

Lecture 6

Ciprian Crainiceanu

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Johns Hopkins University

August 16, 2016

Table of contents

- 1 Table of contents
- 2 Outline
- 3 Defining likelihood
- 4 Interpreting likelihoods
- 5 Plotting likelihoods
- 6 Maximum likelihood
- 7 Interpreting likelihood ratios
- 8 Multiple parameters

Outline

- 1 Define likelihood
- 2 Interpretations of likelihoods
- 3 Likelihood plots
- 4 Maximum likelihood
- 5 Likelihood ratio benchmarks

- A common approach to statistics is to assume that data arise from a family of distributions indexed by a parameter that represents a useful summary of the distribution
- The **likelihood** of the data is the joint density evaluated as a function of the parameters with the data fixed
- Likelihood analysis of data uses the likelihood to perform inference regarding the unknown parameter

Examples: Normal

- $X_1, X_2, X_3 \sim N(\mu, 1)$ are independent identically distributed rvs (conditional on μ)
$$f(x, \mu) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(x-\mu)^2}{2} \right\}$$
- Suppose that we observe $X_1 = 5, X_2 = 2, X_3 = 3$
- $\mathcal{L}(\mu|X_1 = 5, X_2 = 2, X_3 = 3) = f(5, \mu)f(2, \mu)f(3, \mu)$
- $\mathcal{L}(\mu|\mathbf{x}) = \frac{1}{(2\pi)^{3/2}} \exp \left\{ -\frac{(5-\mu)^2 + (2-\mu)^2 + (3-\mu)^2}{2} \right\}$

Examples: Normal

- In general if $X_1 = x_1, \dots, X_n = x_n$

$$\mathcal{L}(\mu|\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} \exp \left\{ -\frac{\sum_i^n (x_i - \mu)^2}{2} \right\}$$

- Note that $f(\mu|\mathbf{x})$ is not a normalized pdf, that is,

$$\int \mathcal{L}(\mu|\mathbf{x}) d\mu \neq 1$$

- Taking logs typically makes log likelihoods better behaved

$$-2 \log \{ \mathcal{L}(\mu|\mathbf{x}) \} = \sum_i^n (x_i - \mu)^2 + \text{const.}$$

Examples: Normal

```
mu=4
```

```
bx=c(5,2,3)
```

```
ebx2=-sum((bx-mu)^2)/2
```

```
like=exp(ebx2)/((2*pi)^(length(bx)/2))
```

```
mu=seq(0,6,length=201)
```

```
likep=rep(0,201)
```

```
for (i in 1:201)
```

```
  {ebx2=-sum((bx-mu[i])^2)/2
```

```
  likep[i]=exp(ebx2)/((2*pi)^(length(bx)/2))}
```

```
plot(mu,likep,type="l",col="blue",lwd=3)
```

```
mle<-mu[which.max(likep)]
```

Lecture 6

Ciprian
Crainiceanu

Table of
contents

Outline

**Defining
likelihood**

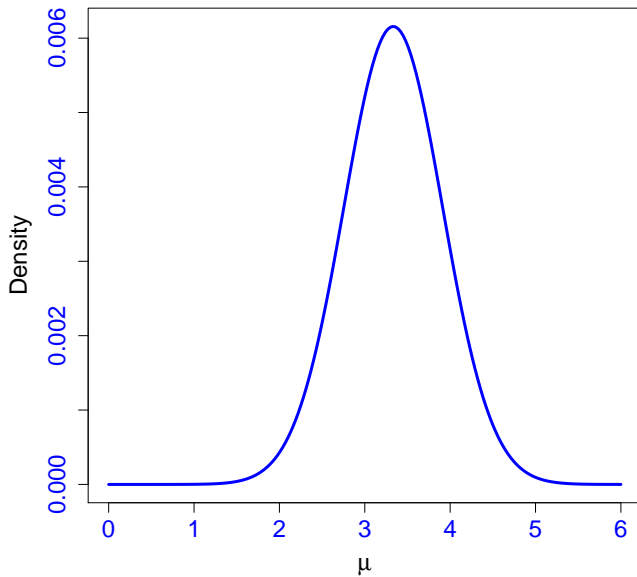
Interpreting
likelihoods

Plotting
likelihoods

Maximum
likelihood

Interpreting
likelihood
ratios

Multiple
parameters



Given a statistical probability mass function or density, say $f(x, \theta)$, where θ is an unknown parameter, the **likelihood** is f viewed as a function of θ for a fixed, observed value of x

$$\mathcal{L}(\theta|x) = f(x, \theta)$$

Interpretations of likelihoods

The law of likelihood requires:

- ① Ratios of likelihood values measure the relative **evidence** of one value of the unknown parameter to another
- ② **Likelihood principle:** Given a statistical model and observed data, all of the relevant information contained in the data regarding the unknown parameter is contained in the likelihood
- ③ If $\{X_i\}$ are independent random variables, then their likelihoods multiply. That is, the likelihood of the parameters given all of the X_i is simply the product of the individual likelihoods

Log likelihood

- Assume X_1, \dots, X_n are iid with pdf $f(x, \theta)$
- Likelihood

$$\mathcal{L}(\theta|\mathbf{x}) = \prod_{i=1}^n f(x_i, \theta)$$

- Log likelihood

$$\log\{\mathcal{L}(\theta|\mathbf{x})\} = \sum_{i=1}^n \log\{f(x_i, \theta)\}$$

Example

- Suppose that we flip a coin with success probability θ
- Recall that the mass function for x

$$f(x, \theta) = \theta^x (1 - \theta)^{1-x} \quad \text{for } \theta \in [0, 1].$$

where x is either 0 (Tails) or 1 (Heads)

- Suppose that the result is a head
- The likelihood is

$$\mathcal{L}(\theta|1) = \theta^1 (1 - \theta)^{1-1} = \theta \quad \text{for } \theta \in [0, 1].$$

- Therefore, $\mathcal{L}(.5|1)/\mathcal{L}(.25|1) = 2$,
- There is twice as much evidence supporting the hypothesis that $\theta = .5$ than the hypothesis that $\theta = .25$

Example continued

- Suppose now that we flip our coin from the previous example 4 times and get the sequence 1, 0, 1, 1
- The likelihood is:

$$\begin{aligned}
 \mathcal{L}(\theta|1, 0, 1, 1) &= \theta^1(1 - \theta)^{1-1}\theta^0(1 - \theta)^{1-0} \\
 &\times \theta^1(1 - \theta)^{1-1}\theta^1(1 - \theta)^{1-1} \\
 &= \theta^3(1 - \theta)^1
 \end{aligned}$$

- This likelihood only depends on the total number of heads and the total number of tails; we might write $\mathcal{L}(\theta|1, 3)$ for shorthand
- Now consider $\mathcal{L}(.5|1, 3)/\mathcal{L}(.25|1, 3) = 5.33$
- There is over five times as much evidence supporting the hypothesis that $\theta = .5$ over the hypothesis that $\theta = .25$

Plotting likelihoods

- Generally, we want to consider all the values of θ between 0 and 1
- A **likelihood plot** displays θ by $\mathcal{L}(\theta|x)$
- Usually, it is divided by its maximum value so that its height is 1
- Because the likelihood measures *relative evidence*, dividing the curve by its maximum value (or any other value for that matter) does not change its interpretation

Lecture 6

Ciprian
Crainiceanu

Table of
contents

Outline

Defining
likelihood

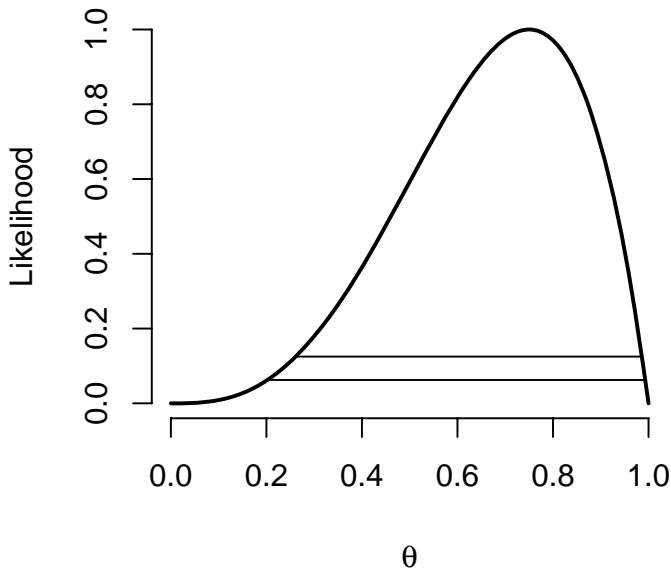
Interpreting
likelihoods

**Plotting
likelihoods**

Maximum
likelihood

Interpreting
likelihood
ratios

Multiple
parameters



Uniform distribution

- Suppose now that we observe three independent realizations from a uniform distribution $U[0, \theta]$
- $X_1 = 5, X_2 = 2, X_3 = 3$
- The likelihood of one observation

$$f(x, \theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

- The likelihood of all three observations

$$\begin{aligned} \mathcal{L}(\theta|\mathbf{x}) &= \frac{1}{\theta^3} I[0 \leq 5 \leq \theta] I[0 \leq 2 \leq \theta] I[0 \leq 3 \leq \theta] \\ &= \frac{1}{\theta^3} I[\theta \geq 5] \end{aligned}$$

Uniform distribution: R

```
theta=seq(1,10,by=0.1)
like=1/theta^3*(theta>=5)
plot(theta,like,type="l",col="blue",lwd=3)

like[theta==6]/like[theta==5]
like[theta==6]/like[theta==4]
theta[which.max(like)] # maximum likelihood

liken=like/max(like)
plot(theta,like,type="l",col="blue",lwd=3)
```

Lecture 6

Ciprian
Crainiceanu

Table of
contents

Outline

Defining
likelihood

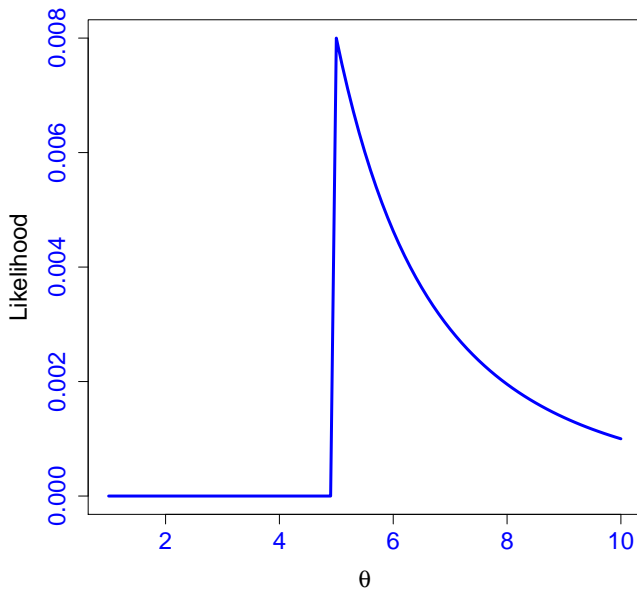
Interpreting
likelihoods

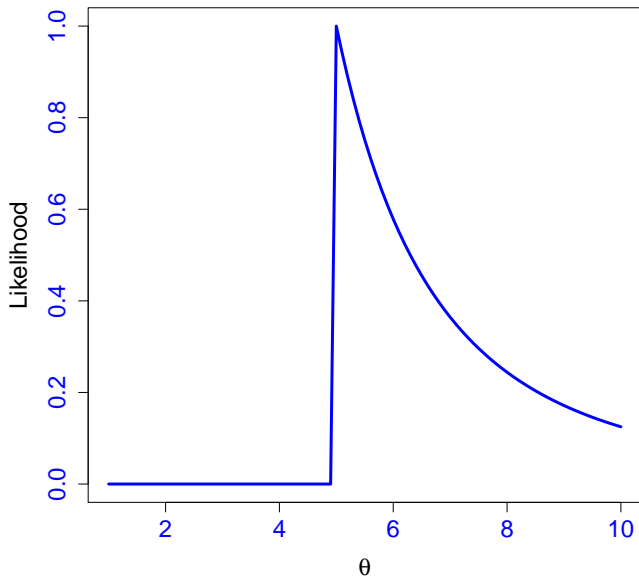
**Plotting
likelihoods**

Maximum
likelihood

Interpreting
likelihood
ratios

Multiple
parameters





Uniform distribution

- Suppose now we observe n independent realizations from a uniform distribution $U[0, \theta]$
- $X_1 = x_1, \dots, X_n = x_n$
- The likelihood for all n observations

$$\begin{aligned}\mathcal{L}(\theta|\mathbf{x}) &= \frac{1}{\theta^n} \prod_{i=1}^n I[0 \leq x_i \leq \theta] \\ &= \frac{1}{\theta^n} I[\theta \geq \max_i x_i]\end{aligned}$$

- Note that often the likelihood depends only on a function of the data (e.g. $\max_i x_i$)
- The evidence is often compressed in a much simpler, easier to understand form

Maximum likelihood

- The value of θ where the curve reaches its maximum has a special meaning
- It is the value of θ that is most well supported by the data
- This point is called the **maximum likelihood estimate** (or MLE) of θ

$$\hat{\theta}_{\text{ML}} = \text{MLE} = \operatorname{argmax}_{\theta} \mathcal{L}(\theta | \mathbf{x})$$

- Another interpretation of the MLE is that it is the value of θ that would make the data that we observed most probable
- Every estimator is a function of the observed data, \mathbf{x}

Maximum likelihood, coin example

Table of
contents

Outline

Defining
likelihoodInterpreting
likelihoodsPlotting
likelihoodsMaximum
likelihoodInterpreting
likelihood
ratiosMultiple
parameters

- The maximum likelihood estimate for θ is always the proportion of heads
- Proof: Let x be the number of heads and n be the number of trials
- Recall

$$\mathcal{L}(\theta|x) = \theta^x(1 - \theta)^{n-x}$$

- It's easier to maximize the **log-likelihood**

$$l(\theta, x) = x \log(\theta) + (n - x) \log(1 - \theta)$$

Continued

- Taking the derivative we get

$$\frac{d}{d\theta} l(\theta, x) = \frac{x}{\theta} - \frac{n-x}{1-\theta}$$

- Setting equal to zero implies

$$\left(1 - \frac{x}{n}\right)\theta = (1 - \theta)\frac{x}{n}$$

- Which is clearly solved at $\theta = \frac{x}{n}$
- Notice that the second derivative

$$\frac{d^2}{d\theta^2} l(\theta, x) = -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2} < 0$$

provided that x is not 0 or n (do these cases on your own)

What constitutes strong evidence?

- Again imagine an experiment where a person repeatedly flips a coin
- Consider the possibility that we are entertaining three hypotheses: $H_1 : \theta = 0$, $H_2 : \theta = .5$, and $H_3 : \theta = 1$

Lecture 6

Ciprian
Crainiceanu

Table of
contents

Outline

Defining
likelihood

Interpreting
likelihoods

Plotting
likelihoods

Maximum
likelihood

Interpreting
likelihood
ratios

Multiple
parameters

Outcome X	$P(X H_1)$	$P(X H_2)$	$P(X H_3)$	$\mathcal{L}(H_1)/\mathcal{L}(H_2)$	$\mathcal{L}(H_3)/\mathcal{L}(H_2)$
H	0	.5	1	0	2
T	1	.5	0	2	0
HH	0	.25	1	0	4
HT	0	.25	0	0	0
TH	0	.25	0	0	0
TT	1	.25	0	4	0
HHH	0	.125	1	8	8
HHT	0	.125	0	0	0
HTH	0	.125	0	0	0
THH	0	.125	0	0	0
HTT	0	.125	0	0	0
THT	0	.125	0	0	0
TTH	0	.125	0	0	0
TTT	1	.125	0	0	8

Benchmarks

- Using this example as a guide, researchers tend to think of a likelihood ratio
 - of 8 as being moderate evidence
 - of 16 as being moderately strong evidence
 - of 32 as being strong evidence
- of one hypothesis over another
- Because of this, it is common to draw reference lines at these values on likelihood plots
- Parameter values above the $1/8$ reference line, for example, are such that no other point is more than 8 times better supported given the data

Likelihood for multiple parameters

- So far, we have focused on the case when θ is a scalar
- Many distributions depend on multiple parameters (normal, gamma, beta, t)
- Definitions remain the same
- Likelihood

$$\mathcal{L}(\theta|\mathbf{x}) = \prod_{i=1}^n f(x_i, \theta)$$

- MLE

$$\hat{\theta}_{\text{ML}} = \text{MLE} = \operatorname{argmax}_{\theta} \mathcal{L}(\theta|\mathbf{x})$$

- It simply requires working with multivariate parameters



Profile likelihood

- Sometimes one is interested in one of the parameters, whereas the others are not the primary focus of the problem
- Example: Effect of air pollution as measured by $PM_{2.5}$ on cardiovascular outcomes in the presence of potential confounders (temperature, secular trends, etc)
- Evidence is hard to visualize with respect to all parameters at once
- Profile likelihood: a way to visualize the evidence with respect to the parameter of interest

Profile likelihood

Table of
contents

Outline

Defining
likelihoodInterpreting
likelihoodsPlotting
likelihoodsMaximum
likelihoodInterpreting
likelihood
ratiosMultiple
parameters

- X_1, \dots, X_n iid with pdf $f(x, \theta)$
- The multivariate parameter can be partitioned in $\theta = (\mu, \eta)$
 - μ is a scalar parameter of interest
 - η are the nuisance parameters
- For each value of μ maximize the likelihood with respect to the rest of the parameters η
- Obtain $\hat{\eta}(\mu, \mathbf{x}) = \max_{\eta} \mathcal{L}(\mu, \eta | \mathbf{x})$
- The profile likelihood is

$$\mathcal{PL}(\mu | \mathbf{x}) = \mathcal{L}\{\mu, \hat{\eta}(\mu, \mathbf{x}) | \mathbf{x}\}$$

Profile likelihood: normal

- In general if $X_1 = x_1, \dots, X_n = x_n$ with mean μ and variance σ^2

$$\mathcal{L}(\mu, \sigma^2 | \mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ -\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2} \right\}$$

- Fix μ and maximize the log likelihood with respect to σ^2
- Log likelihood $\left(-\sum_{i=1}^n 2(x_i - \mu) (-1) = 0 \quad \hat{\mu} = \frac{\sum x_i}{n} \right)$

maximize

$$2 \log \{ \mathcal{L}(\mu, \sigma^2 | \mathbf{x}) \} = - \sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^2} - n \log(\sigma^2) + \text{const.}$$

- Profile estimator of the variance

$$\frac{d}{d\sigma^2} = \frac{\sum_{i=1}^n (x_i - \mu)^2}{(\sigma^2)^2} - \frac{n}{\sigma^2}$$

$$\hat{\sigma}^2(\mu, \mathbf{x}) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \quad \leftarrow = 0$$

Profile likelihood: normal

- The profile log likelihood is

$$2 \log \mathcal{PL}(\mu|\mathbf{x}) = 2 \log \mathcal{L}\{\mu, \hat{\sigma}^2(\mu, \mathbf{x})|\mathbf{x}\}$$

- Plug-in estimator

$$2 \log \mathcal{PL}(\mu|\mathbf{x}) = -n \log \left\{ \sum_{i=1}^n (x_i - \mu)^2 \right\} + \text{const}$$

- Thus, the profile likelihood is, essentially, the sum of squares for the normal distribution

Profile likelihood: R

Table of
contents

Outline

Defining
likelihoodInterpreting
likelihoodsPlotting
likelihoodsMaximum
likelihoodInterpreting
likelihood
ratiosMultiple
parameters

```
bx=c(5,2,3) # Data
mu=seq(0,6,length=201) # Grid for  $\mu$ 
likep=rep(0,201)
for (i in 1:201)
  {likep[i]=-3*log(sum((bx-mu[i])^2))}

plot(mu,likep,type="l",col="blue",lwd=3)
mlep<-mu[which.max(likep)]
```


Lecture 6

Ciprian
Crainiceanu

Table of
contents

Outline

Defining
likelihood

Interpreting
likelihoods

Plotting
likelihoods

Maximum
likelihood

Interpreting
likelihood
ratios

Multiple
parameters

