Topic 3: Estimation

Ethan P. Marzban   University of California, Santa Barbara   PSTAT 120B

Outline

1. Likelihoods

2. Maximum Likelihood Estimation

Likelihoods

UC SANTA BARBARA
Department of Statistics
and Applied Probability

Leadup

- Last lecture, we began discussing the notion of a **likelihood**.
- Recall that, computationally, a likelihood is just a joint PMF/PDF that we now treat as a function of one or more population parameters.
- Conceptually, the likelihood evaluated at a given set of observations represents how *likely* a given value of the parameter is.

Likelihood

Definition (Likelihood)

Let $\vec{\mathbf{y}} := \{y_i\}_{i=1}^n$ be an observed instance of a random sample $\vec{\mathbf{Y}} := \{Y_i\}_{i=1}^n$, whose distribution depends on some parameter θ. The **likelihood** of the sample is simply the joint PMF/PDF of $\vec{\mathbf{Y}}$.

- To avoid having to constantly separate the discrete and continuous cases, we adopt the notation

$$\mathcal{L}_{\vec{\mathbf{y}}}(\theta) \qquad \text{or} \qquad \mathcal{L}(y_1, \cdots, y_n; \theta)$$

to mean the likelihood.

Notation

- A quick note on notation: I will use the notations $\mathcal{L}_{\vec{\boldsymbol{y}}}(\theta)$ and $\mathcal{L}(y_1, \cdots, y_n; \theta)$ interchangeably [though the second notation makes the sample values clearer, it is clunkier than the first].
- Just be aware that the textbook always uses $\mathcal{L}(y_1, \cdots, y_n; \theta)$.
  - Technically the textbook writes $\mathcal{L}(y_1, \cdots, y_n \mid \theta)$, but so as to avoid confusion with conditional distributions I will avoid using this notation for the purposes of this class.
- And, again, to reiterate - the likelihood is nothing more than the joint PMF/PDF of a random sample, evaluated at a particular observed instance $\vec{\boldsymbol{y}}$.

Simplification

- Now, if we assume an i.i.d. sample, we can expand things a bit.
- For instance, if $Y_1, \cdots, Y_n$ are i.i.d. discrete random variables from a distribution with mass function $p(y; \theta)$, then

$$\begin{aligned} \mathcal{L}_{\vec{\mathbf{y}}}(\theta) &= p_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots, x_n) \\ &= p_{X_1}(x_1; \theta) \times p_{X_2}(x_2; \theta) \times \cdots \times p_{X_n}(x_n; \theta) = \prod_{i=1}^{n} p(x_i; \theta) \end{aligned}$$

UC SANTA BARBARA
Department of Statistics and Applied Probability

Simplification

- Similarly, if $Y_1, \cdots, Y_n$ are i.i.d. continuous random variables from a distribution with density function $f(y; \theta)$, then

$$\begin{aligned}\mathcal{L}_{\vec{\mathbf{y}}}(\theta) &= f_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots, x_n) \\ &= f_{X_1}(x_1; \theta) \times f_{X_2}(x_2; \theta) \times \cdots \times f_{X_n}(x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)\end{aligned}$$

Example

Example

The weight of a randomly-selected DSH cat is assumed to be normally distributed about some unknown mean μ and with some known standard deviation $\sigma = 2$ lbs. An i.i.d. random sample of 3 cats is taken; their weights are 8.2 lbs, 16.2 lbs, and 14.1 lbs. What is the likelihood of this sample? (Remember that this will be a function of μ!)

UC SANTA BARBARA
Department of Statistics and Applied Probability

Solution

- Let Y_i denote the weight of a randomly-selected DSH cat; then $Y_1, Y_2, \cdots \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 4)$.
- Hence, the density of Y_i at a point y_i is given by the density of a $\mathcal{N}(\mu, 4)$ distribution, evaluated at y_i:

$$f(y_i; \mu) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{8}(y_i - \mu)^2}$$

UC SANTA BARBARA
Department of Statistics
and Applied Probability

Solution

- Therefore,

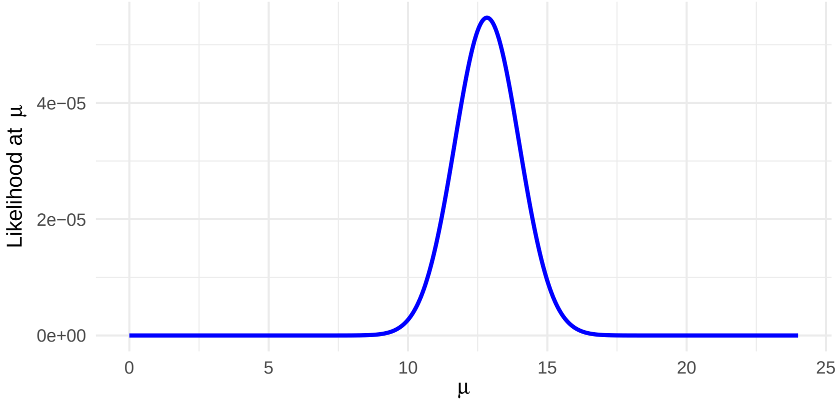
$$\mathcal{L}_{(8.2,16.2,14.1)}(\mu) = f(8.2;\mu) \times f(16.2;\mu) \times f(14.1;\mu)$$

$$= \left(\frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{8}(8.2-\mu)^2} \right) \times \left(\frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{8}(16.2-\mu)^2} \right) \times \left(\frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{8}(14.1-\mu)^2} \right)$$

$$= \left(\frac{1}{2\pi} \right)^3 \exp\left\{ -\frac{1}{8} [(8.2-\mu)^2 + (16.2-\mu)^2 + (14.1-\mu)^2] \right\}$$

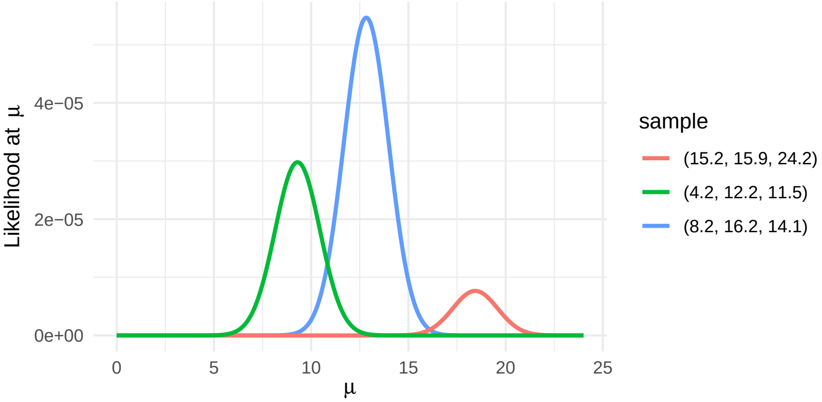
UC SANTA BARBARA
Department of Statistics and Applied Probability

Example



UC SANTA BARBARA
Department of Statistics
and Applied Probability

Example



UC SANTA BARBARA
Department of Statistics and Applied Probability

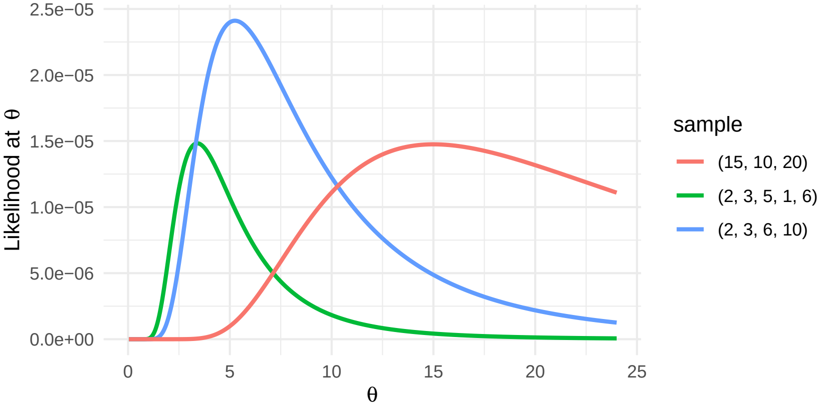
Example

Example

The wait time of a randomly-selected person at the DMV follows an exponential distribution with unknown parameter θ. Assuming an i.i.d. sample $\{Y_i\}_{i=1}^n$ of wait times and their corresponding observed instances $\{y_i\}_{i=1}^n$, what is the likelihood as a function of θ and $\{y_i\}_{i=1}^n$?

- Let's do this one on the board.

Example



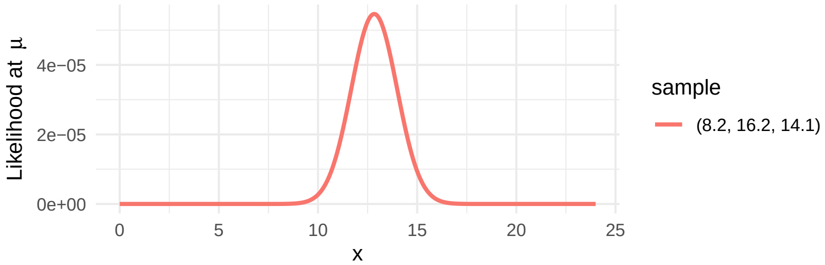
2.5e−05
2.0e−05
1.5e−05
1.0e−05
5.0e−06
0.0e+00
Likelihood at θ
0
5
10
15
20
25
θ
sample
(15, 10, 20)
(2, 3, 5, 1, 6)
(2, 3, 6, 10)

Leadup

- Alright, so that's what a likelihood is. Why do we care?
- Again - I think of the likelihood as, well, the *likelihood* of a particular value of θ, given the data we observed.
    - Given that three randomly-selected cats weigh 8.2, 16.2, and 14.1 lbs, how likely is it that the true average weight of all cats is 10 lbs? 10.2 lbs? 11.4 lbs?
- So, here's the clever idea of how to leverage this to construct an estimator for θ - why don't we choose θ to maximize the likelihood of a particular sample!
- This is the idea behind **maximum likelihood estimation**.

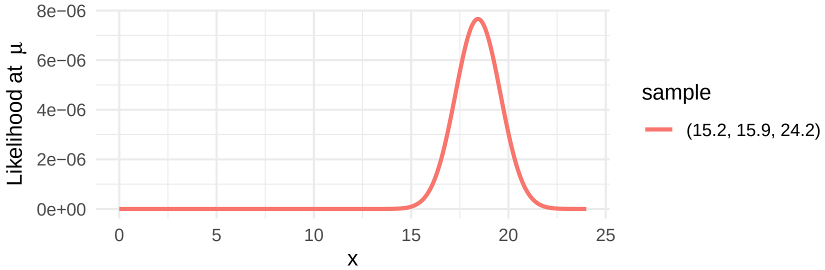
Maximum Likelihood Estimation

Intuition



- Given that we observed cat weights of 8.2, 16.2, and 14.1 lbs, the most plausible value for μ (i.e. the point corresponding to the highest likelihood) is around 13. Hence, a "good" estimate for μ, given the sample we observed, is around 13.

Intuition



- Given that we observed cat weights of 15.2, 15.9, and 24.2 lbs, the most plausible value for μ (i.e. the point corresponding to the highest likelihood) is around 18.5 Hence, a "good" estimate for μ, given the sample we observed, is around 18.5

UC SANTA BARBARA
Department of Statistics
and Applied Probability

Intuition

- The textbook has another (in my opinion) nice way of introducing the notion of maximum likelihood estimation.
- Say we have a bucket containing 3 marbles, some of which are blue and some of which are gold.
- Suppose we take a sample of 2 marbles, and observe that they are both gold. What is a "good" guess for the total number of gold marbles in the bucket?
- Let X denote the number of gold marbles in a sample of 2, taken at random and without replacement from a bucket containing 3 marbles, γ of which are gold. Then

$$X \sim \text{HyperGeom}(3, \gamma, 2)$$

Intuition

- If there are only 2 gold marbles in the bucket, then the probability of observing the 2 gold marbles we did in our sample is given by

$$\mathbb{P}(X = 2) = \frac{\binom{2}{2}\binom{1}{0}}{\binom{3}{2}} = \frac{1}{3}$$

- If there are 3 gold marbles in the bucket, then the probability of observing the 2 gold marbles we did in our sample is given by

$$\mathbb{P}(X = 2) = \frac{\binom{3}{2}}{\binom{3}{2}} = 1$$

- So, $\gamma = 3$ leads to a higher *likelihood* of having observed the 2 gold marbles we did than $\gamma = 2$.

UC SANTA BARBARA
Department of Statistics and Applied Probability

Maximum Likelihood Estimator

Definition (Maximum Likelihood Estimator)

Given a random sample $\vec{\mathbf{Y}} = \{Y_i\}_{i=1}^{n}$ from a population with unknown parameter θ, we define the **maximum likelihood estimator** for θ, denoted $\widehat{\theta}_{\text{MLE}}$, as

$$\widehat{\theta}_{\text{MLE}} = \arg\max_{\theta} \{\mathcal{L}_{\vec{\mathbf{Y}}}(\theta)\}$$

- Notice that when finding the MLE, we evaluate the likelihood at the *random* sample (so that we obtain a *random* estimator). More on that later.

UC SANTA BARBARA
Department of Statistics and Applied Probability

Leadup

- Now, recall that if our sample is i.i.d., then the likelihood becomes a product of several terms.
- Hence, maximizing the likelihood would require us to take the derivative of a function consisting of a product of a bunch of terms, which would therefore require several applications of the product rule (for derivatives).
- As such, the likelihood is somewhat rarely maximized directly. Instead, we make use of a clever fact: given a function $f(x)$ maximized at a point x' and a strictly increasing function $g(\cdot)$, then $(f \circ g)$ is also maximized at x'.

UC SANTA BARBARA
Department of Statistics and Applied Probability

Log-Likelihood

- Motivated by this, we define the following quantity:

Definition (Log-Likelihood)

Given an observation $\vec{\mathbf{y}}$ of a sample $\vec{\mathbf{Y}}$ and the corresponding likelihood $\mathcal{L}_{\vec{\mathbf{y}}}(\theta)$, we define the **log likelihood**, notated $\ell_{\vec{\mathbf{y}}}(\theta)$, to be the natural logarithm of the likelihood. That is,

$$\ell_{\vec{\mathbf{y}}}(\theta) = \ln \mathcal{L}_{\vec{\mathbf{y}}}(\theta)$$

UC SANTA BARBARA
Department of Statistics and Applied Probability

Log-Likelihood

- Since the logarithm is a strictly increasing function, the value that maximizes the log-likelihood will be the same value that maximizes the likelihood.
- That is, the MLE is equivalently given by the maximizing value of the log-likelihood.
- Furthermore, recall that logarithm of products are simply sums of logarithms!
- This is the guiding reason behind why we often maximize the *log*-likelihood, as opposed to the likelihood itself - maximizing the log-likelihood typically involves only taking the *sum* of several derivatives.

UC SANTA BARBARA
Department of Statistics and Applied Probability

Log-Likelihood

- More explicitly, suppose we have a continuous sample $\vec{\mathbf{Y}}$. Then

$$\mathcal{L}_{\vec{\mathbf{Y}}}(\theta) = \prod_{i=1}^{n} f(Y_i; \theta)$$

- Therefore,

$$\ell_{\vec{\mathbf{Y}}}(\theta) = \ln \left[\prod_{i=1}^{n} f(Y_i; \theta) \right] = \sum_{i=1}^{n} \ln f(Y_i; \theta)$$

which is much easier to differentiate than the original likelihood.

UC SANTA BARBARA
Department of Statistics
and Applied Probability

Example

Example

Given $Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(\theta)$, derive an expression for $\widehat{\theta}_{\text{MLE}}$, the maximum likelihood estimator for θ.

Solutions

- We've previously seen that

$$\mathcal{L}_{\vec{\mathbf{Y}}}(\theta) = \left(\frac{1}{\theta} \right)^n \cdot \exp\left\{ -\frac{1}{\theta} \sum_{i=1}^{n} Y_i \right\} \cdot \prod_{i=1}^{n} \mathbb{1}_{\{Y_i \geq 0\}}$$

- The log-likelihood is therefore given by

$$\ell_{\vec{\mathbf{Y}}}(\theta) = -n \ln(\theta) - \frac{1}{\theta} \sum_{i=1}^{n} Y_i + \sum_{i=1}^{n} \ln \mathbb{1}_{\{Y_i \geq 0\}}$$

UC SANTA BARBARA
Department of Statistics and Applied Probability

Solutions

- The derivative of the log-likelihood wrt. θ is:

$$\frac{\partial}{\partial \theta} \ell_{\vec{\mathbf{Y}}}(\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} Y_i$$

- Therefore, $\widehat{\theta}_{\mathrm{MLE}}$ satisfies

$$-\frac{n}{\widehat{\theta}_{\mathrm{MLE}}} + \frac{1}{\widehat{\theta}^2_{\mathrm{MLE}}} \sum_{i=1}^{n} Y_i = 0$$

- Solving and simplifying yields $\widehat{\theta}_{\mathrm{MLE}} = \overline{Y}_n$.

UC SANTA BARBARA
Department of Statistics
and Applied Probability

Multi-Parameter Case

- If the underlying population distribution has multiple parameters, we can still find maximum likelihood estimators for each by *jointly* maximizing the likelihood.
- In practice, this typically amounts to taking derivatives wrt. each of the parameters of interest, setting these derivatives equal to zero, and solving the resulting *system* of equations.

Example

Example

Given $Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ where both $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown parameters, find maximum likelihood estimators for both μ and σ^2.

- You'll work through this during Discussion Section.

Example

Example

Given $Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Unif}[0, \theta]$ where $\theta > 0$ is an unknown parameter, find $\widehat{\theta}_{\text{MLE}}$, the maximum likelihood estimator for θ.

Solution

- Let's begin as we did before, by first finding the likelihood:

$$\mathcal{L}_{\vec{\mathbf{Y}}}(\theta) = \prod_{i=1}^{n} f(Y_i; \theta) = \prod_{i=1}^{n} \left[\frac{1}{\theta} \cdot \mathbb{1}_{\{0 \leq Y_i \leq \theta\}} \right]$$

$$= \left(\frac{1}{\theta} \right)^n \cdot \prod_{i=1}^{n} \mathbb{1}_{\{0 \leq Y_i \leq \theta\}}$$

- First note: the likelihood is **<u>NOT</u>** just equal to $(1/\theta)^n$!!! The product of indicators is **<u>ABSOLUTELY</u>** a part of the likelihood. In fact, let's focus on that product a bit.

UC SANTA BARBARA
Department of Statistics and Applied Probability

Solution

- The entire product (of indicators) is nonzero only when all of the constituent indicators are nonzero. This only happens when all of the Y_i's are greater than 0 and less than θ, which occurs when $Y_{(1)} \geq 0$ and $Y_{(n)} \leq \theta$. Therefore:

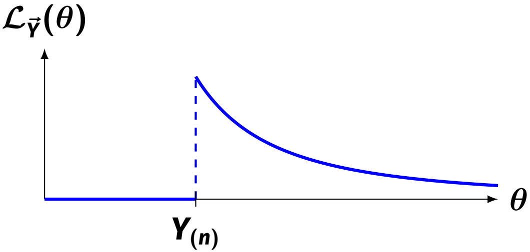
$$\prod_{i=1}^{n} \mathbb{1}_{\{0 \leq Y_i \leq \theta\}} = \mathbb{1}_{\{Y_{(1)} \geq 0\}} \cdot \mathbb{1}_{\{Y_{(n)} \leq \theta\}}$$

and our likelihood can be written as

$$\mathcal{L}_{\vec{\mathbf{Y}}}(\theta) = \left(\frac{1}{\theta} \right)^n \cdot \mathbb{1}_{\{Y_{(1)} \geq 0\}} \cdot \mathbb{1}_{\{Y_{(n)} \leq \theta\}}$$

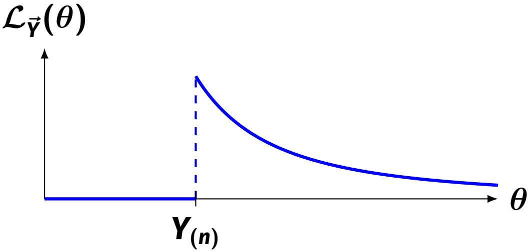
Solution

- Question - is this differentiable in θ?
- The answer is most definitively "no," because of the indicator.
- More specifically, here's a sketch of what the likelihood looks like:



Solution

- Of course, just because the likelihood is nondifferentiable doesn't mean that it doesn't have a maximizing value.
- Indeed, just looking at the graph of $\mathcal{L}_{\vec{\mathbf{Y}}}(\theta)$, we can see that it is maximized when θ equals $Y_{(n)}$:



Solution

- So, we find

$$\arg\max_{\theta} \{\mathcal{L}_{\vec{\mathbf{Y}}}(\theta)\} =: \widehat{\theta}_{\mathrm{MLE}} = Y_{(n)}$$

- How could we have arrived at this conclusion without sketching the likelihood?

UC SANTA BARBARA
Department of Statistics and Applied Probability

Solution

- Here's how I like to think about things. Take a look again at the parts of the likelihood that depend on θ;

$$\mathcal{L}_{\vec{\mathbf{Y}}}(\theta) \propto \left(\frac{1}{\theta} \right)^n \cdot \mathbb{1}_{\{\theta \geq Y_{(n)}\}}$$

- The term $(1/\theta)^n$ is a decreasing function in θ, meaning it is maximized by setting θ to be as small as possible. The term $\mathbb{1}_{\{\theta \geq Y_{(n)}\}}$ constrains θ to be no smaller than $Y_{(n)}$. Hence, combining these two facts, we see that the likelihood is maximized by setting θ to be $Y_{(n)}$, as we saw before.

UC SANTA BARBARA
Department of Statistics and Applied Probability