Likelihood

Bayesian Modeling for Socio-Environmental Data

Chris Che-Castaldo, Mary B. Collins, N. Thompson Hobbs

May 2018

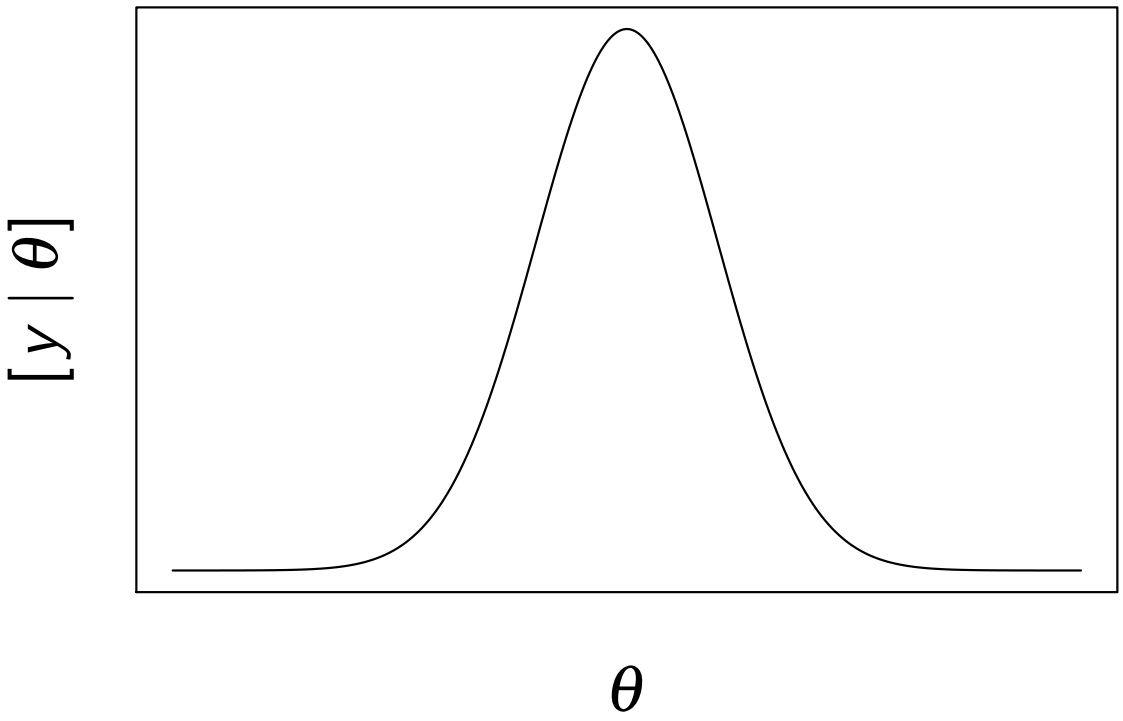
Why Likelihood?

- Likelihood is a component of all Bayesian models.
- Maximum likelihood is an important statistical approach in its own right.

Learning objectives for lecture

- Understand the concept of likelihood and its relationship to the probability of data conditional on parameters.
- Describe a likelihood profile and how it differs from a probability density function.
- Understand how to compose likelihoods for multiple parameters and multiple observations.

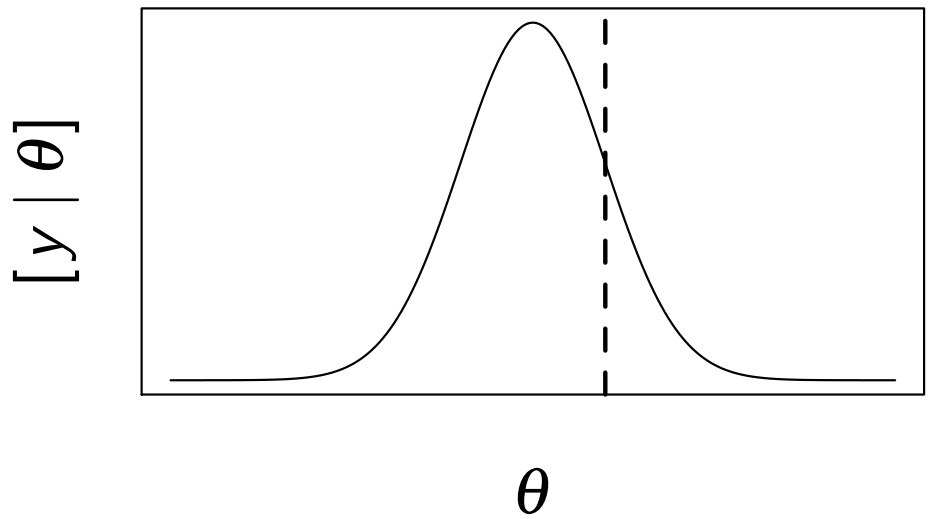
Inference from likelihood is based on $[y \mid \theta]$



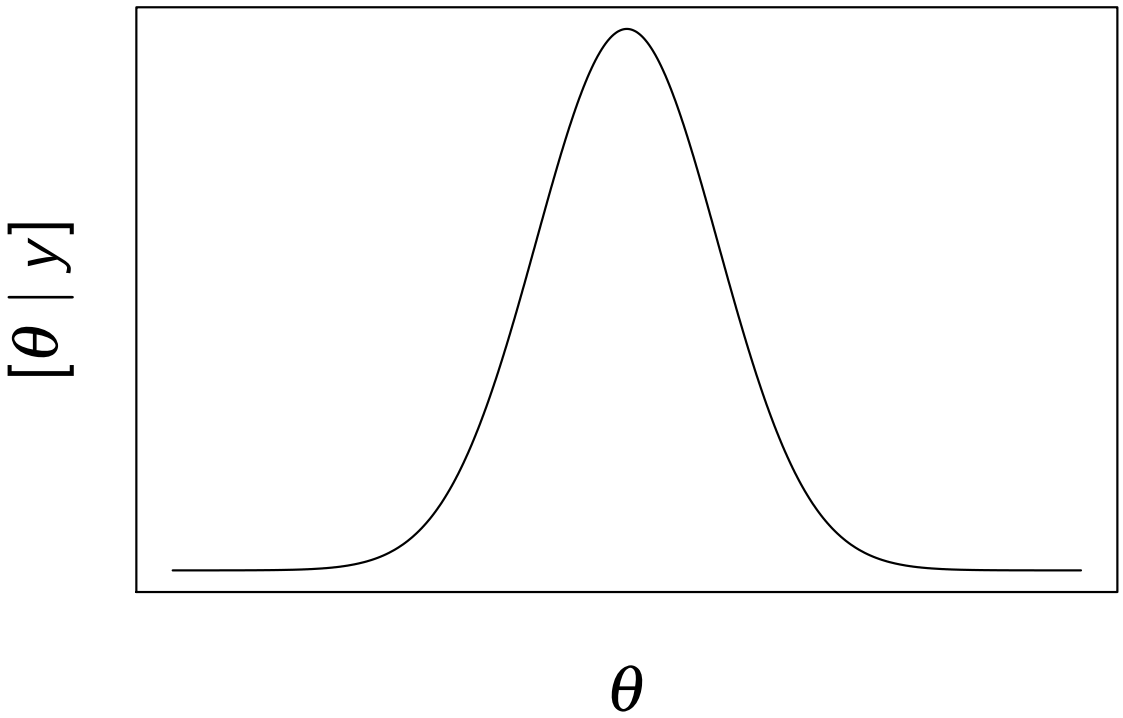
Likelihood allows us to compare alternative parameter values and models by calculating the probability of the data conditional on the parameters $[y \mid \theta]$. As you will see, all evidence based on likelihood is relative.

Data conditional on a parameter: $[y \mid \theta]$

Prevalence is a term used in disease ecology to indicate the proportion of a population that is infected. The true prevalence, θ, of chronic wasting disease in male mule deer on the winter range north of Fort Collins is 12%. A sample of 24 male deer includes 4 infected individuals. What is the probability of obtaining these data if the estimate of prevalence is true?



Bayesian inference is based on $[\theta \mid y]$



Parameter conditional the data: $[\theta \mid y]$

We obtain a new sample of 24 male mule deer on the winter range north of Fort Collins that includes 4 infected individuals. In light of these data, what is the probability that the true value of prevalence, θ, is found in $q_L \leq \theta \leq q_U$?

The key idea in likelihood

- In a probability mass or probability density function, the parameter θ is constant (known) and the data $\boldsymbol{y}$ are random variables. The function sums or integrates to 1 over its support.
- In a likelihood function, the data are constant (known) and the parameter is unknown but fixed. We use $[y \mid \theta]$ to assess the likelihood of different values of θ in light of the data. In this case, the function does not sum or integrate to one over all possible values of the parameter.

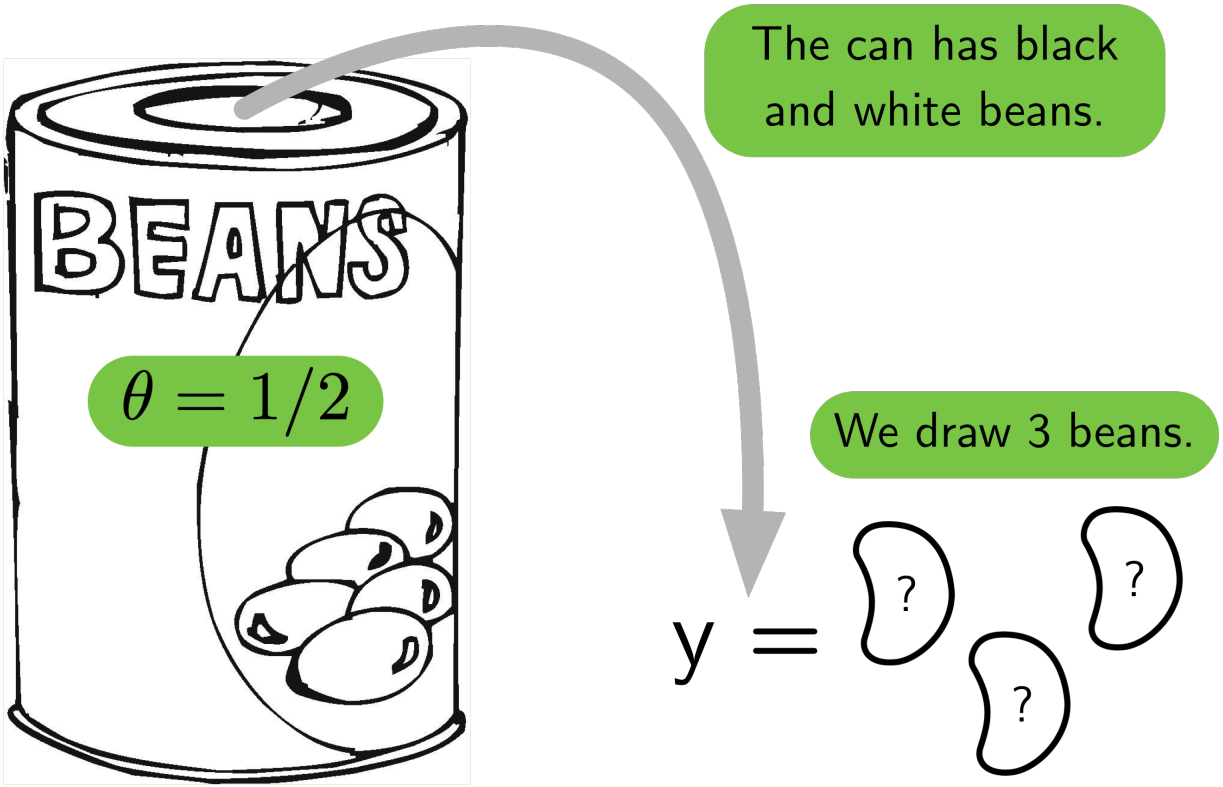
$$\underbrace{L(\theta \mid y)}_{\text{likelihood function}} \propto \underbrace{[y \mid \theta]}_{\text{PDF or PMF}}$$

Likelihood is *proportional* to probability or probability density

Discuss notation

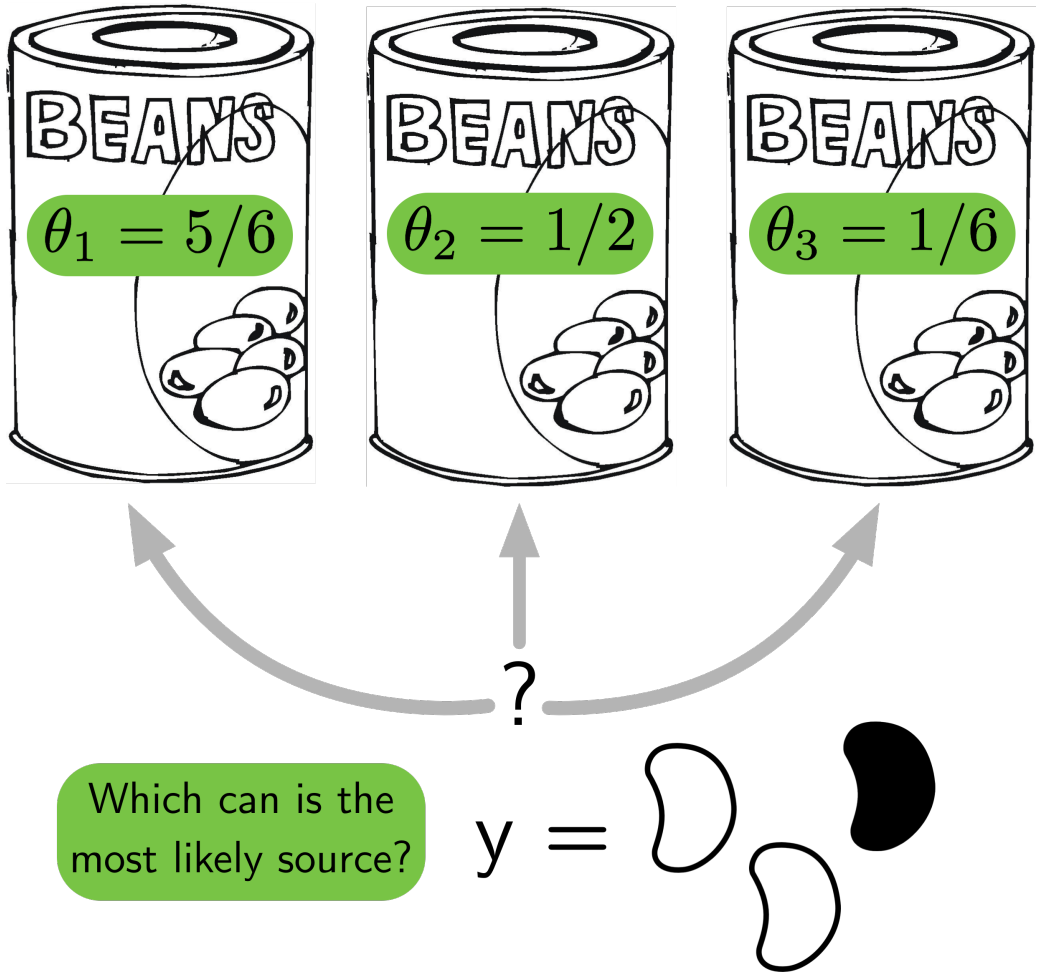
$$L(\theta \mid y) \propto [y \mid \theta] \tag{1}$$
$$L(\theta \mid y) = c[y \mid \theta] \tag{2}$$
$$L(\theta \mid y) = [y \mid \theta] \tag{3}$$

The parameter is known and the data are random



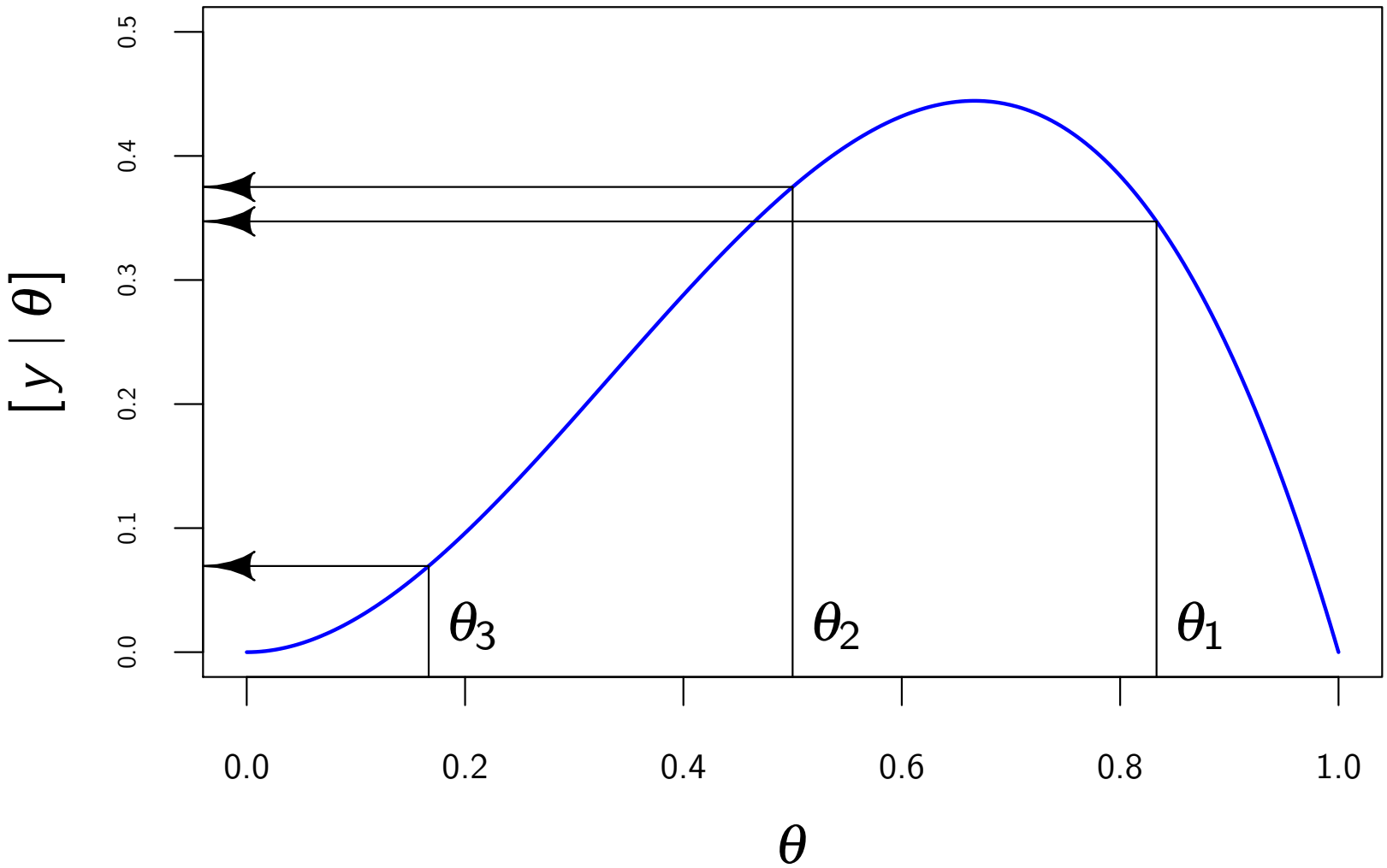
- What are the possible outcomes?
- What probability mass function would you use to model theses data?
- What is the probability of each outcome?
- What is the sum of the individual probabilities?

The data are known and the parameter is random



- What is the likelihood of each parameter value?

A likelihood profile: 2 white beans on 3 draws

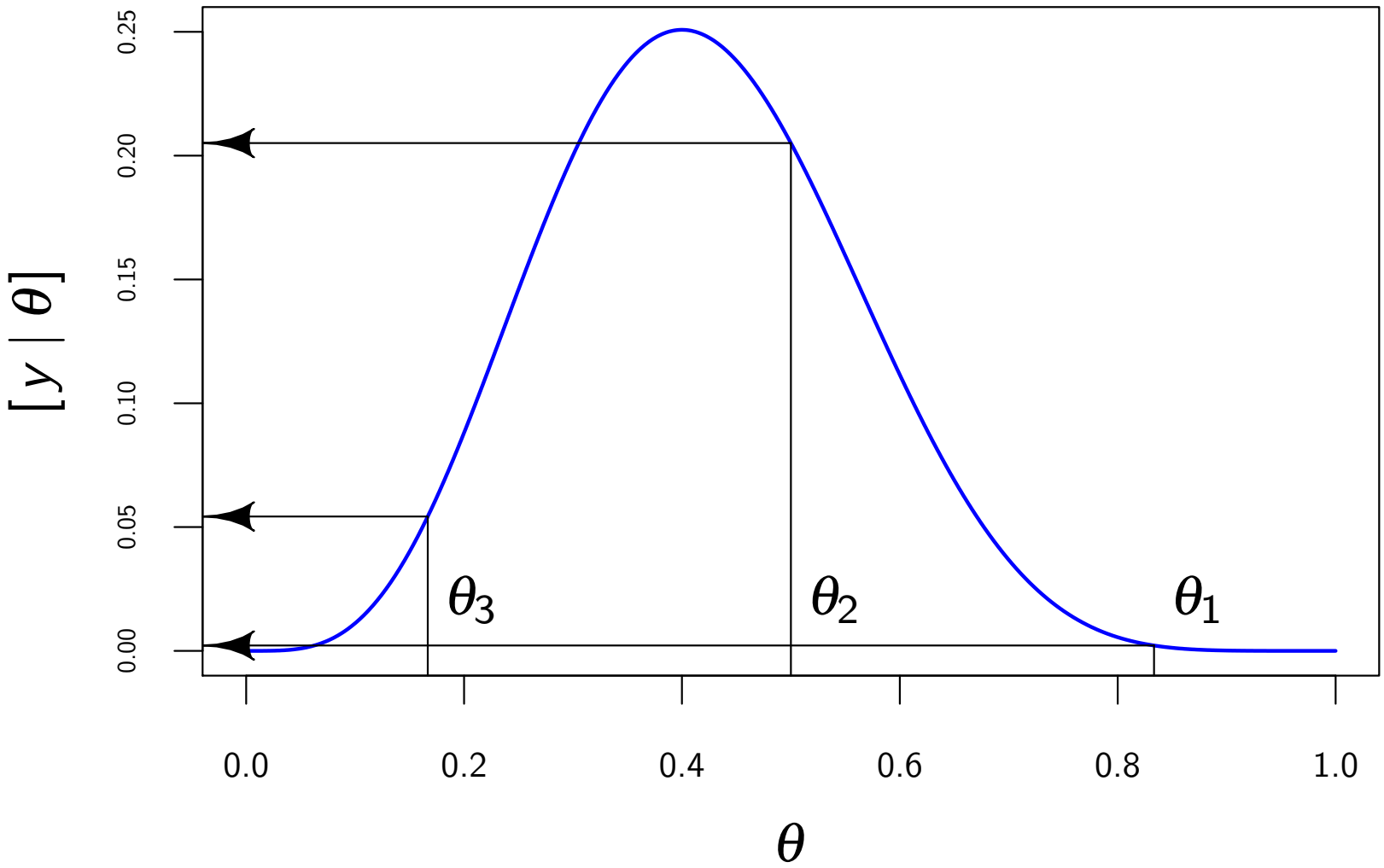


Class exercise: Can of beans

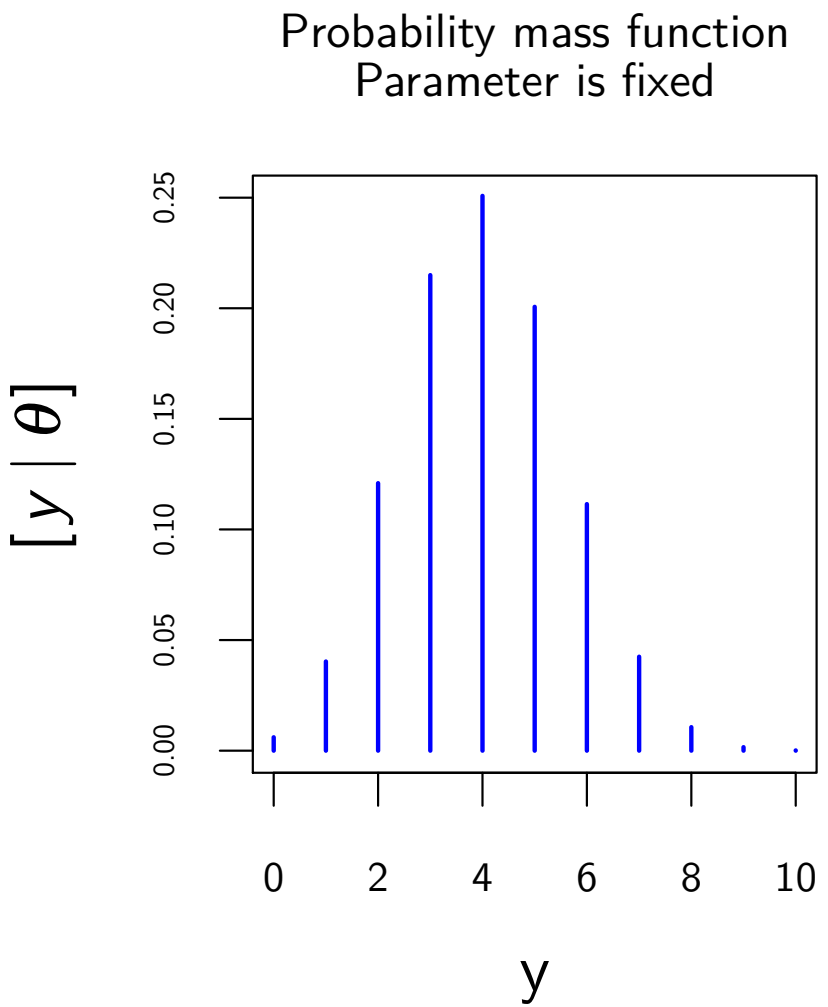
Have someone take a draw of 10 beans from one of the cans where the identity of the can is unknown.

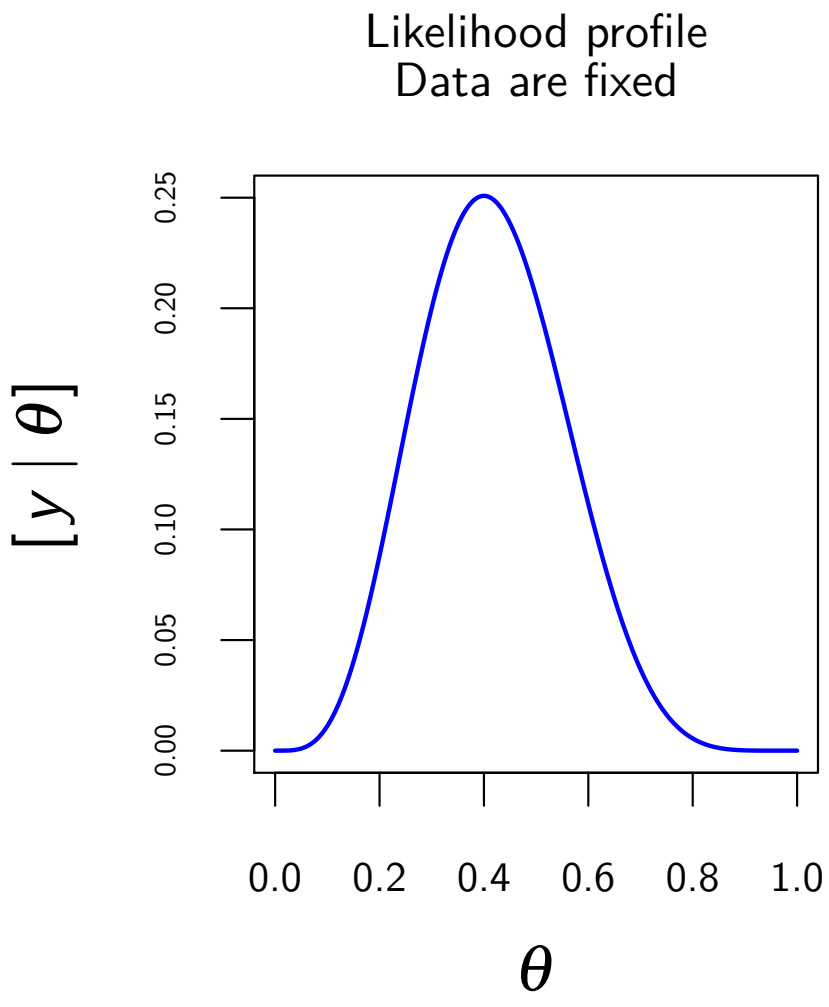
- Which of the three cans is the most likely to have produced this draw?
- How much more likely is this can than the other two?

A likelihood profile: 4 white beans on 10 draws



Likelihood vs Probability:





How do we fit models with multiple parameters?

In the example we had a single parameter, θ, one observation of 4 successes on 10 draws, and a binomial likelihood. However, we could have made the likelihood a function of the *predictions* of a model, and used any probability mass function or probability density function as a "wrapper" for the predictions, i.e.,

$$\begin{aligned} \mu_i &= g(\boldsymbol{\theta}, x_i) \\ \mathcal{L}(\mu_i, \sigma^2 \mid y_i) &\propto \underbrace{[y_i \mid \mu_i, \sigma^2]}_{\text{PDF or PMF}} \end{aligned}$$

Likelihood Surfaces

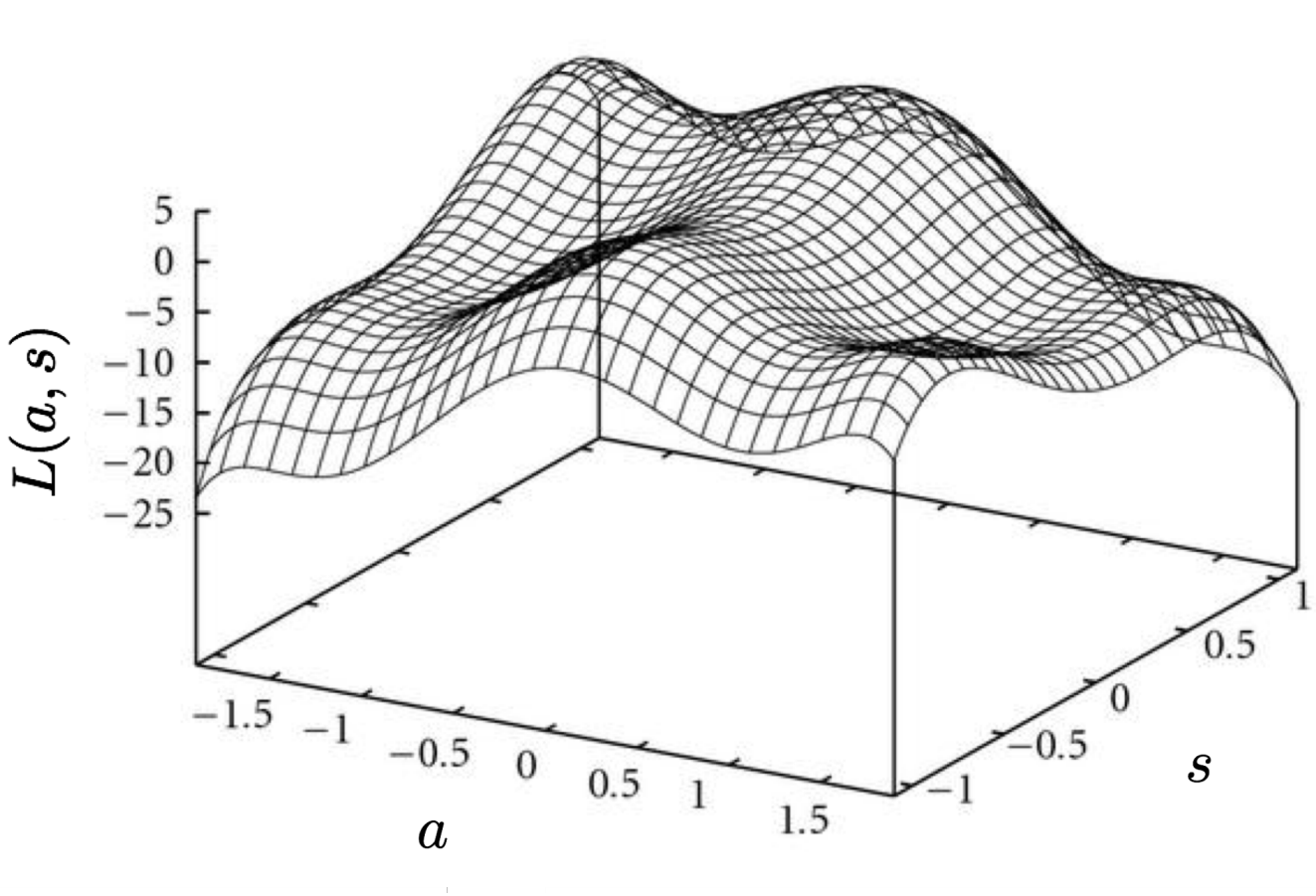


Figure courtesy of Sergiy Nesterko.

How do we fit models with multiple parameters and multiple data points?

The total likelihood is the product of the individual likelihoods, assuming the data are *conditionally independent*:

$$\mathcal{L}(\boldsymbol{\theta}, \sigma^2 \mid \mathbf{y}) \quad = \quad c \prod_{i=1}^{n} [y_i \mid g(\boldsymbol{\theta}, x_i), \sigma^2]$$

What does conditionally independent mean?

Independence is an assumption! Remember from the chain rule:

$$Pr(y_1, \ldots, y_n) = Pr(y_1 \mid y_2 \ldots y_n) Pr(y_2 \mid y_3 \ldots y_n) \ldots Pr(y_n).$$

However, by assuming that these random variables are independent, you can simplyify the joint probability into:

$$Pr(y_1, \ldots, y_n) = Pr(y_1) Pr(y_2) \ldots Pr(y_n),$$

such that the total likelihood a product of the individual likelihoods.

What does conditionally independent mean?

We evaluate the independence assumption by examining the residuals (ε) from a model, where ($\varepsilon_i = y_i - g(\boldsymbol{\theta}, x_i)$).

The independence assumption holds if knowing a residual tells you nothing about the other residuals.

We assess this by ensuring that the residuals:

- do not show a trend, meaning they should be centered on 0 throughout the range of fitted values,
- and are not be autocorrelated.

Log likelihoods:

We often use the sum of the log likelihoods to get the total log likelihood as a basis for fitting models:

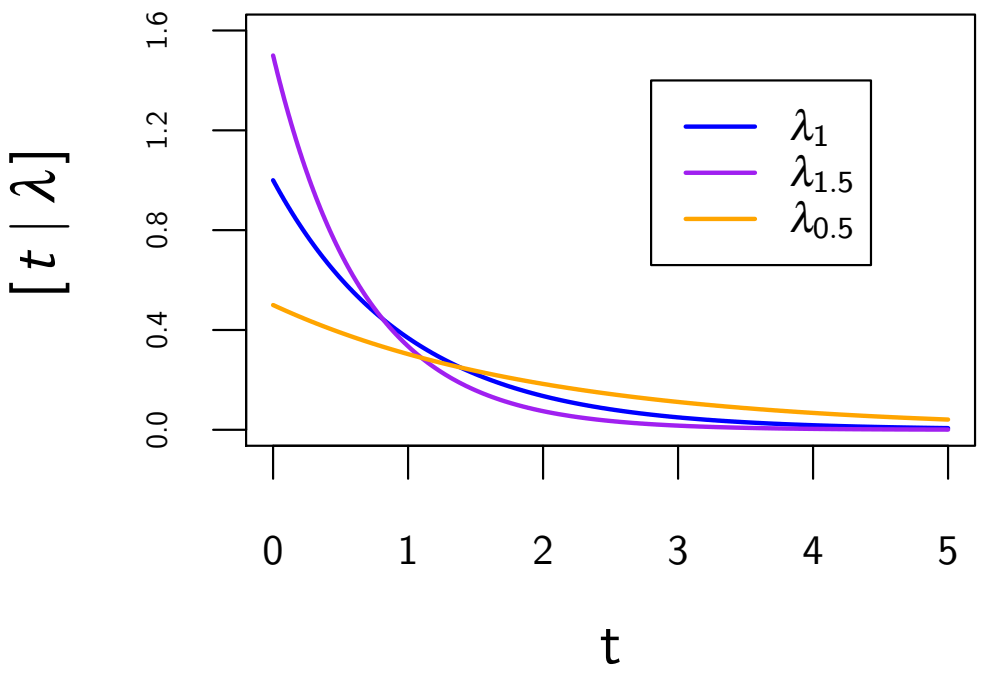
$$\log(\mathcal{L}(\boldsymbol{\theta}, \sigma^2 \mid y)) = L(\boldsymbol{\theta}, \sigma^2 \mid y) = \log(c) + \sum_{i=1}^{n} \log([y_i \mid g(\boldsymbol{\theta}, x_i), \sigma^2])$$

The exponential distribution

$$t_i \sim \text{exponential}(\lambda)$$

$$P(t_i \mid \lambda) = \begin{cases} \lambda e^{-\lambda t_i} & t_i \geq 0 \\ 0 & t_i < 0 \end{cases}$$

if λ is the average number of events time^{-1}, then $1/\lambda$ is the average time between events and mean of the exponential distribution.



- The data, t_i, represent "waiting times" between events happening in a Poisson process. If the number of events per unit time is provided by the Poisson distribution, then the length of time between events is provided by the exponential distribution.
- See this link by John C.B. Cooper for a clear and concise explanation of the connection between these two distributions.

Likelihood exercise for time to tweet

Generate a datum quantifying how long we have to wait for a new tweet from the POTUS. Write out the likelihood function assuming wait times for POTUS tweets are governed by an exponential distribution. Determine the maximum likelihood estimate (MLE) for λ.

Generate the datum to estimate λ for POTUS

- $\frac{991}{151} \sim 6.6$ tweets per day
- $t_1 = \frac{1}{6.6} \sim 0.15$ days between each tweet
- Fun fact: 3.7 hours respite between POTUS tweets!

Analytically solve for λ_{MLE} for POTUS

$$\mathcal{L}(\lambda) = [t_1 \mid \lambda]$$ The likelihood equals the PDF.

$$\mathcal{L}(\lambda) = \lambda e^{-\lambda t_1}$$ Substitude the exponential PDF.

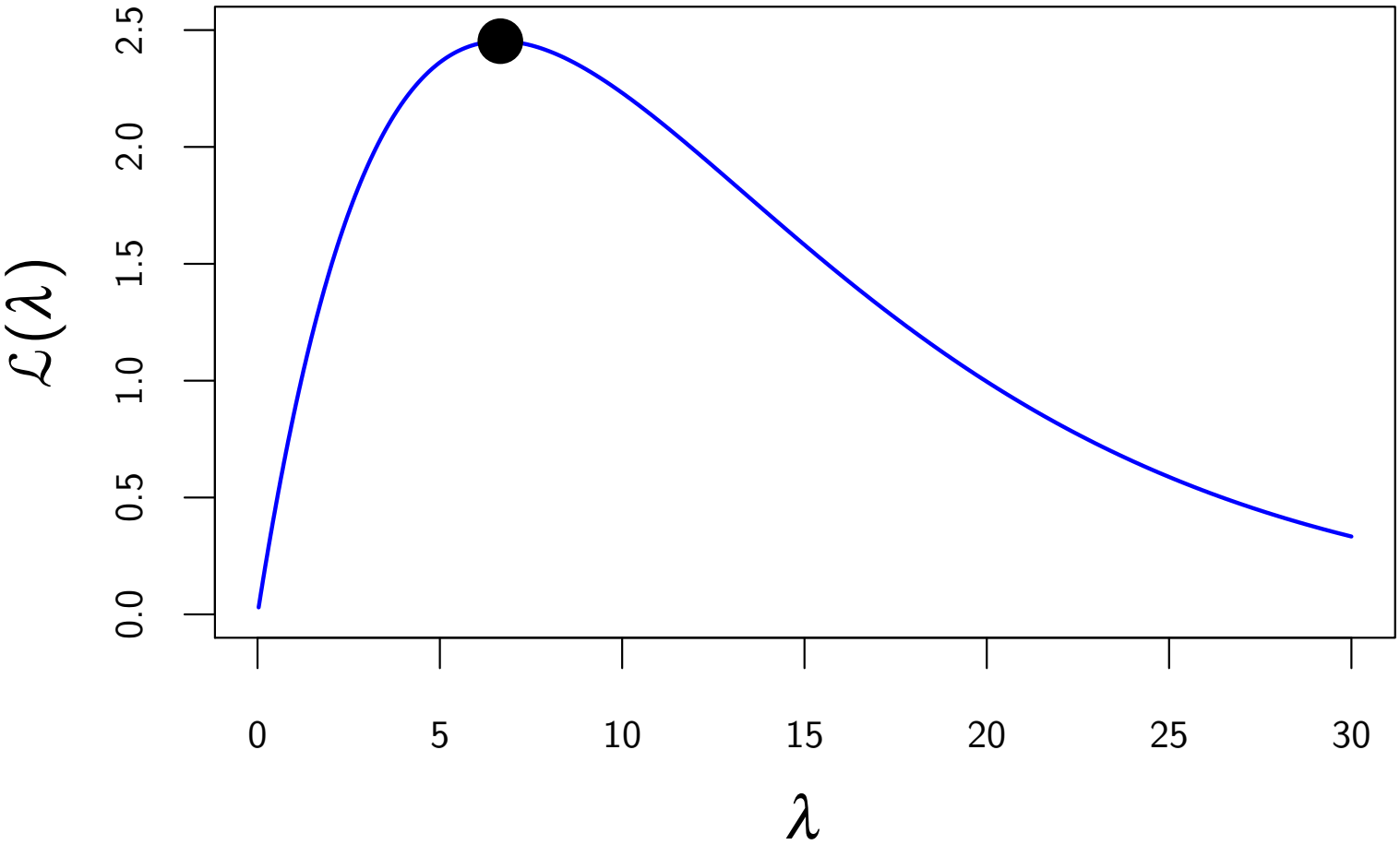
$$L(\lambda) = \log(\lambda) - \lambda t_1$$ Take logs to ease differentiation.

$$\frac{dL(\lambda)}{d\lambda} = \frac{1}{\lambda} - t_1$$ Differentiate.

$$0 = \frac{1}{\lambda_{MLE}} - t_1$$ Set derivative $= 0$ and solve for λ_{MLE}.

$$\lambda_{\text{MLE}} = \frac{1}{0.15} = 6.6$$ Note that λ_{MLE} = Poisson rate parameter.

Likelihood profile of λ for POTUS



Tweeting by heads of state[1]

0.053 0.077 1.1 $-\infty$ 1.5 22

Assuming tweet wait times for heads of state arise from an exponential distribution whose paramater is λ:

- Write the total likelihood and log-likelihood function.
- Advanced Find the maximum likelihood estimate for λ.
- Advanced Plot the likelihood profile for λ using R.

[1] Wait times between tweets collected with R package twitteR.

Main points

- Likelihood allows us to evaluate the relative strength of evidence for one parameter or model relative to another.
- The data are fixed and the parameters are variable in likelihood functions. These functions do not integrate or sum to one over the range of values of the parameter.
- The data are variable and the parameter are fixed in probability mass functions and probability density functions. These functions sum or integrate to one over the support of the random variable, y.

Likelihood ratio confidence intervals

Find the upper and lower bounds of an interval where all λ values within that interval are as consistent with the data as λ_{MLE}.

We compute the likelihood ratio statistic:

$$\mathcal{R} = 2\log\left(\frac{\mathcal{L}(\lambda_{MLE} \mid t_1)}{\mathcal{L}(\lambda_0 \mid t_1)}\right) \sim \chi^2_{k=1}$$

which is distributed χ^2 with 1 degree of freedom. Note that we fail to reject H_0 that $\lambda = \lambda_0$ if $\mathcal{R} < \chi^2_{k=1}(1-\alpha)$.

Likelihood ratio confidence intervals

We determine the $(1-\alpha = 0.95)$ likelihood ratio confidence interval by finding the upper and lower bounds for all values of λ_0 where we would fail to reject H_0.

$$2\log\left(\frac{\mathcal{L}(\lambda_{MLE} \mid t_1)}{\mathcal{L}(\lambda \mid t_1)}\right) < \chi^2_{k=1}(0.95)$$

$$L(\lambda_{MLE}) - \frac{3.84}{2} < L(\lambda \mid t_1)$$

$$L(\lambda \mid t_1) > L(\lambda_{MLE}) - 1.92$$

Likelihood ratio confidence intervals

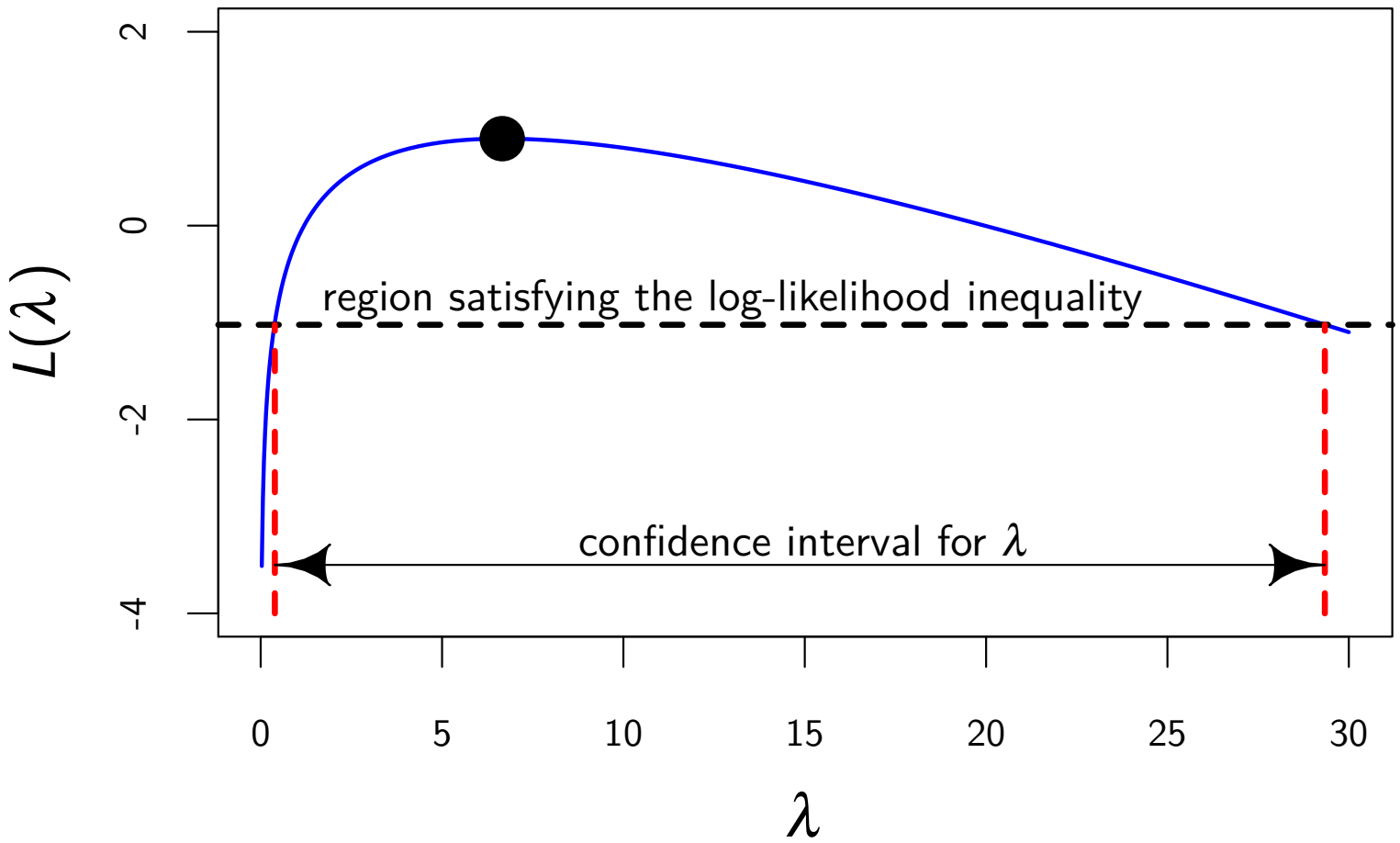
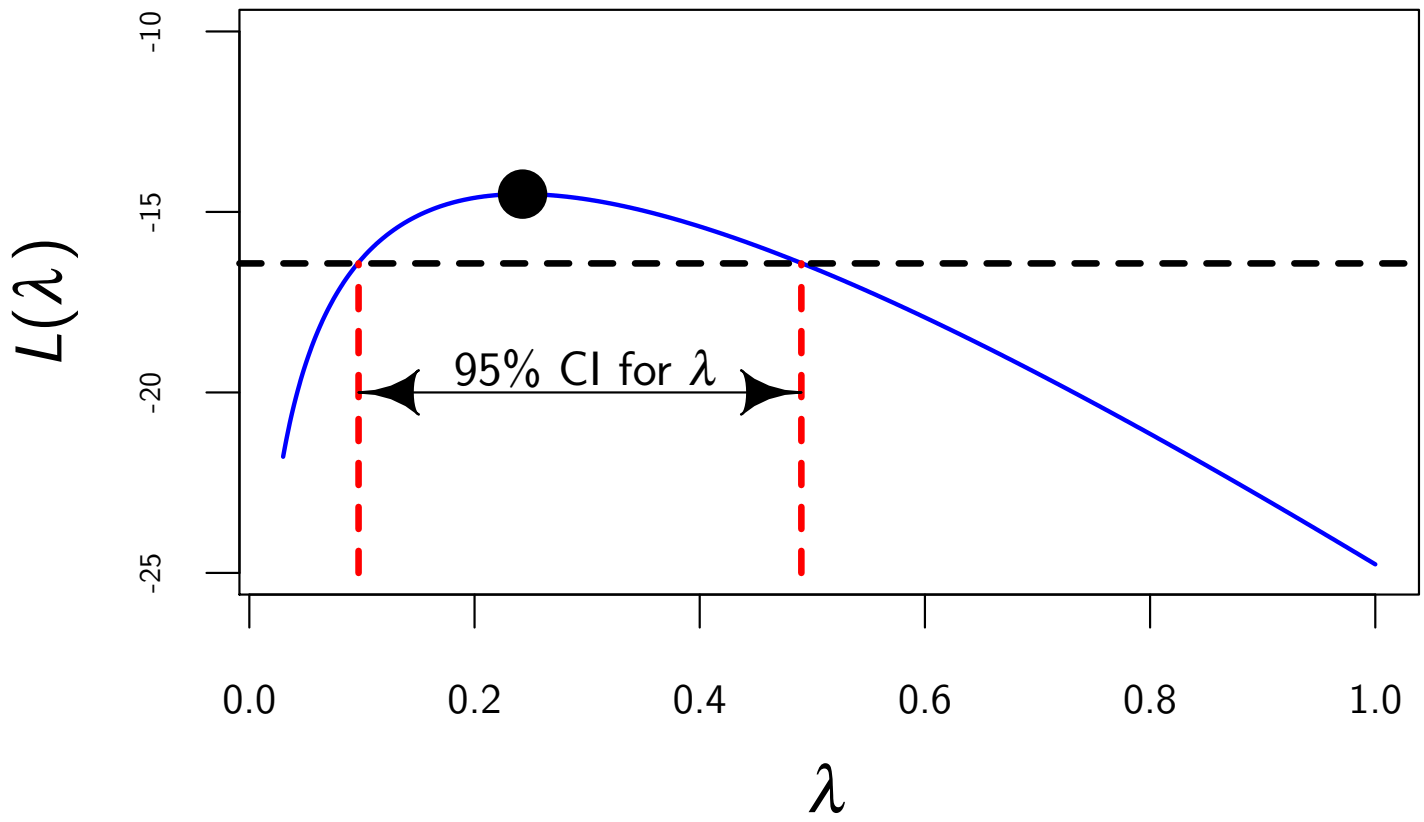


Figure courtesy of the UNC Biology Department.

Likelihood profile of λ for heads of state




```
lambda <- seq(.03, 1, length = 1000)
y <- NA
for(i in 1:length(lambda)) {y[i] <- log(prod(dexp(c(.15, 1.5, 22, .077, .0053, 1.1), lambda[i])))}
```