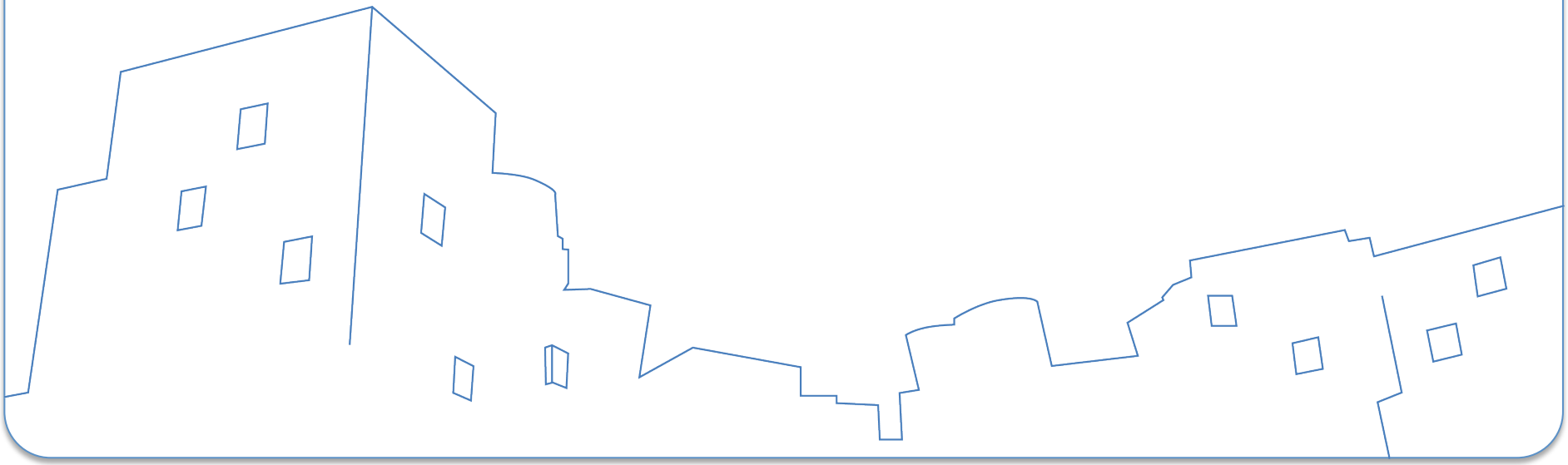
LIDS

6.434/16.391 Statistics for Engineers and Scientists

Lecture 10 10/07/2013

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology



Lecture 10 10/07/2013

INVARIANCE PROPERTY OF MAXIMUM LIKELIHOOD ESTIMATE

Induced likelihood function

- Definition (induced likelihood function, Casella & Berger):
  Let $\tau(\theta)$ be a function of θ, the induced likelihood function is defined as

$$L^*(\eta|\mathbf{x}) = \sup_{\{\theta:\tau(\theta)=\eta\}} L(\theta|\mathbf{x})$$

- The value $\hat{\eta}$ that maximizes $L^*(\eta|\mathbf{x})$ is called the MLE of $\eta = \tau(\theta)$

- The maxima of $L^*(\eta|\mathbf{x})$ and $L(\theta|\mathbf{x})$ coincide

Invariance property of MLE

- Theorem (Invariance property of MLE, Casella & Berger): If $\widehat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\widehat{\theta})$
- Proof: (Page 320, Casella & Berger) Let $\widehat{\eta}$ be the value that maximizes $L^*(\eta|\mathbf{x})$. We have

$$L^*(\widehat{\eta}|\mathbf{x}) = \sup_{\eta} \sup_{\{\theta:\tau(\theta)=\eta\}} L(\theta|\mathbf{x}) \qquad \text{definition of } L^*$$

$$\overset{(a)}{=} \sup_{\theta} L(\theta|\mathbf{x})$$

$$= L(\widehat{\theta}|\mathbf{x}) \qquad \text{definition of } \hat{\theta}$$

Furthermore, $L(\widehat{\theta}|\mathbf{x}) = \sup_{\{\theta:\tau(\theta)=\tau(\widehat{\theta})\}} L(\theta|\mathbf{x})$

$$= L^*(\tau(\widehat{\theta})|\mathbf{x}) \qquad \text{definition of } L^*$$

Therefore, $\tau(\widehat{\theta})$ is the MLE of $\tau(\theta)$

(a): iterative maximization = unconditional maximization

Lecture 10 10/07/2013

EXAMPLES OF MAXIMUM LIKELIHOOD ESTIMATOR

Example 1

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with $X_i \sim N\left(\mu, \sigma^2\right)$ where μ is known and σ^2 is unknown. Find the MLE of $\theta = \sigma^2$
- Note: Here $\Theta = \{r : r > 0\}$ since $0 < \sigma^2 < \infty$.
- First we have

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

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$

and

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

Example 1

- We choose σ^2 that maximizes $\ln L(\sigma^2|\mathbf{x})$, which is equivalent to minimize $g(\sigma^2)$, given by

$$g(\sigma^2) = n \ln \sigma + \frac{1}{2} \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^2$$

- Using the invariance property, we will find MLE of σ, i.e., $\hat{\sigma}$, instead of $\widehat{\sigma^2}$, since it is easier to differentiate $g(\sigma^2)$ with respect to σ, then set $\widehat{\sigma^2} = \hat{\sigma}^2$

Example 1

- Differentiating $g(\sigma^2)$ with respect to σ,

$$\frac{d}{d\sigma} g(\sigma^2) = \frac{n}{\sigma} - \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\sigma^3}$$

and

$$\frac{d^2}{d\sigma^2} g(\sigma^2) = -\frac{n}{\sigma^2} + \frac{3\sum_{i=1}^{n} (x_i - \mu)^2}{\sigma^4}$$

By setting $\frac{d}{d\sigma} g(\sigma^2) = 0$, we obtain

$$\sigma_* = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2}$$

Example 1

- And we have

$$\left.\frac{d^2}{d\sigma^2} g(\sigma^2)\right|_{\sigma=\sigma_*} = -\frac{n}{\sigma_*^2} + \frac{3n\sigma_*^2}{\sigma_*^4}$$

$$= \frac{2n}{\sigma_*^2} > 0$$

- Therefore, $g(\sigma^2)$ is minimized at σ_*, and MLE for σ is given by

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2}$$

and MLE for σ^2 is

$$\widehat{\sigma^2} = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

Please check: $\widehat{\sigma^2}$ is unbiased and consistent

Example 2

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. log-normal random variables, i.e.,

$$Y_i = \ln X_i \sim N(\mu, \sigma^2)$$

(Shadowing in mobile radio link can be modeled as log-normal random variable)

- Let

$$v = \mathbb{E}\{X_i\} = e^{\mu + \frac{\sigma^2}{2}}$$

$$\tau^2 = \mathbb{V}\{X_i\} = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right)$$

Find the MLEs of v and τ^2

- MLE of (μ, σ^2) are

$$\hat{\mu} = \overline{Y} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

Example 2

- By invariance property,

$$\hat{v} = e^{\overline{Y} + \frac{1}{2n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

$$\widehat{\tau^2} = \left[e^{2\overline{Y} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2}\right] \cdot \left[e^{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2} - 1\right]$$

Example 3

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli random variables with success probability p, i.e.,

$$\mathbb{E}\{X_i\} = p$$
$$\mathbb{V}\{X_i\} = p(1-p)$$

Find the MLE of mean and variance

- The likelihood function is

$$L(p|\mathbf{x}) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

and

$$\ln L(p|\mathbf{x}) = \sum x_i \ln p + \left(n - \sum x_i\right) \ln(1-p)$$

We have

$$\frac{\partial}{\partial p} \ln L(p|\mathbf{x}) = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p}$$
$$\frac{\partial^2}{\partial p^2} \ln L(p|\mathbf{x}) = -\frac{\sum x_i}{p^2} - \frac{n - \sum x_i}{(1-p)^2}$$

Example 3

- By setting $\frac{\partial}{\partial p} \ln L(p|\mathbf{x}) = 0$, we obtain

$$p_* = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and we also have

$$\frac{\partial^2}{\partial p^2} \ln L(p|\mathbf{x}) \Big|_{p=p_*} < 0$$

Therefore, p_* maximizes $\ln L(p|\mathbf{x})$. By the invariance property, the MLEs for the mean and variance are

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}$$

$$\widehat{\mathrm{Var}\{X_i\}} = \hat{p}\,(1 - \hat{p}) = \overline{X}(1 - \overline{X})$$

Indifferentiable likelihood function

- So far we differentiate the likelihood function with respect to the unknown parameter θ to find MLE.
- What if $L(\theta|\mathbf{x})$ is not differentiable?
- Recall: $L(\theta|\mathbf{x})$ is a function of θ. The X_i (after they have been observed) are held fixed in the entire process of finding MLE's
- Useful tool: Indicator function $\mathbb{I}_{\mathcal{S}}(t)$ of a set $\mathcal{S}$ is

$$\mathbb{I}_{\mathcal{S}}(t) = \begin{cases} 1, & \text{if } t \in \mathcal{S} \\ 0, & \text{otherwise} \end{cases}$$

Example 4

- Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from displaced exponential distribution with

$$f(x|\theta) = \begin{cases} e^{-(x-\theta)}, & x \geq \theta \\ 0, & \text{otherwise} \end{cases}$$

Find $\hat{\theta}$, the MLE of θ.

- First, the likelihood function can be written as

$$f(x|\theta) = e^{-(x-\theta)} \mathbb{I}_{[\theta,\infty)}(x)$$

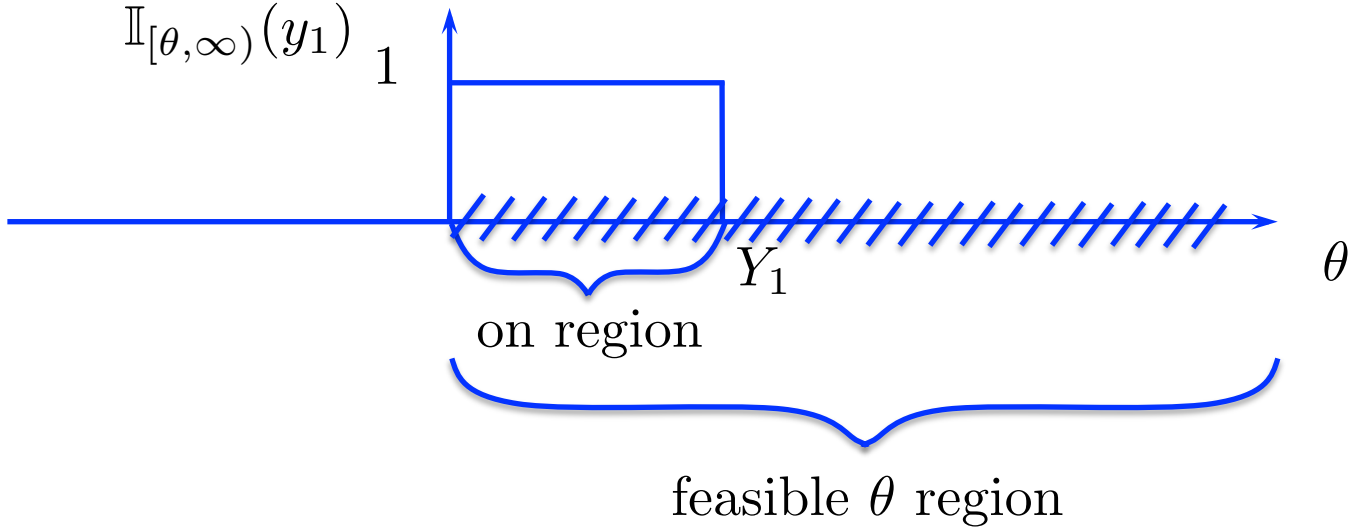
and

$$\begin{aligned} L(\theta|\mathbf{x}) &= e^{-\left(\sum_{i=1}^{n} x_i - n\theta\right)} \prod_{i=1}^{n} \mathbb{I}_{[\theta,\infty)}(x_i) \\ &= e^{-\sum_{i=1}^{n} x_i + n\theta} \mathbb{I}_{[\theta,\infty)}(y_1) \end{aligned}$$

where $y_1 = \min\{x_1, x_2, \ldots, x_n\}$

Example 4

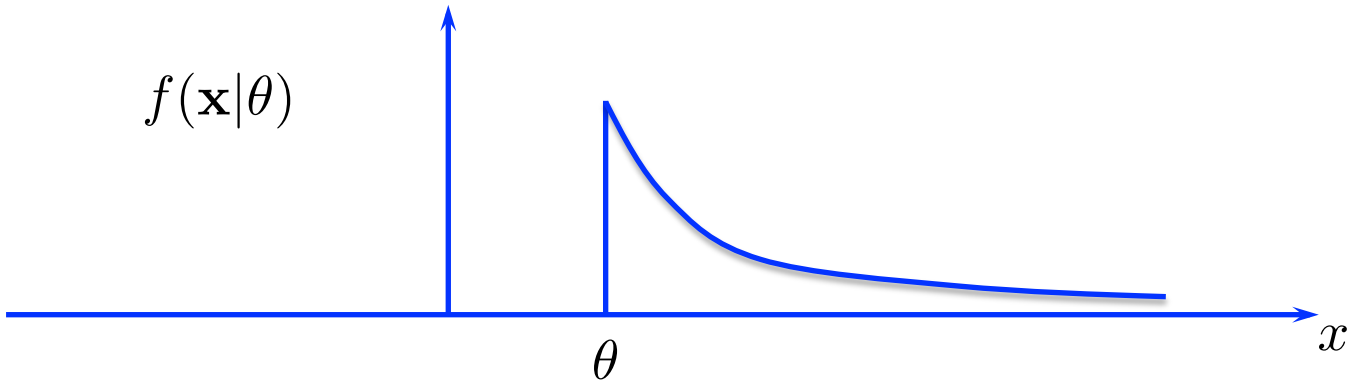
- $L(\theta|\mathbf{x})$ is the product of
  - increasing function of θ
  - and Indicator function.



- Therefore, $L(\theta|\mathbf{x})$ is maximized at $\hat{\theta} = y_1 = \min\{x_1, x_2, \ldots, x_n\}$ and MLE is $\hat{\theta}(\mathbf{X}) = \min\{X_1, X_2, \cdots, X_n\}$

Example 4

- Intuition: We draw samples from distribution with possible ranges $[\theta, \infty)$
  - X_i cannot be smaller than θ
  - X_i is more "likely" to be close to θ
  - Estimate the value of θ as large as possible that "explain" all of the observed data $X_1, X_2, \ldots, X_n$



Example 4

- Suppose $X_1, X_2, \ldots, X_n$ are independent random variables, each with uniform distribution on $[-d, d]$ where d is unknown, and $d \in \Theta = \{r : \ r > 0\}$. Find MLE of d
- First, we have

$$f_X(x|d) = \begin{cases} \frac{1}{2d} & -d \leq x \leq d \\ 0 & \text{otherwise} \end{cases}$$

which can be rewritten as

$$f_X(x|d) = \frac{1}{2d}\mathbb{I}_{[0,d]}(|x|)$$

Example 4

- The likelihood function is given by

$$L(d|\mathbf{x}) = \left(\frac{1}{2d}\right)^n \mathbb{I}_{[0,d]}(|x_1|) \ldots \mathbb{I}_{[0,d]}(|x_n|)$$

where $\mathbb{I}_{[0,d]}(|x_1|) \ldots \mathbb{I}_{[0,d]}(|x_n|) = 1$ iff $0 \leq |x_i| \leq d, \ \forall i$

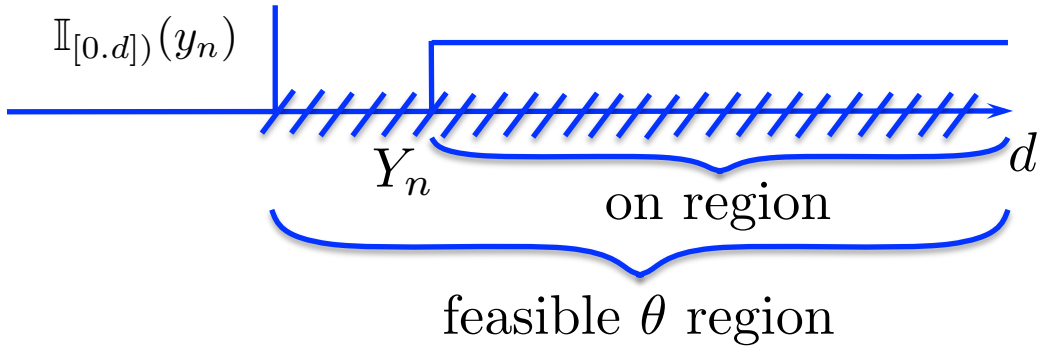
- Therefore, the likelihood function can be expressed as

$$L(d|\mathbf{x}) = \left(\frac{1}{2d}\right)^n \mathbb{I}_{[0,d]}(y_n)$$

where $y_n = \max\{|x_1|, |x_2|, \ldots, |x_n|\}$.

Example 4

- It is seen that $L(d|\mathbf{x})$ is a product of
  - a decreasing function of d
  - and an indicator function.



- Thus, the maximum likelihood estimate is given by

$$\widehat{d}(x_1, x_2, \ldots, x_n) = y_n = \max\{|x_1|, |x_2|, \ldots, |x_n|\}$$

and maximum likelihood estimator (MLE) is

$$\widehat{d}(X_1, X_2, \ldots, X_n) = Y_n = \max\{|X_1|, |X_2|, \ldots, |X_n|\}$$

Example 4

- Intuition: We observed sample from uniform distribution with possible ranges $[-d, d]$
  - $|X_i|$ cannot be larger than d (therefore we know that $d \geq |X_i|$)
  - Estimate the value of d as small as possible that "explains" the data

Example 5

- We draw samples from uniform distribution centered at a (a known parameter) with width d, i.e.,

$$X_i \sim \text{uniform}[a-d,\, a+d]$$
$$X_i - a \sim \text{uniform}[-d,\, d]$$

- Following similar analysis, MLE of d is

$$\widehat{d} = \max_i \{|X_i - a|\}$$

Example 6

- Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with uniform distribution in $[c-d,\, c+d]$, where d is known, but c is unknown, and $c \in \Theta = \mathbb{R}$. Find the MLE of c
- The likelihood function is

$$L(c|\mathbf{x}) = \prod_{i=1}^{n} \left(\frac{1}{2d}\right)^n \mathbb{I}_{[c-d,\, c+d]}(x_i)$$

$$= \left(\frac{1}{2d}\right)^n \mathbb{I}_{[c-d,\, c+d]}(y_1)\, \mathbb{I}_{[c-d,\, c+d]}(y_n)$$

where

$$y_1 = \min\{x_1, x_2, \ldots, x_n\}$$

$$y_n = \max\{x_1, x_2, \ldots, x_n\}$$

Example 6

- The likelihood function is a product of
  - a constant (i.e., constant function of c) and
  - two indicator functions.
- Therefore, MLE is any c such that $c - d \leq Y_1$ and $Y_n \leq c + d$
- Finally, the MLE satisfies

$$Y_n - d \leq \widehat{c} \leq Y_1 + d$$

$$\max\{X_1, X_2 \ldots, X_n\} - d \leq \widehat{c} \leq \min\{X_1, X_2, \ldots, X_n\} + d$$

- Note: In this example, MLE is not unique

Lecture 10 10/07/2013

EXAMPLES WITH MORE THAN ONE UNKNOWN PARAMETERS

Example 7

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. each with exponential distribution

$$f(x) = \begin{cases} \frac{1}{\theta_1} e^{-\frac{(x-\theta_2)}{\theta_1}} & x \geq \theta_2 \\ 0 & x < \theta_2 \end{cases}$$

  where $-\infty < \theta_2 < +\infty$ and $\theta_1 > 0$ are both unknown. Find MLEs of θ_1 and θ_2. Obtain $\mathbb{E}\{\widehat{\theta_1}\}$ and $\mathbb{E}\{\widehat{\theta_2}\}$

- The likelihood function is given by

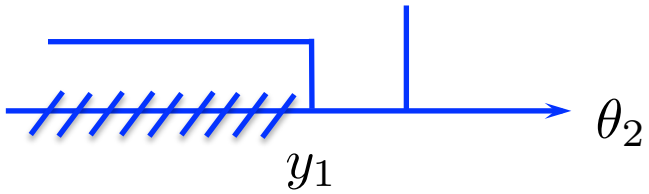
$$\begin{aligned} L(\theta_1, \theta_2 | \mathbf{x}) &= \prod_{i=1}^{n} f(x_i | \theta_1, \theta_2) \\ &= \frac{1}{\theta_1^n} e^{-\sum_{i=1}^{n} \frac{(x_i - \theta_2)}{\theta_1}} \mathbb{I}_{[\theta_2, \infty)}(x_1) \cdots \mathbb{I}_{[\theta_2, \infty)}(x_n) \\ &= \frac{1}{\theta_1^n} e^{-\sum_{i=1}^{n} \frac{(x_i - \theta_2)}{\theta_1}} \mathbb{I}_{[\theta_2, \infty)}(y_1) \end{aligned}$$

  where $y_1 = \min\{x_1, x_2, \ldots, x_n\}$

Example 7

- It can be seen that $L(\theta_1, \theta_2|\mathbf{x})$ is a product of
  - decreasing function of θ_1
  - increasing function of θ_2
  - and an indicator function

- For any fixed $\theta_1 > 0$, $L(\theta_1, \theta_2|\mathbf{x})$ is maximized by choosing
  - θ_2 as large as possible
  - $\mathbb{I}_{[\theta_2, \infty)}(y_1)$ is equal to1
  - Hence, $\widehat{\theta}_2 = Y_1 = \min\{X_1, X_2, \ldots, X_n\}$



Example 7

- Then we differentiate $L(\theta_1, \widehat{\theta}_2|\mathbf{x})$ or $\ln L(\theta_1, \widehat{\theta}_2|\mathbf{x})$ with respect to θ_1:

$$L(\theta_1, \widehat{\theta}_2) = \frac{1}{\theta_1^n} e^{-\sum_{i=1}^{n} \frac{(x_i - \widehat{\theta}_2)}{\theta_1}}$$

$$\ln L(\theta_1, \widehat{\theta}_2) = -\, n \ln \theta_1 - \frac{\sum_{i=1}^{n} (x_i - \widehat{\theta}_2)}{\theta_1}$$

$$\frac{\partial}{\partial \theta_1} \ln L(\theta_1, \widehat{\theta}_2) = -\, \frac{n}{\theta_1} + \frac{\sum_{i=1}^{n} (x_i - \widehat{\theta}_2)}{\theta_1^2}$$

$$\frac{\partial^2}{\partial \theta_1^2} \ln L(\theta_1, \widehat{\theta}_2) = +\, \frac{n}{\theta_1^2} - 2\frac{\sum_{i=1}^{n} (x_i - \widehat{\theta}_2)}{\theta_1^3}$$

Example 7

- Let the first derivative be zero, i.e.,

$$\frac{\partial}{\partial \theta_1} \ln L(\theta_1, \widehat{\theta}_2 | \mathbf{x}) = 0 \Rightarrow \theta_1^* = \frac{1}{n} \sum_{i=1}^{n} (x_i - \widehat{\theta}_2)$$

- The second derivative

$$\frac{\partial^2}{\partial \theta_1^2} \ln L(\theta_1, \widehat{\theta}_2 | \mathbf{x}) \Big|_{\theta_1 = \theta_1^*} = \frac{n}{\theta_1^{*2}} - \frac{2n}{\theta_1^{*2}}$$
$$< 0$$

Therefore, θ_1^* achieves maximum, and the MLE for θ_1 is

$$\begin{aligned} \widehat{\theta}_1 &= \frac{1}{n} \sum_{i=1}^{n} X_i - \widehat{\theta}_2 \\ &= \overline{X} - \min \{X_1, X_2, \ldots, X_n\} \end{aligned}$$

Example 7

- Next, we calculate $\mathbb{E}\{\widehat{\theta}_1\}$ and $\mathbb{E}\{\widehat{\theta}_2\}$
- First,

$$\mathbb{E}\{\overline{X}\} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{X_i\} = \theta_1 + \theta_2 \quad (\text{left as exercise})$$

Secondly,

$$\begin{aligned}
F_{Y_1}(y) =& \mathbb{P}\{Y_1 \leq y\} \\
=& 1 - \mathbb{P}\{Y_1 > y\} \\
=& 1 - \mathbb{P}\{\min\{X_1, X_2, \ldots, X_n\} > y\} \\
=& 1 - \mathbb{P}\{X_1 > y, X_2 > y, \ldots, X_n > y\} \\
=& 1 - \mathbb{P}\{X_1 > y\}\mathbb{P}\{X_2 > y\}\ldots\mathbb{P}\{X_n > y\} \\
=& 1 - [1 - F_X(y)]^n \\
=& 1 - \left[1 - \left(1 - e^{-\frac{X-\theta_2}{\theta_1}}\right)\right]^n, \quad \text{if } y \geq \theta_2 \\
=& 1 - e^{-n\left(\frac{y-\theta_2}{\theta_1}\right)}
\end{aligned}$$

Example 7

- Then we have

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{n}{\theta_1} e^{-n\left(\frac{y-\theta_2}{\theta_1}\right)}, \quad y \geq \theta_2$$

and

$$\begin{aligned}
\mathbb{E}\{Y_1\} &= \int_{\theta_2}^{\infty} y \frac{n}{\theta_1} e^{-n\left(\frac{y-\theta_2}{\theta_1}\right)} dy \\
&= \int_0^{\infty} (u + \theta_2) \frac{n}{\theta_1} e^{-\frac{nu}{\theta_1}} du \\
&= \theta_2 \int_0^{\infty} \frac{n}{\theta_1} e^{-\frac{nu}{\theta_1}} du + \int_0^{\infty} u \frac{n}{\theta_1} e^{-\frac{nu}{\theta_1}} du \\
&= \theta_2 \int_0^{\infty} e^{-z} dz + \frac{\theta_1}{n} \int_0^{\infty} z e^{-z} dz \\
&= \theta_2 + \frac{\theta_1}{n}
\end{aligned}$$

Example 7

- Thus,

$$\mathbb{E}\{\widehat{\theta_1}\} = \mathbb{E}\left\{\overline{X} - \min\{X_1, X_2, \ldots, X_n\}\right\}$$
$$= (\theta_1 + \theta_2) - \left(\theta_2 + \frac{\theta_1}{n}\right)$$
$$= \theta_1\left(1 - \frac{1}{n}\right)$$
$$\mathbb{E}\{\widehat{\theta_2}\} = \mathbb{E}\{\min\{X_1, X_2 \ldots, X_n\}\}$$
$$= \theta_2 + \frac{\theta_1}{n}$$

- Therefore, the estimator is biased, but asymptotically unbiased

Example 8

- Wildlife sampling
- Ecologist often desire to estimate the total population of a certain kind of animal in a given area
- Capture-Recapture method is often used (assuming that population size remain constant)
  - Sample "a" animals (capture a animals), tag them and release them
  - Wait sufficient time for thorough mixing with untagged population
  - Recapture n animals and count the number of tagged animals say X

Example 8

- We have

$$P_X(x) = \frac{\binom{a}{X}\binom{N-a}{n-X}}{\binom{N}{n}}$$

where we assume that probability of recapture is not dependent on whether an animal is tagged or untagged

  - Such assumption is often true if recapture is in the form of a "net" that captures all in a given sampling area
  - Fake in some degree if animals become shy of the apparatus used in the capture

Example 8

- The likelihood function of N is

$$L(N) = \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}$$

How do we maximize it?

- We look at the ratio

$$\begin{aligned} r(N) &= \frac{L(N)}{L(N-1)} \\ &= \frac{\binom{N-a}{n-x}}{\binom{N}{n}} \frac{\binom{N-1}{n}}{\binom{N-1-a}{n-x}} \\ &= \frac{\frac{(N-a)!}{(n-x)!(N-a-(n-x))!} \frac{(N-1)!}{n!(N-1-n)!}}{\frac{N!}{n!(N-n)!} \frac{(N-1-a)!}{(n-x)!(N-1-a-(n-x))!}} \\ &= \frac{\frac{N-a}{N-a-n+x}}{\frac{N}{N-n}} \end{aligned}$$

Example 8

- Finally, the ratio is given by

$$r(N) = \frac{N-a}{N-a-n+x}\frac{N-n}{N}$$

- Ratio $r(N) < 1$ if and only if

$$(N-a)(N-n) < N(N-a-n+x)$$
$$N^2 - Nn - aN + an < N^2 - Na - Nn + Nx$$
$$an < Nx$$
$$\frac{an}{x} < N$$

and $r(N) > 1$ if and only if

$$\frac{an}{x} > N$$

Example 8

- Thus, $L(N)$ decreases if $N > \frac{an}{x}$ and increases if $N < \frac{an}{x}$

- If $\frac{an}{x}$ is integer, $\widehat{N} = \frac{an}{x}$
  otherwise, $\widehat{N} = \lceil \frac{an}{X} \rceil$ or $\lfloor \frac{an}{X} \rfloor$, depending on which results in larger $L(N)$