ST5215 Advanced Statistical Theory, Lecture 22

HUANG Dongming

National University of Singapore

3 Nov 2020

Overview

Last time

- Roots of the Likelihood Equation (RLE)
- Asymptotic Normality of RLEs and MLEs

Today

- Asymptotic Efficiency
- Linear Models

Recap: Asymptotics of RLEs

- Under some basic regularity conditions, there exists a sequence of roots of the likelihood equation (RLEs) $\hat{\theta}_n$ that is stronger consistent and $\sqrt{n}\left(\tilde{\theta}_n - \theta_*\right) \xrightarrow{\mathcal{D}} N(\mathbf{0}, [I(\theta_*)]^{-1})$
- For the consistency, we use USLLN to get uniform control of $\frac{1}{n}\log\frac{L_n(\theta)}{L_n(\theta_*)}$ over $\{\theta : \|\theta - \theta_*\| = \rho\}$
- For the asymptotic normality,
  - we use the mean-value theorem (or Taylor expansion) to make a connection between $\left(\tilde{\theta}_n - \theta_*\right)$ and the score function

$$s_n(\theta) = \frac{\partial}{\partial\theta}\log L_n(\theta)$$

  - By CLT, $n^{-1/2}s_n(\theta_*)$ is asymptotically normal
  - Then we use USLLN to control $\left\|n^{-1}\nabla s_n(\theta) - I(\theta_*)\right\|$ over any small closed ball around θ_*

Asymptotic Efficiency (1)

- Let $\{\hat{\theta}_n\}$ be a sequence of estimators of θ based on a sequence of samples $\{X = (X_1, ..., X_n) : n = 1, 2, ...\}$ and the distributions of the samples are in a parametric family indexed by $\theta \in \Theta \subset \mathcal{R}^k$
- Suppose that
$$[V_n(\theta)]^{-1/2}(\hat{\theta}_n - \theta) \xrightarrow{D} N_k(0, I_k), \tag{1}$$
where $V_n(\theta)$ is a $k \times k$ positive definite matrix depending on θ, and is called *the asymptotic covariance matrix* (or *asymptotic variance* if $k = 1$)
- Since the asymptotic covariance matrices are unique only in the limiting sense, we have to make our comparison based on their limits.
- When X_i's are i.i.d., $V_n(\theta)$ is usually of the form $n^{-\delta}V(\theta)$ for some $\delta > 0$ (and $= 1$ in the majority of cases) and a positive definite matrix $V(\theta)$ that does not depend on n

Asymptotic Efficiency (2)

- Suppose two estimators $\hat{\theta}_{1n}$ and $\hat{\theta}_{2n}$ satisfy Equation (1) with $V_{1n}(\theta)$ and $V_{1n}(\theta)$. If $V_{1n}(\theta) \preceq V_{2n}(\theta)$ for all $\theta \in \Theta$ and all large n and $V_{1n}(\theta) \prec V_{2n}(\theta)$ for at least one $\theta \in \Theta$, then $\hat{\theta}_{1n}$ is said to be *asymptotically more efficient* than $\hat{\theta}_{2n}$
  - For two $k \times k$ matrices, $A \preceq B$ means $B - A$ is positive semi-definite; $A \prec B$ means $B - A$ is positive definite
- The Cramèr-Rao lower bound says that, if $\hat{\theta}_n$ is unbiased, then under some regularity conditions,

$$Var(\hat{\theta}_n) \succeq [I_n(\theta)]^{-1},$$

  where $I_n(\theta)$ is the Fisher information matrix with n samples.
- If $\hat{\theta}_n$ satisfies Equation (1), it is asymptotically unbiased, but the following may not hold even if the regularity conditions in the Cramèr-Rao lower bound are satisfied:

$$V_n(\theta) \succeq [I_n(\theta)]^{-1} \tag{2}$$

Example: Hodges' estimator

Let $X_1, ..., X_n$ be i.i.d. from $N(\theta, 1)$, $\theta \in \mathcal{R}$. Then $I_n(\theta) = n$, and the CR-lower bound for estimating θ is $1/n$

- For any constant $0 < t < 1$, define

$$\hat{\theta}_n = \begin{cases} \bar{X}_n & |\bar{X}_n| \geq n^{-1/4} \\ t\bar{X}_n & |\bar{X}_n| < n^{-1/4}, \end{cases} \tag{3}$$

- By Proposition 3.2, the conditions about exchanging the differentiation and the integration in C-R lower bound are satisfied
- If $\theta \neq 0$, we have
$\sqrt{n}(\hat{\theta}_n - \theta) = \sqrt{n}(\bar{X}_n - \theta) - (1-t)\sqrt{n}\bar{X}_n I_{|\bar{X}_n| < n^{-1/4}} \xrightarrow{\mathcal{D}} N(0, 1)$,
because the second term $\xrightarrow{\mathcal{P}} 0$ and by using Slutsky's theorem
- If $\theta = 0$, then $\sqrt{n}(\hat{\theta}_n - \theta) = t\sqrt{n}\bar{X}_n \xrightarrow{\mathcal{D}} N(0, t^2)$
- So

$$V_n(\theta) = \begin{cases} 1/n & \text{if } \theta \neq 0 \\ t^2/n, & \text{if } \theta = 0 \end{cases}$$

Remark

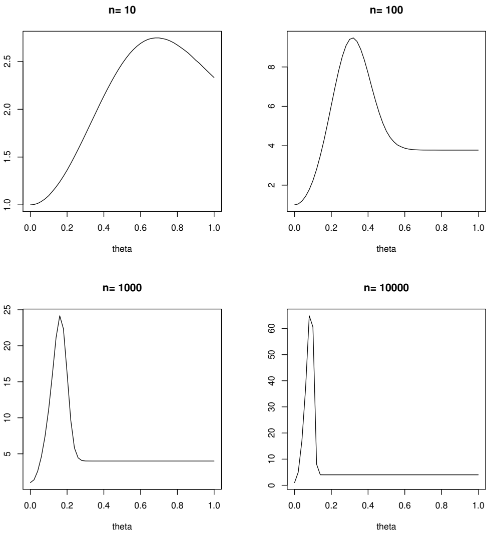
At first sight, $\hat{\theta}_n$ is an improvement on $\bar{X}_n$:

- For every $\theta \neq 0$, the estimators behave the same, while for $\theta = 0$, the sequence $\hat{\theta}_n$ has an smaller amse

However, this reasoning is a bad use of asymptotics

- The risk of $\bar{X}_n$ is $1/n$, which is a constant in θ
- The risk function of $\hat{\theta}_n$ is $R_n(\theta) = E_\theta(\hat{\theta}_n - \theta)^2$

$R_n(\theta)/R_n(0)$



- The peak of $R_n(\theta)$ over θ is much higher than $R_n(0)$; as n increases, the ratio goes to ∞
- Compared to the UMVUE ($\bar{X}_n$), $\hat{\theta}_n$ buys its better asymptotic behavior at 0 at the expense of worse performance for other θ's with any n
- Because the values of θ at which $\hat{\theta}_n$ is bad differ from n to n, the erratic behavior is not visible in the point-wise limit distributions under fixed θ

- Points at which the information inequality (2) fails are called *points of superefficiency*
- The following result says that the set of points of superefficiency is often of Lebesgue measure 0 under some regularity conditions

Theorem (Theorem 4.16 in JS)

Under the same conditions in the theorem "Asymptotic Normality of RLEs" in Lecture 21, if $\hat{\theta}_n$ is an estimator of θ satisfies Equation (1), then there is a $\Theta_0 \subset \Theta$ with Lebesgue measure 0 such that the information inequality (2) holds for any $\theta \notin \Theta_0$.

Asymptotic efficiency (3)

Definition

Assume that the Fisher information matrix $I_n(\theta)$ is well defined and positive definite for every n. A sequence of estimators $\{\hat{\theta}_n\}$ that satisfies Equation (1) is said to be *asymptotically efficient* or *asymptotically optimal* if and only if $V_n(\theta) = [I_n(\theta)]^{-1}$

- Suppose that we are interested in estimating $\vartheta = g(\theta)$, where g is a differentiable function from Θ to $\mathcal{R}^p$, $1 \leq p \leq k$
- If $\hat{\theta}_n$ satisfies Equation (1), then for $\hat{\vartheta}_n = g(\hat{\theta}_n)$, $\left([\nabla g(\theta)]^\top V_n(\theta) \nabla g(\theta)\right)^{-1/2} (\hat{\vartheta}_n - \vartheta) \xrightarrow{\mathcal{D}} N(0, I_p)$
- If $p = k$ and g is one-to-one, we can check that the information inequality for ϑ is equivalent to the one for θ
- For this reason, for general g, $\hat{\vartheta}_n$ is defined to be *asymptotically efficient* if and only if $\hat{\theta}_n$ is asymptotically efficient

Asymptotically efficient estimators

- In lecture 21, we have show that under some regularity conditions, any consistent sequence of RLEs satisfies

$$\sqrt{n}\left(\tilde{\theta}_n - \theta\right) \xrightarrow{\mathcal{D}} N(\mathbf{0}, [I(\theta)]^{-1}),$$

which implies that this sequence of RLEs is asymptotically efficient
- The following *one-step MLE* is often also asymptotically efficient

We begin with any estimator $\hat{\theta}_n^{(0)}$, and define an estimator

$$\hat{\theta}_n^{(1)} = \hat{\theta}_n^{(0)} - \left[\nabla s_n\left(\hat{\theta}_n^{(0)}\right)\right]^{-1} s_n\left(\hat{\theta}_n^{(0)}\right),$$

where $s_n = \frac{\partial}{\partial\theta}\log L_n(\theta)$ is the score function.

- This is the first iteration in computing an MLE (or RLE) using the Newton–Raphson iteration method with $\hat{\theta}_n^{(0)}$ as the initial value.
- Under the same regularity conditions as before, and if $\hat{\theta}_n^{(0)}$ is $\sqrt{n}$-consistent, then $\hat{\theta}_n^{(1)}$ is asymptotically efficient (left for exercise; use the same techniques in Lecture 21 or check Theorem 4.19 (i) in JS)

Linear models

A linear model is given below:

$$X_i = Z_i^\top \boldsymbol{\beta} + \epsilon_i, \qquad i = 1, ..., n, \tag{4}$$

- X_i is the value of a response variable observed on the ith individual;
- Z_i is the value of a p-vector of explanatory variables (non-random covariates) observed on the ith individual;
- $\boldsymbol{\beta}$ is a p-vector of unknown parameters (main parameters of interest), $p < n$;
- ϵ_i is a random error (not observed) associated with the ith individual.

Suppose that the range of $\boldsymbol{\beta}$ in model (5) is $B \subset \mathcal{R}^p$.

Matrix Forms

Let

- $X = (X_1, ..., X_n)^\top$: the vector of responses
- $\epsilon = (\epsilon_1, ..., \epsilon_n)^\top$: the vector of noise
- Z be the $n \times p$ matrix whose ith row is the vector $Z_i^\top$, $i = 1, ..., n$: the design matrix, or the matrix of covariates

A matrix form of the model is

$$X = Z\boldsymbol{\beta} + \epsilon. \tag{5}$$

LSE

- A *least squares estimator (LSE)* of $\boldsymbol{\beta}$ is defined to be any $\hat{\boldsymbol{\beta}} \in B$ such that
$$\|X - Z\hat{\boldsymbol{\beta}}\|^2 = \min_{\mathbf{b} \in B} \|X - Z\mathbf{b}\|^2. \tag{6}$$
- For any $a \in \mathcal{R}^p$, $a^\top \hat{\beta}$ is called an *LSE of* $a^\top \beta$.
- From now on, assume $B = \mathcal{R}^p$ unless otherwise stated.
- Differentiating $\|X - Z\mathbf{b}\|^2$ w.r.t. $\mathbf{b}$, we obtain the normal equation
$$Z^\top Z\mathbf{b} = Z^\top X. \tag{7}$$
    - $g(\mathbf{b}) = \|X - Z\mathbf{b}\|^2 = (X - Z\mathbf{b})^\top (X - Z\mathbf{b})$ is a quadratic form
    - $\frac{\partial}{\partial \mathbf{b}}(\mathbf{b}^\top A\mathbf{b}) = 2A\mathbf{b}$ and $\frac{\partial}{\partial \mathbf{b}}(\mathbf{b}^\top A\mathbf{c}) = A\mathbf{c}$
- Any solution of the normal equation is an LSE of β.

Expression for a LSE

The case of full rank Z: If the rank of the matrix Z is p, in which case $(Z^\top Z)^{-1}$ exists and Z is said to be of full rank, then there is a unique LSE, which is

$$\hat{\beta} = (Z^\top Z)^{-1} Z^\top X. \tag{8}$$

The case of non full rank Z: If Z is not of full rank, then $\boldsymbol{\beta}$ is *not identifiable* because there exist $\tilde{\boldsymbol{\beta}} \neq \boldsymbol{\beta}$ but $Z\boldsymbol{\beta} = Z\tilde{\boldsymbol{\beta}}$

- In terms of estimation, there are infinitely many LSE's of β.
- Any LSE of β is of the form

$$\hat{\beta} = (Z^\top Z)^{-} Z^\top X, \tag{9}$$

where $(Z^\top Z)^{-}$ is called a *generalized inverse* of $Z^\top Z$ and satisfies

$$Z^\top Z (Z^\top Z)^{-} Z^\top Z = Z^\top Z. \tag{10}$$

Some properties of generalized inverse

- If Z is of full rank, $(Z^\top Z)^-$ is unique and equal to $(Z^\top Z)^{-1}$
- If Z is not of full rank, generalized inverse matrices are not unique
- If the singular value decomposition of Z is $UDV^\top$, where U is an $n \times n$ orthogonal matrix, V is an $p \times p$ orthogonal matrix, and D is of the form $\begin{pmatrix} D_* & 0 \\ 0 & 0 \end{pmatrix}$, where $D_* = \mathrm{diag}\{\lambda_1, \ldots, \lambda_r\}$ with λ_i's positive.
  Then $(Z^\top Z)^-$ can be constructed by
  $$V \begin{pmatrix} D_*^{-2} & A \\ A^\top & \tilde{D} \end{pmatrix} V^\top,$$
  where $\tilde{D}$ is any $(p-r) \times (p-r)$ matrix and A is any $r \times (p-r)$ matrix
- $Z(Z^\top Z)^- Z^\top$ is a projection matrix in to the column space of Z
  - $[Z(Z^\top Z)^- Z^\top]^2 = Z(Z^\top Z)^- Z^\top$
  - $Z(Z^\top Z)^- Z^\top Z = Z$
  - The rank of $Z(Z^\top Z)^- Z^\top$ is $\mathrm{tr}(Z(Z^\top Z)^- Z^\top) = r$

Simple linear regression

- Suppose $p = 2$. Let $\beta = (\beta_0, \beta_1) \in \mathcal{R}^2$ and $Z_i = (1, t_i), t_i \in \mathcal{R}, i = 1, \ldots, n$.
- Then model (5) is called a simple linear regression model.
- It turns out that

$$Z^\top Z = \begin{pmatrix} n & \sum_{i=1}^n t_i \\ \sum_{i=1}^n t_i & \sum_{i=1}^n t_i^2 \end{pmatrix},$$

  which is invertible if and only if some t_i 's are different.
- If some t_i 's are different, then the LSE $\hat{\beta}$ of β is given by $\left(Z^\top Z\right)^{-1} Z^\top X$
- If we assume ϵ_i's are i.i.d. normal, then $\hat{\beta}$ has the normal distribution. Furthermore, we can check that this LSE is also the MLE for β

The result can be easily extended to the case of polynomial regression of order p in which $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})$ and $Z_i = \left(1, t_i, \ldots, t_i^{p-1}\right)$

Tutorial

1. Exercise 1.6.146 in JS
2. Exercise 1.6.155 in JS
3. [Neyman and Scott (1948)] Suppose we have a sample of size d from each of n normal populations with common unknown variance but possibly different unknown means
$$X_{ij} \in \mathcal{N}\left(\mu_i, \sigma^2\right), \quad I = 1, \ldots, n, \quad j = 1, \ldots, d$$
where all the X_{ij} are independent.
(a) Find the maximum-likelihood estimate of σ^2.
(b) Show that for d fixed, the MLE of σ^2 is not consistent as $n \to \infty$. Why doesn't Theorem of Consistency of MLE apply?
(c) Find a consistent estimate of σ^2.
4. Let $X = (X_1, \ldots, X_n)$ be a random sample of random variables with probability density f_θ. Find an MLE of θ and its asymptotic distribution in each of the following cases
(i) $f_\theta(x) = e^{-(x-\theta)} I_{(\theta,\infty)}(x), \theta > 0$
(ii) $f_\theta(x) = \theta(1-x)^{\theta-1} I_{(0,1)}(x), \theta > 1$

Exercise 1.6.146 in JS

Let $U_1, U_2, \ldots$ be i.i.d. random variables having the uniform distribution on $[0,1]$ and $Y_n = \left(\prod_{i=1}^n U_i\right)^{-1/n}$. Show that $\sqrt{n}\,(Y_n - e) \to_d N\left(0, e^2\right)$

Proof:

- Let $X_i = -\log U_i$. Then $X_1, X_2, \ldots$ are independent and identically distributed random variables with $EX_1 = 1$ and $\operatorname{Var}(X_1) = 1$.
- By the CLT, $\sqrt{n}\left(\bar{X}_n - 1\right) \to_d N(0,1)$,
- Note that $Y_n = e^{\bar{X}_n}$. Applying the δ-method with $g(t) = e^t$ to $\bar{X}_n$, we obtain that $\sqrt{n}\,(Y_n - e) \xrightarrow{\mathcal{D}} N\left(0, e^2\right)$, since $g(1) = e$ and $g'(1) = e$

Exercise 1.6.155 in JS

Let $\{X_n\}$ be a sequence of random variables and let $\bar{X} = \sum_{i=1}^n X_i/n$
(a) Show that if $X_n \to_{a.s.} 0$, then $\bar{X} \to_{a.s.} 0$

Proof: Part (a)

- Fixed any ω for which $X_n(\omega) \to 0$. Let $x_n = X_n(\omega)$.
- Generally, if $x_n \to 0$, then $\frac{|\sum_{i=1}^n x_i|}{n} \to 0$
- For any $\epsilon > 0$, there exists a N_1 s.t. for any $n \geq N_1$, $|x_n| < \epsilon$.
- Let N_2 to be larger than $N_1 \max_{1 \leq i \leq N_1} |x_i|/\epsilon$.
- For any $n > \max(N_1, N_2)$,

$$\frac{|\sum_{i \leq n} x_i|}{n} \leq n^{-1} \sum_{N_1 \leq i \leq n} |x_i| + n^{-1} \sum_{i < N_1} |x_i| \leq \epsilon + \epsilon,$$

which implies that $\frac{|\sum_{i \leq n} x_i|}{n} \to 0$

Part (b)

Show that if $X_n \to_{L_r} 0$, then $\bar{X} \to_{L_r} 0$, where $r \geq 1$ is a constant.

- When $r \geq 1$, $|x|^r$ is a convex function.
- By Jensen's inequality, $E\left|\bar{X}_n\right|^r \leq n^{-1}\sum_{i=1}^n E\left|X_i\right|^r$.
- Since $\lim_n E\left|X_n\right|^r = 0$, by the result in part (a), $\lim_n n^{-1}\sum_{i=1}^n E\left|X_i\right|^r = 0$
- Hence, $\lim_n E\left|\bar{X}_n\right|^r = 0$.

Part (c)

Show that the result in (b) may not be true for $r \in (0,1)$

- Let $U \sim \text{Unif}(0,1)$.
- Let $A_n = \{U \in (1/(n+1), 1/n)\}$ and $X_n = n(n+1)^{1/r} I_{A_n}$ for each $n = 1, 2, \ldots$
- Then $P(X_n \neq 0) = 1/(n(n+1))$ and $E|X_n|^r = n^{r-1} \to 0$ since $r < 1$
- Note that A_n's are disjoint, so

$$E|\sum_{i \leq n} X_i|^r = \sum_{i \leq n} E I_{A_i} |X_i|^r = \sum_{i \leq n} n^{r-1} \geq \int_1^{n+1} x^{r-1} \, \mathrm{d}x = [(n+1)^r - 1]$$

- This implies that

$$E|\bar{X}_n|^r = n^{-r} E|\sum_{i \leq n} X_i|^r \geq \left[(1+1/n)^r - n^{-r}\right]/r \to 1/r$$

Part (d)

(d) Show that $X_n \to_p 0$ may not imply $\bar{X} \to_p 0$

- Construct independent X_n's such that $P(X_n = n) = 1 - P(X_n = 0) = 1/n$.
- $P(|X_n| > 0) = 1/n \to 0$
- Note that $EX_n = 1$ and $Var(X_n) = n - 1$
- For any small t such that $1 - t > 1/\sqrt{2}$, Chebyshev's inequality implies that

$$
\begin{aligned}
P(\sum_{i \le n} X_i < tn) &= P\left(n - \sum_{i \le n} X_i > (1-t)n\right) \\
&\le \mathrm{Var}\left(\sum_{i \le n} X_i\right) / \left((1-t)^2 n^2\right) \\
&= \frac{n(n-1)/2}{(1-t)^2 n^2} < \left[2(1-t)^2\right]^{-1} < 1
\end{aligned}
$$

- So $P(\bar{X}_n \ge t) > 1 - \left[2(1-t)^2\right]^{-1} > 0$ for all n

Exercise 3

Suppose we have a sample of size d from each of n normal populations with common unknown variance but possibly different unknown means

$$X_{ij} \in \mathcal{N}\left(\mu_i, \sigma^2\right), \quad I = 1, \ldots, n, \quad j = 1, \ldots, d$$

where all the X_{ij} are independent.

(a) Find the maximum-likelihood estimate of σ^2.

Proof:

- The maximization over μ_i can be done as usual, and $\hat{\mu}_i = \bar{X}_i = d^{-1} \sum_{j \le d} X_{ij}$
- The derivate of $\log L_n$ w.r.t. σ is

$$\frac{\partial}{\partial \sigma} \log L = -\frac{nd}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} \sum_{j=1}^{d} (X_{ij} - \mu_i)^2 = 0,$$

whose solution is

$$\widehat{\sigma^2} = \frac{1}{nd} \sum_{i=1}^{n} \sum_{j=1}^{d} (X_{ij} - \hat{\mu}_i)^2 = \frac{1}{n} \sum_{i=1}^{n} s_i^2$$

- Here s_i^2 is the empirical variance for the ith population and $= (1/d) \sum_{j=1}^{d} \left(X_{ij} - \bar{X}_i\right)^2$

Part (b) and (c)

Show that for d fixed, the MLE of σ^2 is not consistent as $n \to \infty$. Why doesn't the Theorem of Consistency of MLE apply? Find a consistent estimate of σ^2.

- The s_i^2 are i.i.d. with mean $Es_i^2 = ((d-1)/d)\sigma^2$.
- By SLLN, $\hat{\sigma}^2 \stackrel{a.s.}{\to} \frac{d-1}{d}\sigma^2$ almost surely. So $\hat{\sigma}^2$ is not consistent.
- Here the number of parameters grows to infinity as $n \to \infty$, so the structure of the problem differs from that of Theorem of Consistency of MLE.
- A consistent estimator is given by $\frac{d}{d-1}\hat{\sigma}^2$

Exercise 4

Let $X = (X_1, \ldots, X_n)$ be a random sample of random variables with probability density f_θ. Find an MLE of θ and its asymptotic distribution in each of the following cases
(i) $f_\theta(x) = e^{-(x-\theta)} I_{(\theta,\infty)}(x), \theta > 0$
(ii) $f_\theta(x) = \theta(1-x)^{\theta-1} I_{(0,1)}(x), \theta > 1$

Proof: Part (i)

- Let $X_{(1),n}$ be the smallest order statistic for data of size n.
- The likelihood function is $\ell(\theta) = \exp\left\{-\sum_{i=1}^n (X_i - \theta)\right\} I_{\left(0, X_{(1),n}\right)}(\theta)$, which is 0 when $\theta > X_{(1),n}$ and increasing on $\left(0, X_{(1),n}\right)$.
- Hence, the MLE of θ is $X_{(1),n}$
- For any δ,
  $P(X_{(1),n} > \theta + \delta) = \prod_{i \le n} P(X_i - \theta > \delta) = \left(\int_\delta^\infty e^{-x}\,\mathrm{d}x\right)^n = e^{-n\delta}$
- By the second Borel-Cantelli lemma, we can show that $P(X_{(1),n} > \theta + \delta, i.o.) = 0$ so $X_{(1),n} \overset{a.s.}{\to} \theta$
- So for any $t > 0$, $P(n\left[X_{(1),n} - \theta\right] \le t) = 1 - e^{-t}$, or

$$n\left[X_{(1),n} - \theta\right] \overset{\mathcal{D}}{\to} \mathrm{E}(0,1)$$

Exercise 4 Part (ii)

$f_\theta(x) = \theta(1-x)^{\theta-1} I_{(0,1)}(x), \theta > 1$

- Note that $\ell(\theta) = \theta^n \prod_{i=1}^n (1-X_i)^{\theta-1} I_{(0,1)}(X_i)$ and, when $\theta > 1$

$$\frac{\partial \log \ell(\theta)}{\partial \theta} = \frac{n}{\theta} + \sum_{i=1}^n \log(1-X_i) \quad \text{and} \quad \frac{\partial^2 \log \ell(\theta)}{\partial \theta^2} = -\frac{n}{\theta^2} < 0$$

- The equation $\frac{\partial \log \ell(\theta)}{\partial \theta} = 0$ has a unique solution $\hat{\theta} = -n / \sum_{i=1}^n \log(1-X_i)$
- If $\hat{\theta} > 1$, then it maximizes $\ell(\theta)$ and is the MLE
- If $\hat{\theta} \leq 1$, then $\ell(\theta)$ is decreasing on the interval $(1, \infty)$ and the MLE does not exist (or is 1 according to the general definition)
- Let $Y_i = -\log(1-X_i)$. Then p.d.f. of Y_i is $\theta \exp(-\theta y) I_{y>0}$
- By CLT, $\sqrt{n}\left(\bar{Y}_n - 1/\theta\right) \xrightarrow{\mathcal{D}} N(0, 1/\theta^2)$
- By δ-method with $g(t) = 1/t$ and $g'(t) = -1/t^2$, $\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{\mathcal{D}} N(0, \theta^2)$