4-1 Maximum likelihood estimation

Food for thought: measuring asbestos concentration

A study done at the National Institute of Standards and Technology to develop measurement standards for asbestos concentration (Steel et al. 1980)

- Asbestos dissolved in water was spread on a filter.
- 3-mm diameter punches were taken from the filter.
- The number of fibers was counted in each of 23 grid squares.

31	29	19	18	31	28	34	27	34	30	16	18
26	27	27	18	24	22	28	24	21	17	24	

It is known that the Poisson model is a good model for count data of this type.

In Lecture 1-1, the method of moments gave $\widehat{\lambda} = \bar{X}$, which is 24.91 for these data.

By contrast, the method of maximum likelihood is based on the idea of picking the value that makes the observed data "most likely."

Unit 4 roadmap

Lecture 4-1 Maximum likelihood estimation

Lecture 4-2 Functional invariance and consistency

Lecture 4-3 Asymptotic normality

Today's plan

1. Definitions

2. Examples
    - Measuring asbestos concentration (Poisson)
    - Normal distribution
    - Uniform distribution

Learning objectives

By the end of this lecture, you should be able to:

1. Explain the idea behind maximum likelihood estimation.
2. Form the likelihood and the log-likelihood for a given set-up, and use them to find the mle.
3. Recognize when the mle can be found by setting derivatives to zero, and when it must be found from first principles.

Definitions

Recall: statistical inference

Suppose

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(x; \theta), \quad \theta \in \Theta$$

for some pdf/pmf $f(x; \theta)$.

We know what f is except for θ, which can take any value in the parameter space Θ. We obtain a random sample $\mathbf{X}_n$. We want to estimate θ from the observed values $\mathbf{x}_n$.

- random sample $\mathbf{X}_n = (X_1, \ldots, X_n)$
- observed sample $\mathbf{x}_n = (x_1, \ldots, x_n)$

The joint distribution of $\mathbf{X}_n$ is characterized by the joint pdf/pmf, which in the case of an i.i.d. sample is

Idea
Pick $\theta \in \Theta$ that is "most likely" given the observed values!

Likelihood and log-likelihood

The **likelihood** function is the joint pdf/pmf of a random sample regarded as a function of $\theta \in \Theta$:

$$L_n(\theta) = L(\theta; \mathbf{X}_n) = \prod_{i=1}^{n} f(X_i; \theta), \quad \theta \in \Theta$$

The **log-likelihood** function is the log of the likelihood:

$$\ell_n(\theta) = \log L(\theta; \mathbf{X}_n) = \sum_{i=1}^{n} \log f(X_i; \theta), \quad \theta \in \Theta$$

Example 4.1.1 Food for thought revisited (1/3) (Rice, Section 8.5, Example A)

In the Food for thought, the asbestos counts were modeled as $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$ with $\Theta = \{\lambda : \lambda > 0\}$ and the pmf

$$f(x; \lambda) =$$

The likelihood is

$$L_n(\lambda) =$$

The log-likelihood is

$$\ell_n(\lambda) =$$

$$=$$

$$=$$

Maximum likelihood estimator (mle)

Definition 4.1.1 (Definition 6.1.1 in HMC)
The **maximum likelihood estimator (mle)** of θ is

$$\widehat{\theta}_n = \arg\max_{\theta \in \Theta} L_n(\theta) = \arg\max_{\theta \in \Theta} \ell_n(\theta)$$

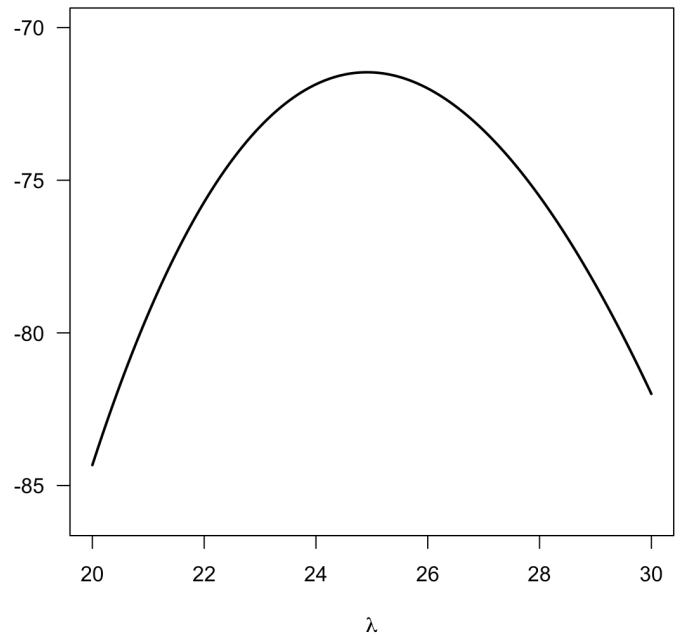
i.e., the value of θ that maximizes the likelihood—or equivalently, the log-likelihood—over the parameter space Θ.

Observation
L_n and ℓ_n are maximized at the same θ. However, ℓ_n will often be easier to work with than L_n. ℓ_n is also an i.i.d. sum; this has implications for the asymptotic normality.

Example 4.1.1 Food for thought revisited (2/3) (Rice, Section 8.5, Example A)

$$\ell_n(\lambda) = \log \lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n} \log(X_i!)$$

For the asbestos data, ℓ_n looks like the graph on the right. In general,



$$\ell_n'(\lambda) =$$

$$\ell_n''(\lambda) =$$

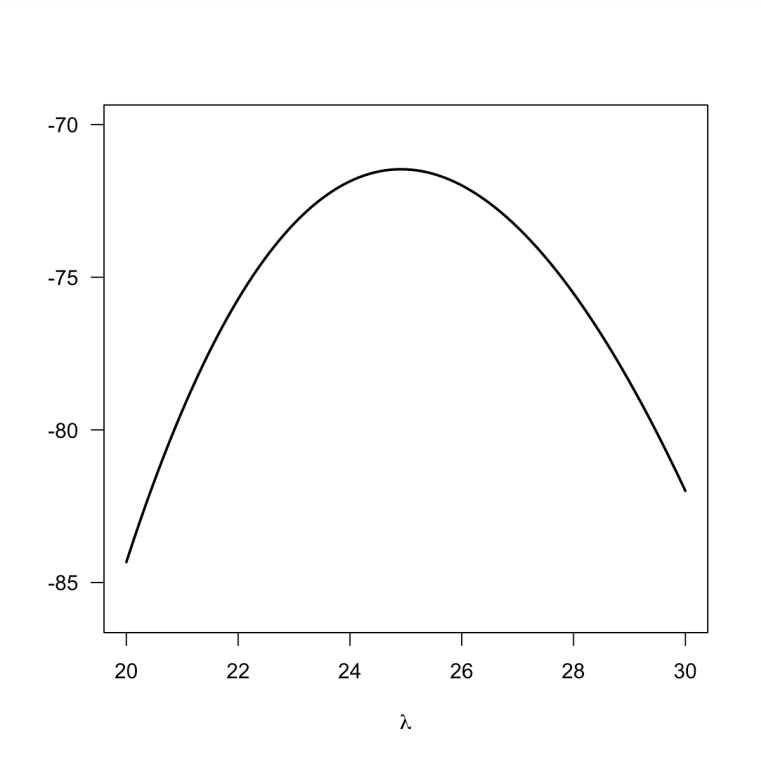
So, ℓ_n is smooth and *concave*[1] on Θ. In particular, a local maximum point is the global maximum point.

[1] A function with a graph that lies below all of its tangents.

Example 4.1.1 Food for thought revisited (3/3) (Rice, Section 8.5, Example A)

Thus, we can find the arg max of ℓ_n by solving

$$\ell_n'(\lambda) = \frac{1}{\lambda}\sum_{i=1}^{n} X_i - n = 0$$



The solution is

$$\widehat{\lambda}_n =$$

Thus, for the Poisson model, the mle of λ is , which is both

- the intuitive estimator, because
- the method of moments estimator from Lecture 1-1

Benefits

$$\widehat{\theta}_n = \arg\max_{\theta \in \Theta} L_n(\theta) = \arg\max_{\theta \in \Theta} \ell_n(\theta)$$

The framework of maximum likelihood estimation turns a statistical estimation problem into an *optimization* problem.

The benefit of this change in perspective is that we know a lot about optimization, and hence, we can simply borrow its tools to estimate parameters in many situations.

Examples

Examples

Measuring asbestos concentration (Poisson)

Example 4.1.2 Measuring asbestos concentration (Rice, Section 8.4, Example A)

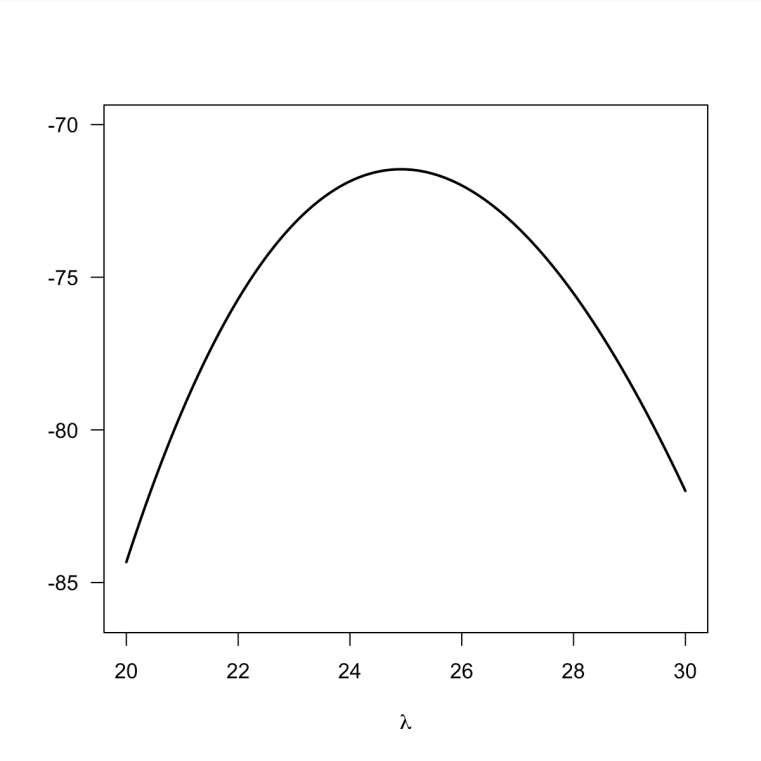
Recall the asbestos data from the Food for thought.

Model: $X_1, \ldots, X_{23} \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$ with $\Theta = \{\lambda : \lambda > 0\}$. By Example 4.1.1,

$$\ell_n(\lambda) = \log \lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n} \log(X_i!), \quad \widehat{\lambda}_n = \bar{X}_n$$

Here, $n =$ and $\sum_{i=1}^{n} x_i =$, so

$$\widehat{\lambda}_n =$$



WARNING! Heed the parameter space (1/3)

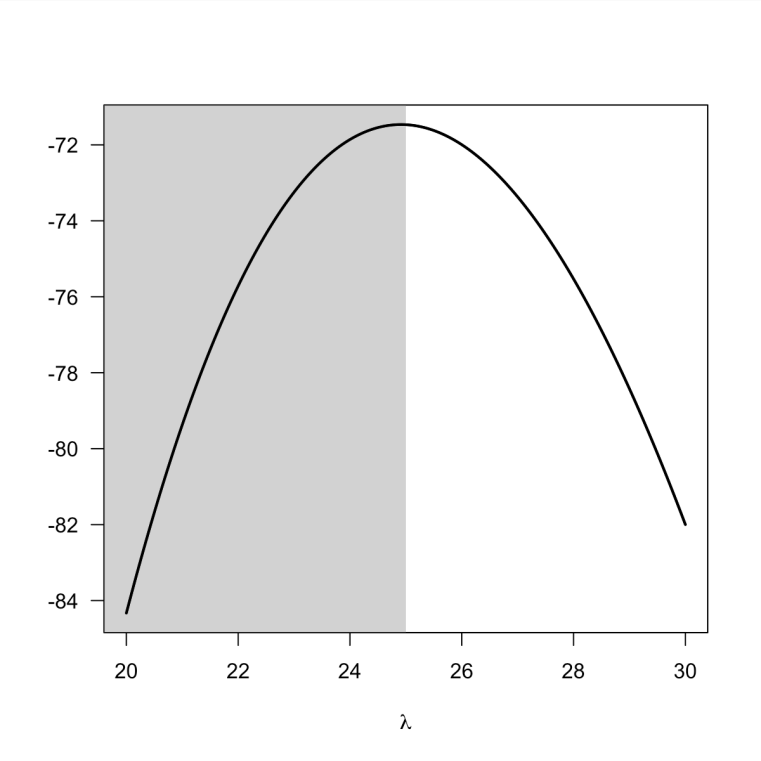
In Example 4.1.2, we found the mle $\widehat{\lambda}_n = 24.91$, which maximizes ℓ_n over

$$\Theta = \{\lambda : \lambda > 0\}$$

Now, suppose it is understood, for scientific reasons, that λ must be at least 25, i.e.,

$$\Theta' = \{\lambda : \lambda \geq 25\}$$

Then, $\widehat{\lambda}_n = 24.91$ is not a good estimate, because it is not a possible value of λ.



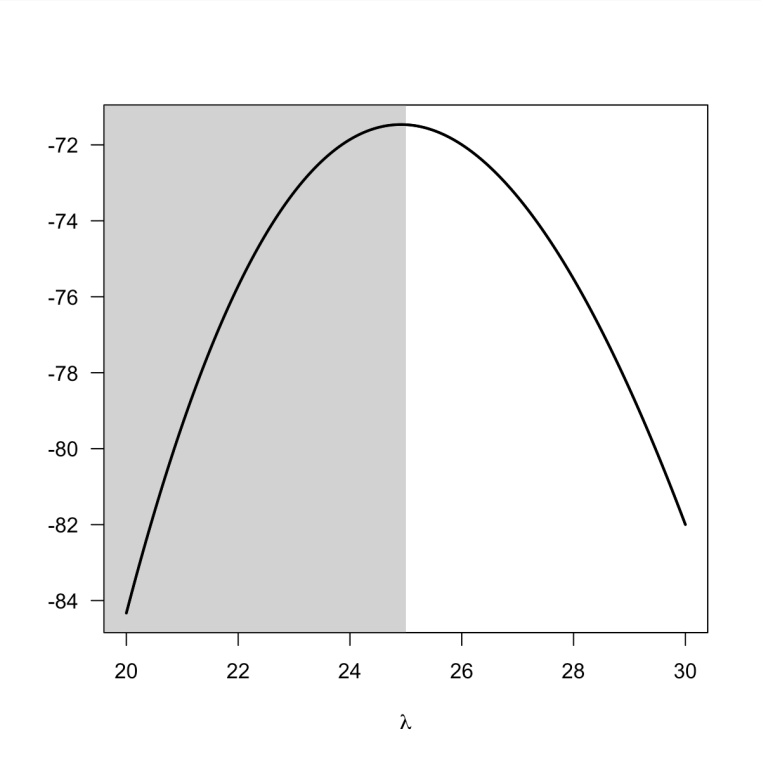
WARNING! Heed the parameter space (2/3)

It makes intuitive sense to do the following:

$$\widehat{\lambda}'_n =$$

For the asbestos data, $\widehat{\lambda}_n = 24.91 < 25$, so the mle $\widehat{\lambda}'_n$ over Θ' is

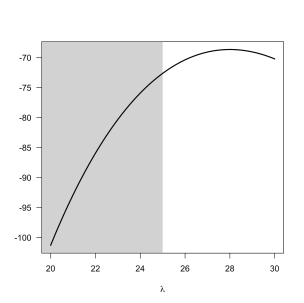
However, had the observed sample mean been greater than or equal to 25, we would have gone with the sample mean.



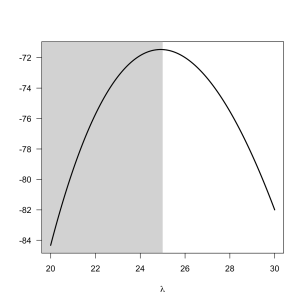
WARNING! Heed the parameter space (3/3)

$$\widehat{\lambda}_n' = \arg\max_{\lambda \geq 25} \ell_n(\lambda), \quad \ell_n(\lambda) = \log \lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n} \log(X_i!)$$

Case 1. $\widehat{\lambda}_n = \bar{X}_n \geq 25$



Case 2. $\widehat{\lambda}_n = \bar{X}_n < 25$



Examples

Normal distribution

Example 4.1.3 Normal distribution (1/2) (Rice, Section 8.5, Example B)

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \textbf{Normal}(\mu, \sigma^2)$, where $\theta = (\mu, \sigma^2)$ is unknown and the parameter space is

$$\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R},\ \sigma^2 > 0\}$$

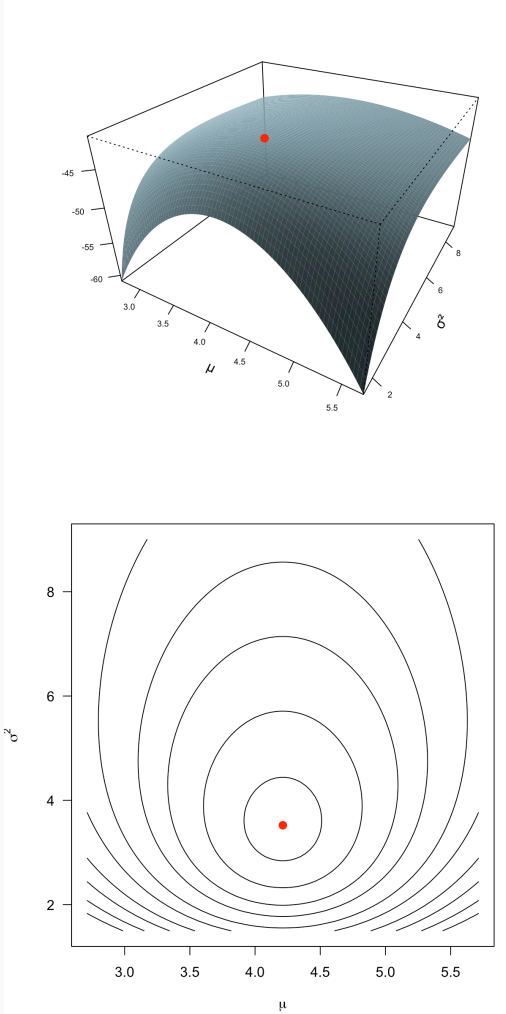
The likelihood is

$$L_n(\mu, \sigma^2) =$$

$$=$$

The log-likelihood is

$$\ell_n(\mu, \sigma^2) =$$



Example 4.1.3 Normal distribution (2/2) (Rice, Section 8.5, Example B)

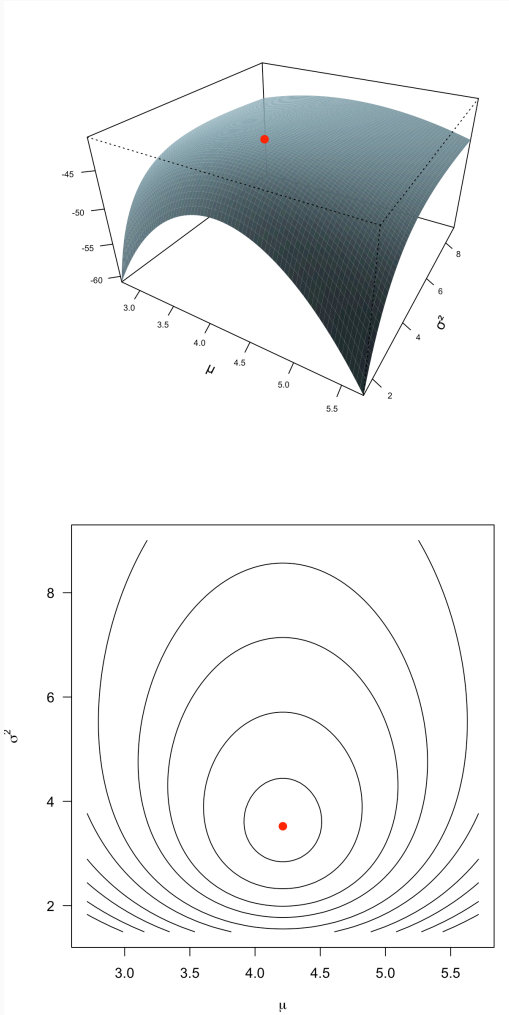
ℓ_n is smooth with a single peak, so we set the partial derivatives to zero:

$$\frac{\partial \ell_n}{\partial \mu} = \qquad\qquad = 0$$

$$\frac{\partial \ell_n}{\partial \sigma^2} = \qquad\qquad = 0$$

Solving the two equations in turn gives

$$\widehat{\mu}_n = \quad , \quad \widehat{\sigma}_n^2 =$$



Examples

Uniform distribution

Indicator function

Suppose $X \sim \textsf{Uniform}(0, \theta)$, $\theta > 0$. Then, X has the pdf

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

The **indicator function** of $A \subseteq \mathbb{R}$ is the function defined as

$$\mathbb{1}\{x \in A\} = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

One useful property of indicator functions is that for any $A, B \subseteq \mathbb{R}$,

$$\mathbb{1}\{x \in A \cap B\} = \mathbb{1}\{x \in A\}\mathbb{1}\{x \in B\}$$

Thus, the pdf of X can equivalently be expressed as

$$f(x; \theta) =$$

Example 4.1.4 Uniform$(0, \theta)$ (1/3) (HMC, Example 4.1.4)

Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \textsf{Uniform}(0, \theta)$, $\theta \in \Theta = \{\theta : \theta > 0\}$. Then, X_i has the pdf

$$f(x; \theta) = \frac{1}{\theta} \mathbb{1}\{x \leq \theta\} \mathbb{1}\{x \geq 0\}$$

The likelihood is

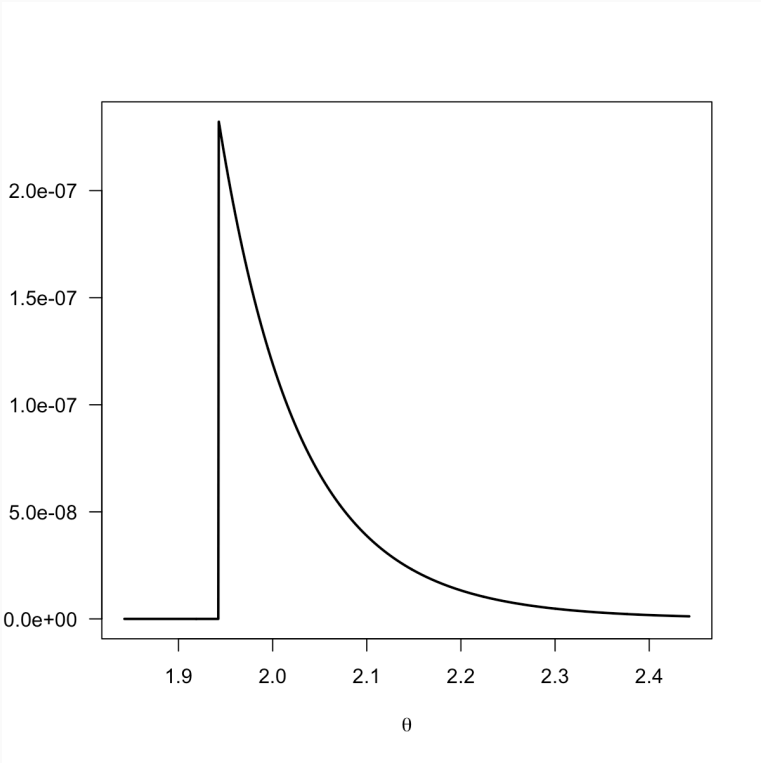
$$\begin{aligned} L_n(\theta) &= \\ &= \\ &= \\ &= \end{aligned}$$

Example 4.1.4 Uniform$(0, \theta)$ (2/3) (HMC, Example 4.1.4)

On the right, I generated a random sample from a uniform distribution and plotted L_n.

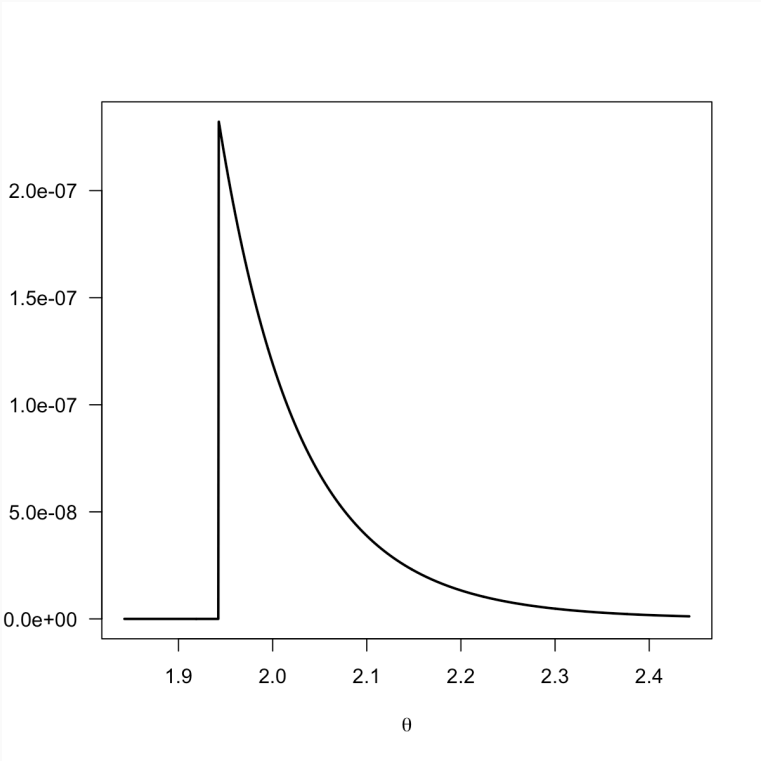
Note that $L_n(\theta)$ has a discontinuity.

$$L_n(\theta) = \frac{1}{\theta^n} \mathbb{1}\{X_{(n)} \leq \theta\} \mathbb{1}\{X_{(1)} \geq 0\}$$



Example 4.1.4 Uniform$(0, \theta)$ (3/3) (HMC, Example 4.1.4)

$$L_n(\theta) = \frac{1}{\theta^n} \mathbb{1}\{X_{(n)} \leq \theta\} \mathbb{1}\{X_{(1)} \geq 0\}$$



2

[2] We first saw this example in Lecture 1-1 (Example 1.1.4).

Wrapping up

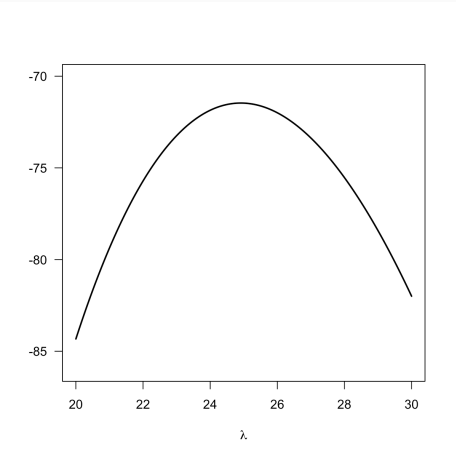
Key takeaways

- *mle*

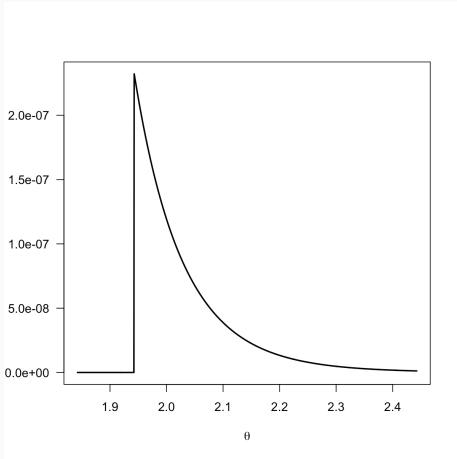
$$\widehat{\theta}_n = \arg\max_{\theta \in \Theta} L_n(\theta) = \arg\max_{\theta \in \Theta} \ell_n(\theta)$$

- *Finding the mle* Solve $\ell_n'(\theta) = 0$ when ℓ_n is smooth and concave on Θ. Otherwise, argue from first principles.

Regular case



Irregular case



Suggested exercises
Exercises 6.1.1 and 6.1.2(b) in HMC.

Next time: Lecture 4-2 Functional invariance and consistency

We will show that the mle of $g(\theta)$ is $g(\widehat{\theta}_n)$, and that, under regularity conditions, the mle is consistent:

$$\widehat{\theta}_n \xrightarrow{p} \theta$$