)^2$

A familiar approach

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\varepsilon_i \sim normal(0, \sigma^2)$$

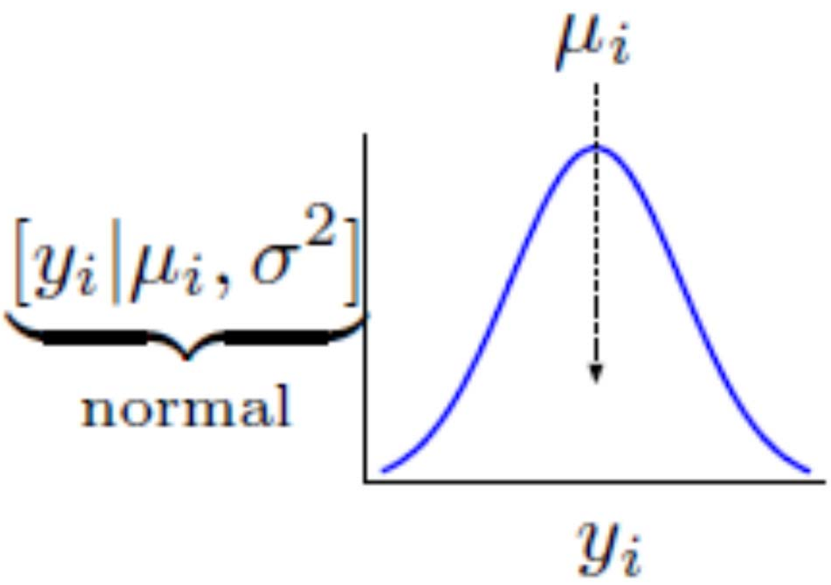
which is identical to

$$\mu_{\mathrm{i}} = \beta_0 + \beta_1 x_i$$

$$y_i \sim normal(\mu_i, \sigma^2)$$

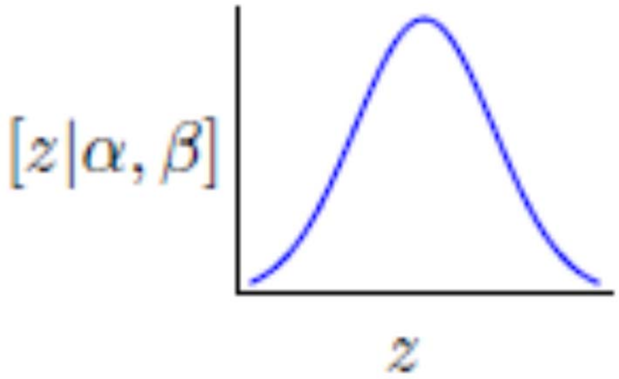
A familiar approach

$$\mu_{\mathrm{i}} = \beta_0 + \beta_1 x_i$$



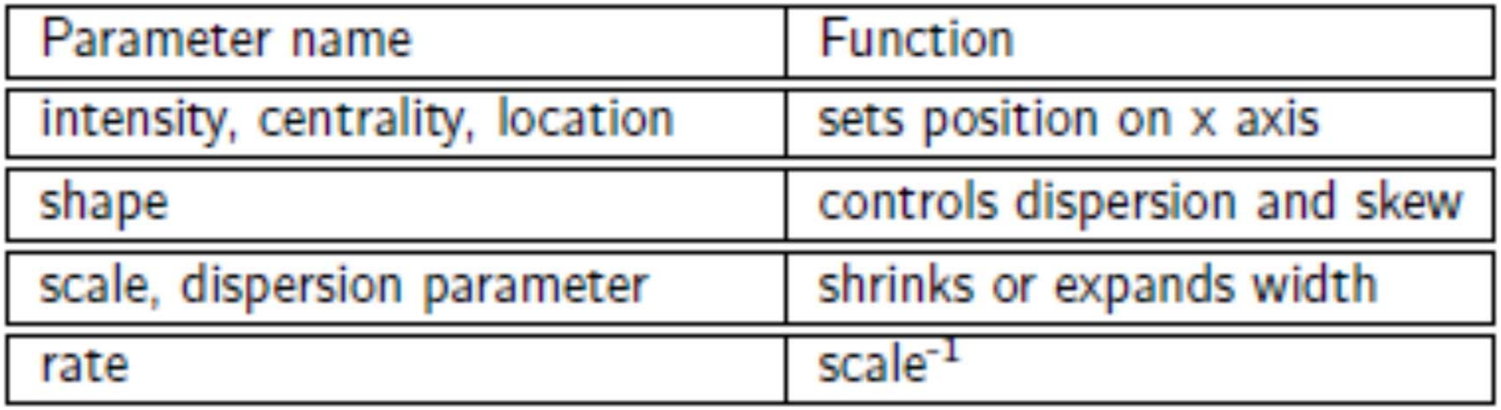
The problem

All distributions have parameters



α and β are parameters of the distribution of the random variable z.

Types of parameters



Parameter name	Function
intensity, centrality, location	sets position on x axis
shape	controls dispersion and skew
scale, dispersion parameter	shrinks or expands width
rate	scale^{-1}

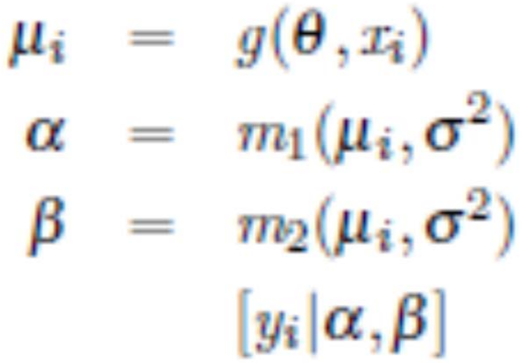
The problem

The normal and the Poisson are the only distributions for which the parameters of the distribution are the same as the moments. For all other distributions, the parameters are functions of the moments.

$$\begin{aligned} \alpha &= m_1(\mu,\sigma^2) \\ \beta &= m_2(\mu,\sigma^2) \end{aligned}$$

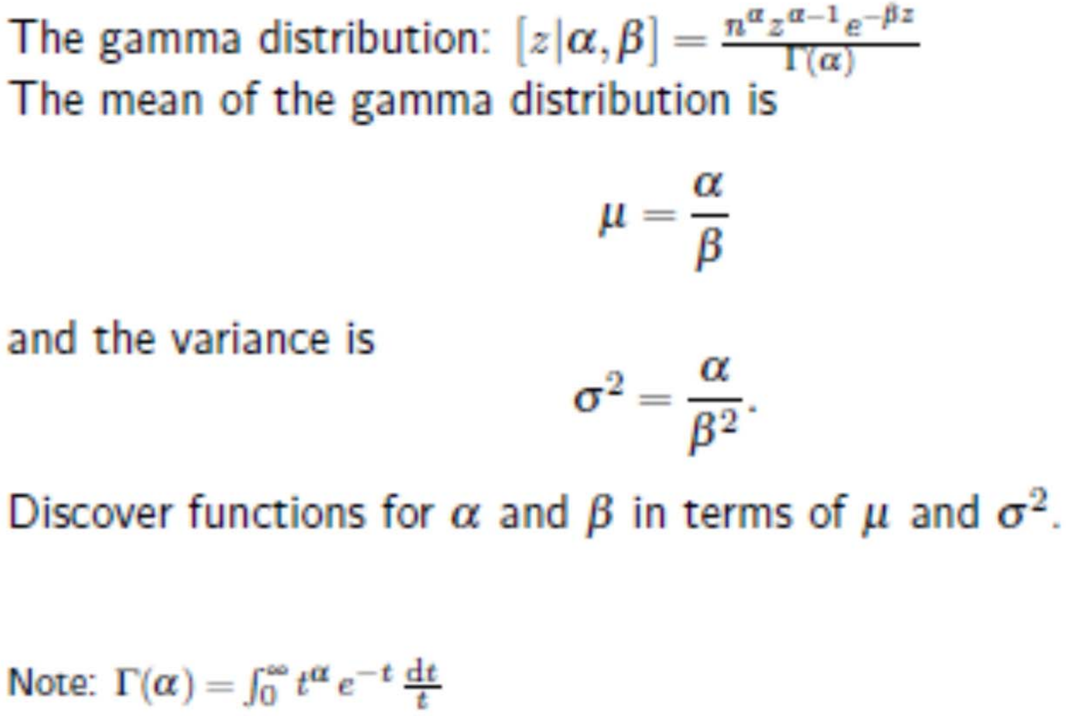
We can use these functions to “match” the moments to the parameters.

Moment matching



$$\begin{aligned}
\boldsymbol{\mu}_i &= g(\boldsymbol{\theta}, x_i) \\
\boldsymbol{\alpha} &= m_1(\boldsymbol{\mu}_i, \boldsymbol{\sigma}^2) \\
\boldsymbol{\beta} &= m_2(\boldsymbol{\mu}_i, \boldsymbol{\sigma}^2) \\
& [y_i|\boldsymbol{\alpha}, \boldsymbol{\beta}]
\end{aligned}$$

Moment matching the gamma distribution



The gamma distribution: $[z|\alpha,\beta] = \frac{n^{\alpha} z^{\alpha-1} e^{-\beta z}}{\Gamma(\alpha)}$
The mean of the gamma distribution is

$$\mu = \frac{\alpha}{\beta}$$

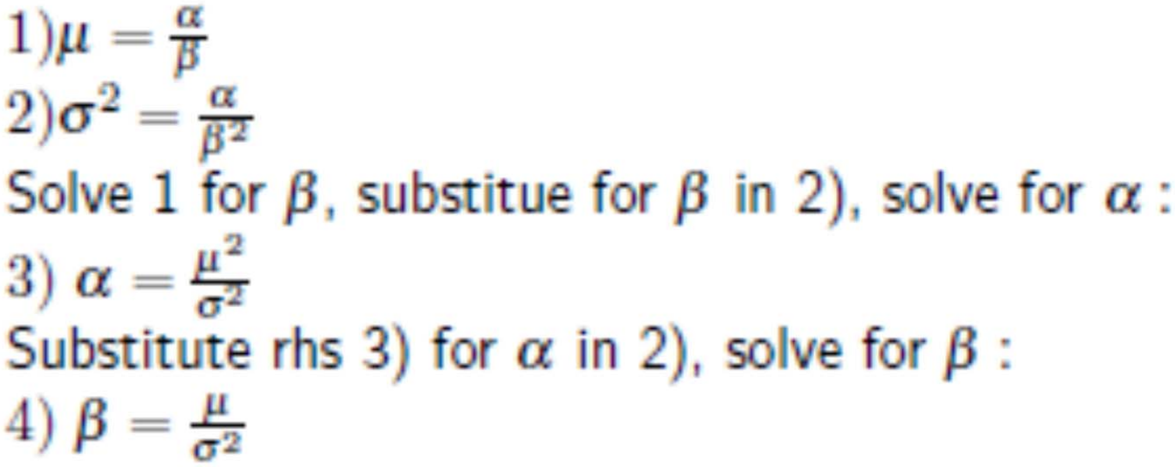
and the variance is

$$\sigma^2 = \frac{\alpha}{\beta^2}.$$

Discover functions for α and β in terms of μ and σ^2.

Note: $\Gamma(\alpha) = \int_0^\infty t^\alpha e^{-t} \frac{dt}{t}$

Answer



1)$\mu = \frac{\alpha}{\beta}$

2)$\sigma^2 = \frac{\alpha}{\beta^2}$

Solve 1 for β, substitue for β in 2), solve for α :

3) $\alpha = \frac{\mu^2}{\sigma^2}$

Substitute rhs 3) for α in 2), solve for β :

4) $\beta = \frac{\mu}{\sigma^2}$

Moment matching the beta distribution

The beta distribution gives the probability density of random variables with support on 0, …, 1.

$$[z|\alpha,\beta] = \frac{z^{\alpha-1}(1-z)^{\beta-1}}{B(\alpha,\beta)}$$

$$B = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

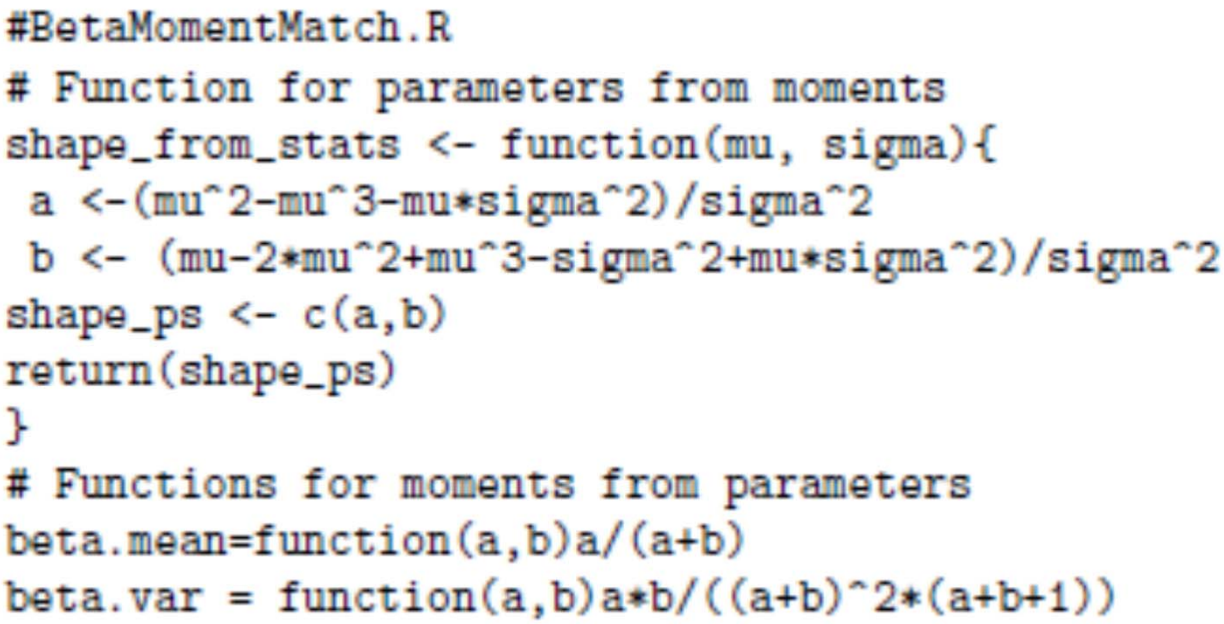
$$\mu = \frac{\alpha}{\alpha+\beta}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$\alpha = \frac{\mu^2 - \mu^3 - \mu\sigma^2}{\sigma^2}$$

$$\beta = \frac{\mu - 2\mu^2 + \mu^3 - \sigma^2 + \mu\sigma^2}{\sigma^2}$$

You need some functions



```
#BetaMomentMatch.R
# Function for parameters from moments
shape_from_stats <- function(mu, sigma){
 a <-(mu^2-mu^3-mu*sigma^2)/sigma^2
 b <- (mu-2*mu^2+mu^3-sigma^2+mu*sigma^2)/sigma^2
shape_ps <- c(a,b)
return(shape_ps)
}
# Functions for moments from parameters
beta.mean=function(a,b)a/(a+b)
beta.var = function(a,b)a*b/((a+b)^2*(a+b+1))
```

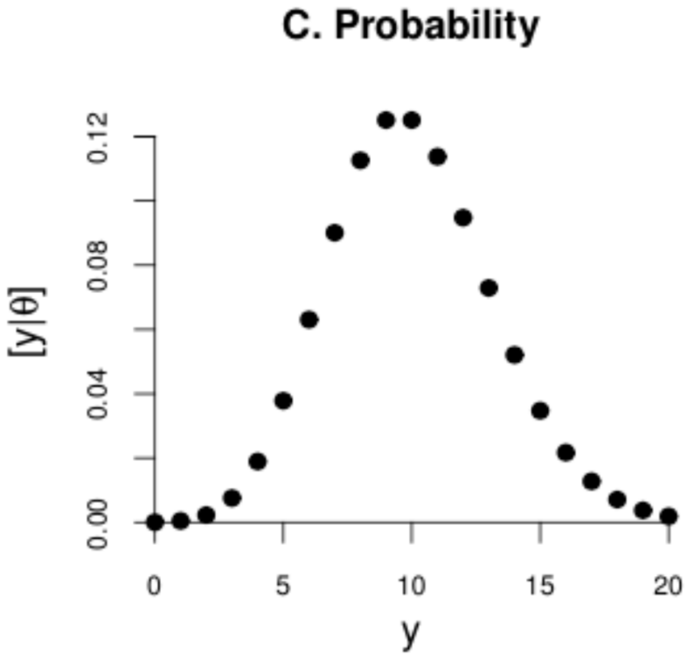
likelihood

Likelihood forms the fundamental link between models and data in a Bayesian framework. In addition, ***maximum likelihood*** *is a widely used alternative to Bayesian methods for estimating parameters in socio-ecological models.*

-Hobbs and Hooten 2015

Probability functions

For a discrete random variable, Y, the probability that the random variable Y takes on a specific value y is a probability function.



Tadpole example

You collect data on the number of tadpoles per volume of water in a pond.
You observe 14 tadpoles in one litter sample.

You know the true average number of tadpoles per liter of water to b 23.

The probability of your data is

$[y_i|\lambda] =$

In this example, what did we treat as fixed and what did we treat as random?

In this example, what did we treat as fixed and what did we treat as random?

Parameter values (λ or θ) are fixed and the data (y) are random.

What if, instead, you want to know the likelihood of the parameter given the observed data

The evaluation can be accomplished using a likelihood function $L(\theta|y)$

What is a likelihood function

$L(\theta|y) = [y|\theta]$

$L(\theta|\mathbf{y}) = \prod_{i=1}^{n} [y_i \mid \theta]$

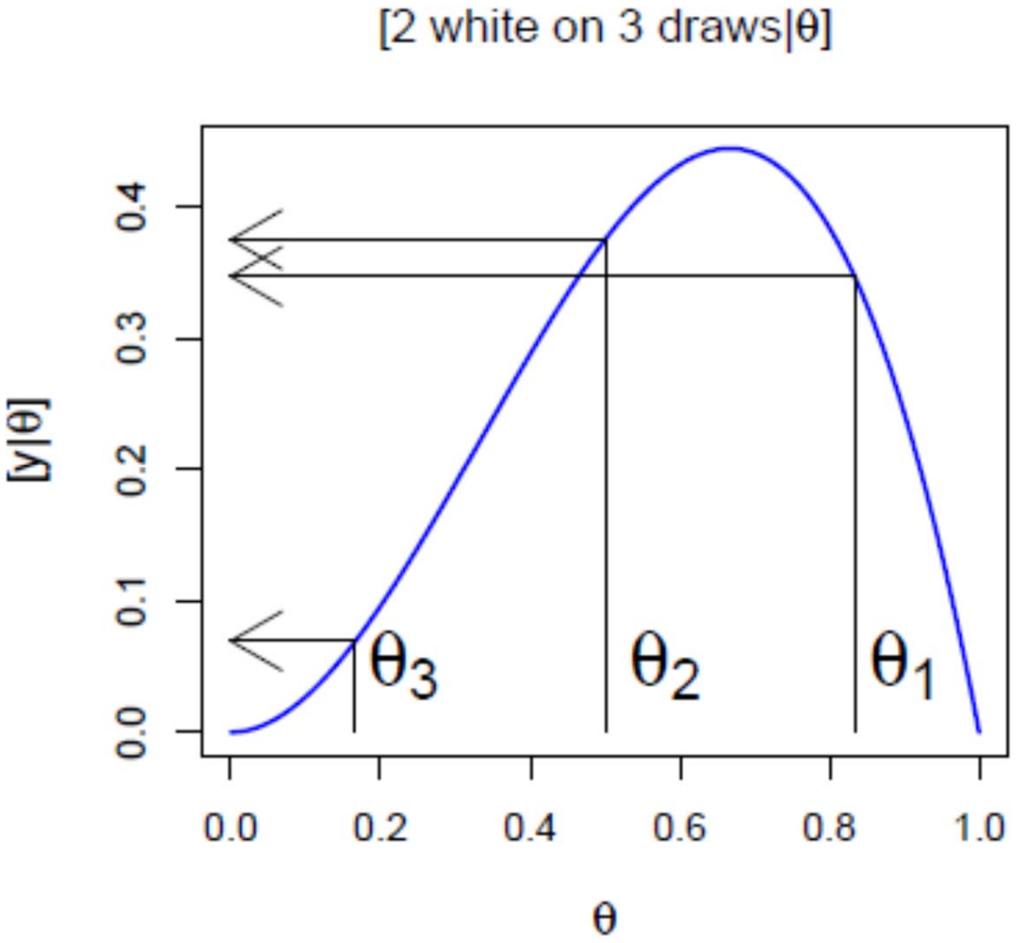
What is a likelihood function

$$\underbrace{\mathrm{L}(\theta|\boldsymbol{y})}_{\text{``Likelihood Function"}} = \overbrace{\prod_{i=1}^{n}[y_i|\theta]}^{\text{``Likelihood Model" or ``Data Model"}}$$

Parameter	Likelihood $[y \vert \theta_i]$
θ_1	.347
θ_2	.375
θ_3	.069
$\Sigma_{i=1}^{3}$	.791

Table 1: Probability of two whites on three draws conditional on θ_i

Likelihood profile



Likelihood as a concept

- Likelihood is the [y|θ].

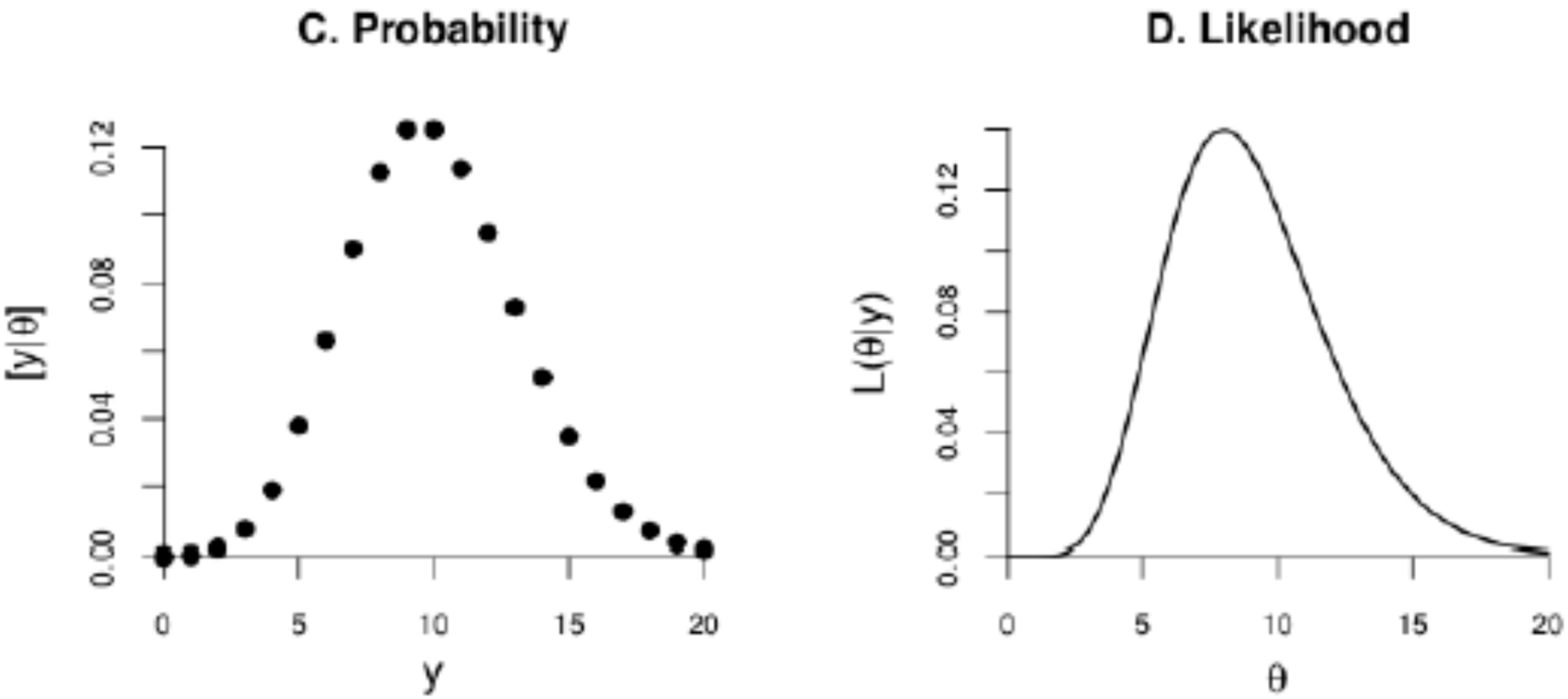
Likelihood as a concept

- Likelihood is the [y|θ].
- Likelihood is the chance of observing your data given θ.

Likelihood as a concept

- Likelihood is the [y|θ].
- Likelihood is the chance of observing your data given θ.
- Likelihood is the probability of observing your data conditional on your hypothesis θ.

What are the main differences between a PMF (or PDF) and a likelihood profile?



What are the main differences between a PMF (or PDF) and a likelihood profile?

Probability density/mass

- Data are treated as random variables.

Likelihood

What are the main differences between a PMF (or PDF) and a likelihood profile?

Probability density/mass

- Data are treated as random variables.
- Parameters are fixed.

Likelihood

What are the main differences between a PMF (or PDF) and a likelihood profile?

Probability density/mass

- Data are treated as random variables.
- Parameters are fixed.
- Areas under the curve = 1

Likelihood

- Data are fixed.

What are the main differences between a PMF (or PDF) and a likelihood profile?

Probability density/mass

- Data are treated as random variables.
- Parameters are fixed.
- Areas under the curve = 1

Likelihood

- Data are fixed.
- Parameters are varied.

What are the main differences between a PMF (or PDF) and a likelihood profile?

Probability density/mass

- Data are treated as random variables.
- Parameters are fixed.
- Areas under the curve = 1

Likelihood

- Data are fixed.
- Parameters are varied.
- Areas under the curve ≠1.

What are the main differences between a PMF (or PDF) and a likelihood profile?

Probability density/mass

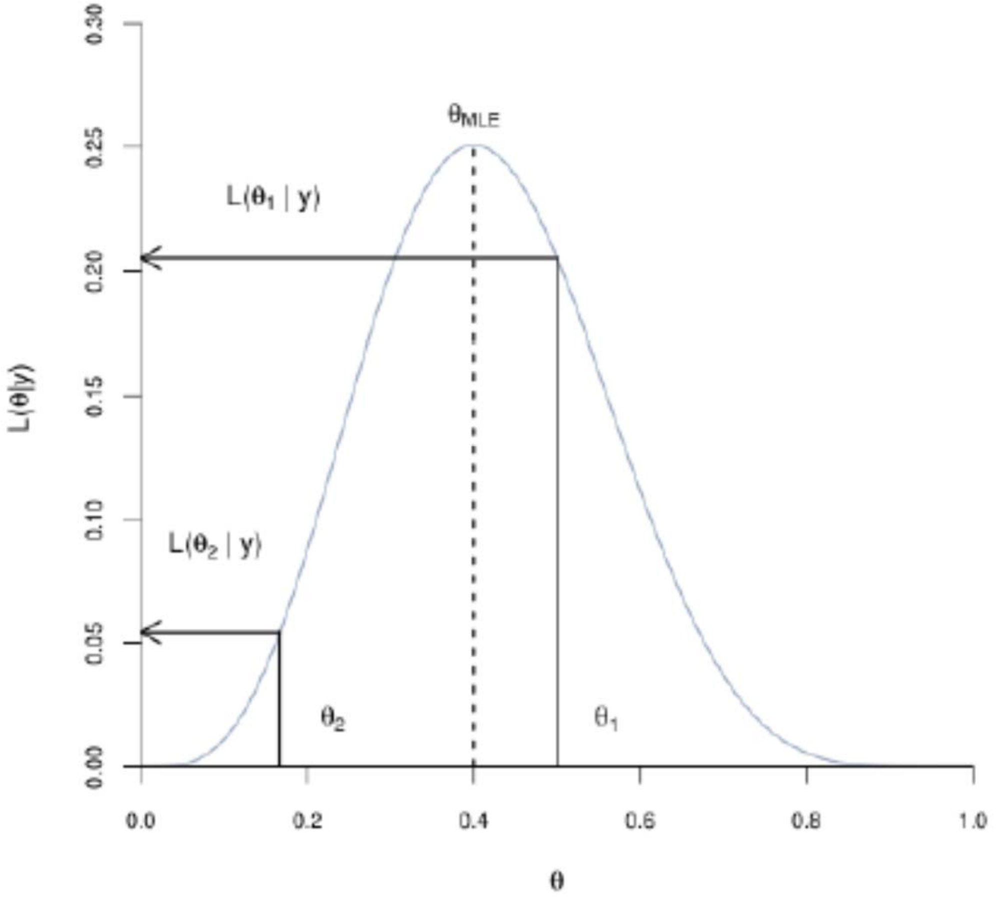
- Data are treated as random variables.
- Parameters are fixed.
- Areas under the curve = 1

Likelihood

- Data are fixed.
- Parameters are varied.
- Areas under the curve ≠1.
- Y-axis values are arbitrary and scalable.

Understanding the likelihood profile

What is the meaning of any one point on the likelihood profile curve?



Maximum likelihood

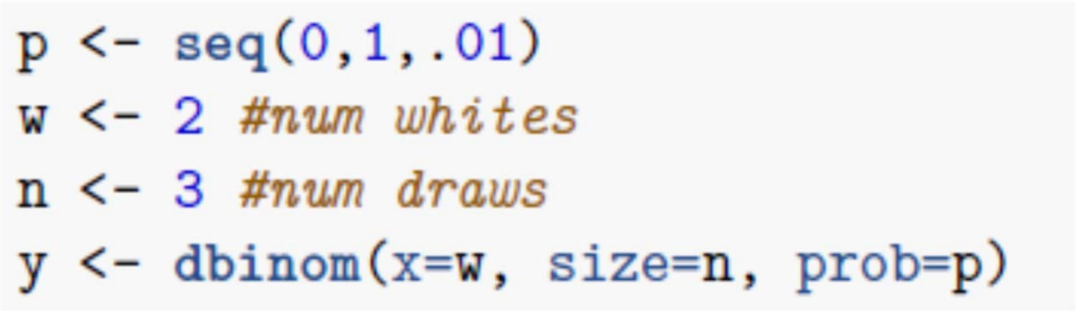
Knowing the likelihood of a specific parameter value doesn’t tell us anything useful in the absence of a comparison value. Therefore the evidence provided by data is expressed as the likelihood ratio.

$$\frac{L(\theta_1 \mid y)}{L(\theta_2 \mid y)} = \frac{[y \mid \theta_1]}{[y \mid \theta_2]}$$

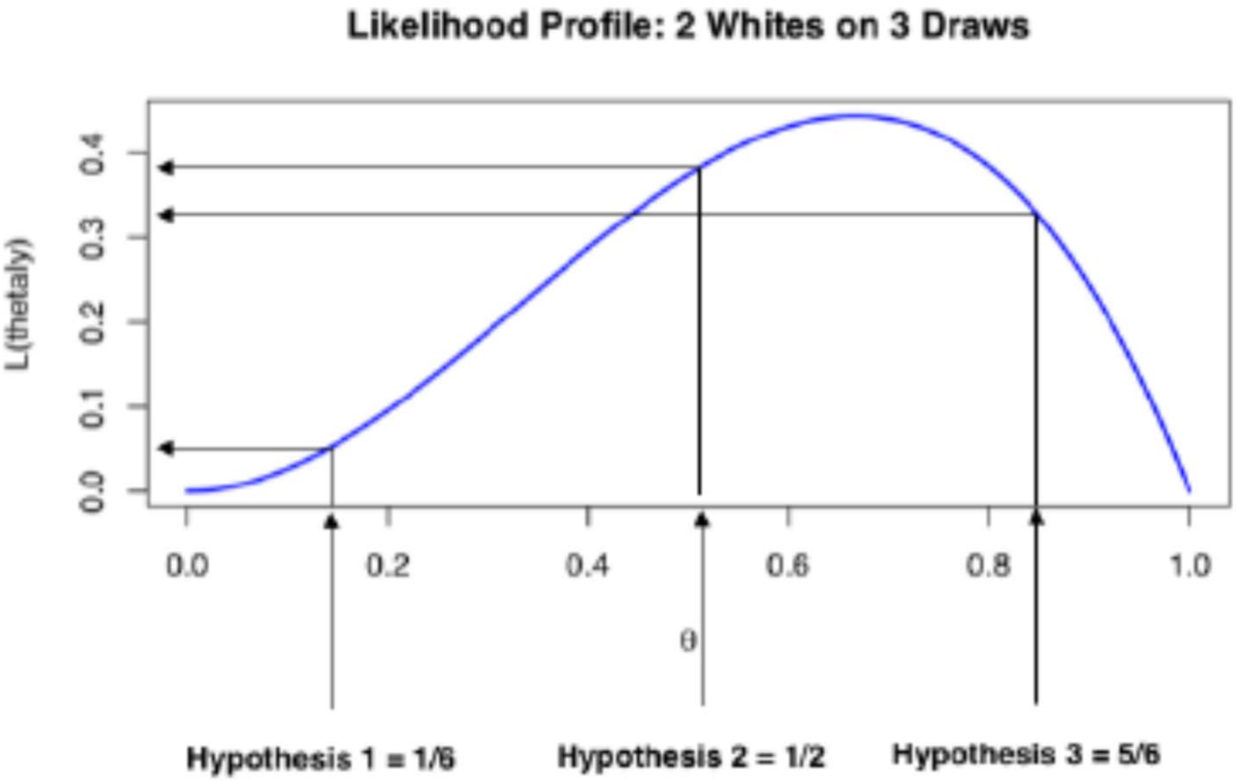
Practically, we often want to know the value of parameter θ that has the maximum support in the data, which is the peak of the likelihood profile. This is the value of θ that maximizes the likelihood function.

Likelihood example

Consider we have a jar of white and black beans and want to estimate the probability, p, of choosing a white bean. We draw 3 beans and 2 are white. Plot the probability of the data conditional on θ as a function of all possible θ.



```
p <- seq(0,1,.01)
w <- 2 #num whites
n <- 3 #num draws
y <- dbinom(x=w, size=n, prob=p)
```


Likelihood Profile: 2 Whites on 3 Draws
L(theta|y)
0.0
0.1
0.2
0.3
0.4
0.0
0.2
0.4
0.6
0.8
1.0
θ
Hypothesis 1 = 1/6
Hypothesis 2 = 1/2
Hypothesis 3 = 5/6