统计极限理论

作者：Yukina

时间：2025 年 12 月 31 日

目录

Chapter 1

测度理论

§1.1 集类与其运算

1.1.1 随机事件

Definition 1.1 设 $\{A_n\}_{n=1}^{\infty}$ 是一个集序列，则

1. A_n 的上极限定义为
$$\limsup_n A_n := \{\omega\text{属于无穷多个}A_n\}$$
2. A_n 的下极限定义为
$$\liminf_n A_n := \{\omega\text{不属于有限多个}A_n\}$$

Proposition 1.2 设 $\{A_n\}_{n=1}^{\infty}$ 是一个集序列

1. $\liminf_n A_n = \bigcup_{n=1}^{\infty}\bigcap_{m=n}^{\infty} A_m, \limsup_n A_n = \bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty} A_m$
2. $\liminf_n A_n \subseteq \limsup_n A_n$

Definition 1.3 设 $\mathcal{C}$ 为一集类.

1. 称 $\mathcal{C}$ 为$\underline{\pi\text{ 类}}$，如果它对有限交封闭.
2. 称 $\mathcal{C}$ 为<u>半环</u>，如果 $\varnothing \in \mathcal{C}$ ，且有
$$A, B \in \mathcal{C} \implies A \cap B \in \mathcal{C}, A \backslash B \in \mathcal{C}_{\Sigma f}$$
3. 称 $\mathcal{C}$ 为<u>半代数</u>，如果它是半环，且 $\Omega \in \mathcal{C}$.
4. 称 $\mathcal{C}$ 为<u>单调类</u>，如果它对单调序列极限封闭.

Definition 1.4 设 $\mathcal{C}$ 为一集类，称 $\mathcal{C}$ 为<u>代数</u>，如果

1. $\varnothing \in \mathcal{C}, \Omega \in \mathcal{C}$

2. 对有限并和补运算封闭，即

$$E_1, E_2 \in \mathcal{C} \implies E_1 \cup E_2 \in \mathcal{C}, \quad E \in \mathcal{C} \implies E^c \in \mathcal{C}$$

如果上述运算可以操作可数次，则称 $\mathcal{C}$ 为 σ 代数.

Remark 1.5 由定义可以立即推断代数对“交”，“并”，“差”，“补”都封闭.

Proposition 1.6 设 $\mathcal{C}$ 为半代数，则 $\mathcal{A}(\mathcal{C}) = \mathcal{C}_{\Sigma f}$

Definition 1.7 设 $\mathcal{C}$ 是一集类，则存在唯一的 σ-代数 (记为 $\sigma(\mathcal{C})$)，具有下列性质：

1. $\mathcal{C} \subset \sigma(\mathcal{C})$

2. $\sigma(\mathcal{C})$ 是包含 $\mathcal{C}$ 的最小 σ-代数.

称其为 $\mathcal{C}$ 生成的 σ-代数.

Proof. 验证；良定义：由于 $\mathcal{P}(\Omega)$ 是一 σ 代数，于是包含 $\mathcal{C}$ 的 σ-代数全体构成的集合是非空的. 考虑该集合中所有元素的交，自动满足存在唯一性. □

1.1.2 乘积空间与乘积 σ 代数

Definition 1.8 集合 A_1 和 A_2 的乘积集记为 $A_1 \times A_2$ ，是指的下列集合：

$$A_1 \times A_2 = \{(\omega_1, \omega_2) : \omega_1 \in A_1, \omega_2 \in A_2\}.$$

称 $\Omega_1 \times \Omega_2$ 为乘积空间，$A_1 \times A_2$ 为其中的矩形.

Definition 1.9 设 $(\Omega_1, \mathcal{A}_1)$ 和 $(\Omega_2, \mathcal{A}_2)$ 是两个可测空间. 定义集合

$$E = \{A_1 \times A_2 : A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2\}$$

称包含 E 的最小 σ-代数 $\sigma(E)$ 为 $\mathcal{A}_1$ 和 $\mathcal{A}_2$ 的乘积 σ-代数，记为 $\mathcal{A}_1 \times \mathcal{A}_2$.

Remark 1.10 E 本身是一个半集代数，但不是 σ-代数.

Definition 1.11 设 $(\Omega_i, \mathcal{A}_i), i = 1, \ldots, n$ 是 n 个可测空间. 定义

$$E_n = \left\{ \prod_{i=1}^{n} A_i : A_i \in \mathcal{A}_i, i = 1, \ldots, n \right\}$$

乘积 σ-代数定义为

$$\bigotimes_{i=1}^{n} \mathcal{A}_i = \sigma(E_n)$$

Example 1.12 (Borel σ-代数的乘积) 当 $\Omega_1 = \Omega_2 = \mathbb{R}, \mathcal{A}_1 = \mathcal{A}_2 = \mathcal{B}(\mathbb{R})$ 时，有

$$\mathcal{B}(\mathbb{R}^2) = \mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R})$$

即二维 Borel σ-代数等于两个一维 Borel σ-代数的乘积 σ-代数.

§1.2 集函数，测度，概率

1.2.1 集函数与测度

Definition 1.13 (集函数) 设 C 为任一集类. 映射

$$\mu : C \to [-\infty, \infty]$$

称为 C 上的<u>集函数</u>. 若 μ 值域为 $[0,\infty]$ 则称其为 C 上的<u>非负集函数</u>.

Remark 1.14 在测度论中，我们一般约定非负集函数 μ 满足

1. $\mu(\varnothing) = 0$
2. 单调性:

$$A, B \in C, A \subset B \implies \mu(A) \leqslant \mu(B)$$

Theorem 1.15 如果 μ 是 σ-代数 $\mathcal{A}$ 上的 σ-可加集函数，则 μ 有限可加且连续.

<u>Proof.</u> 设 $\{A_n\}_{n=1}^{\infty} \subset \mathcal{A}$ 是非降序列，即 $A = \bigcup\limits_{n=1}^{\infty} A_n \in \mathcal{A}$. 定义序列 $\{B_n\}_{n=1}^{\infty}$ 如下:

$$B_1 = A_1, \quad B_n = A_n \setminus A_{n-1}\ (n \geqslant 2)$$

由于 $\mathcal{A}$ 是 σ-代数，对差运算封闭，所以 $B_n \in \mathcal{A}$. 显然，B_n 两两不交，且:

$$A_n = \bigsqcup_{k=1}^{n} B_k, \quad A = \bigcup_{k=1}^{\infty} B_k$$

由 σ-可加性:

$$\mu(A) = \mu\left(\bigsqcup_{k=1}^{\infty} B_k\right) = \sum_{k=1}^{\infty} \mu(B_k)$$

另一方面，由有限可加性:

$$\mu(A_n) = \mu\left(\bigsqcup_{k=1}^{n} B_k\right) = \sum_{k=1}^{n} \mu(B_k)$$

因此:

$$\lim_{n\to\infty} \mu(A_n) = \lim_{n\to\infty} \sum_{k=1}^{n} \mu(B_k) = \sum_{k=1}^{\infty} \mu(B_k) = \mu(A)$$

这就证明了 μ 是下连续的.

设 $\{A_n\}_{n=1}^{\infty} \subset \mathcal{A}$ 是非增序列，即 $A_1 \supset A_2 \supset \cdots$，且 $A = \bigcap\limits_{n=1}^{\infty} A_n \in \mathcal{A}$，且存在 m 使得 $\mu(A_m) < \infty$. 由于序列是非增的，我们可以考虑补集序列 $\{A_m \setminus A_n\}_{n=m}^{\infty}$. 定义 $C_n = A_m \setminus A_n$，则 $\{C_n\}_{n=m}^{\infty}$ 是非降序列，且:

$$\bigcup_{n=m}^{\infty} C_n = A_m \setminus \bigcap_{n=m}^{\infty} A_n = A_m \setminus A$$

由下连续性（已证）：

$$\lim_{n\to\infty}\mu(C_n)=\mu\left(\bigcup_{n=m}^{\infty}C_n\right)=\mu(A_m\setminus A)$$

由于 $\mu(A_m)<\infty$，我们可以使用可减性：

$$\mu(C_n)=\mu(A_m\setminus A_n)=\mu(A_m)-\mu(A_n)\quad(\text{因为}A_n\subset A_m)$$

$$\mu(A_m\setminus A)=\mu(A_m)-\mu(A)\quad(\text{因为}A\subset A_m)$$

因此：

$$\lim_{n\to\infty}[\mu(A_m)-\mu(A_n)]=\mu(A_m)-\mu(A)$$

即：

$$\mu(A_m)-\lim_{n\to\infty}\mu(A_n)=\mu(A_m)-\mu(A)$$

所以：

$$\lim_{n\to\infty}\mu(A_n)=\mu(A)$$

这就证明了 μ 是上连续的. □

Definition 1.16 设 $\{A_t,t\in T\}$ 是 $\mathscr{A}$ 中的事件类，若对于 T 的任意子集列 $\{t_1,\dots,t_n\}$ 均有

$$P\left(\bigcap_{j=1}^{n}A_{t_j}\right)=\prod_{j=1}^{n}P\left(A_{t_j}\right)$$

则称 $\{A_t,t\in T\}$ 独立.

§1.3 随机变量及其分布函数

Definition 1.17 设 $(\Omega,\mathcal{F},\mathbb{P})$ 是概率空间，定义函数

$$\begin{aligned}X:\Omega&\to\mathbb{R}\\ \omega&\mapsto X(\omega)\end{aligned}$$

并且满足

$$\{\omega:X(\omega)\leqslant x\}\in\mathcal{F}$$

则称 X 是 $(\Omega,\mathcal{F},\mathbb{P})$ 上的实随机变量.

Definition 1.18 设 $(\Omega,\mathcal{F},\mathbb{P})$ 是概率空间. 定义集函数

$$\begin{aligned}P_X:\mathcal{B}&\to[0,1]\\ B&\mapsto P_X(B)\end{aligned}$$

其中

$$P_X(B)=\mathbb{P}\left(\{\omega:X(\omega)\in B\}\right)$$

则称 P_X 为随机变量 X 的概率分布. 由此，X 诱导出一个新的概率空间 $(\mathbb{R},\mathscr{B},P_X)$.

Definition 1.19 记

$$F(x) = P_X\{(-\infty, x]\} = \mathbb{P}(X \leqslant x)\,(x \in \mathbb{R})$$

称它是 X 的分布函数(简记为 d.f.) 易知它具有性质：

1. $F(x)$ 是不减、右连续的；
2. $\lim_{x\to-\infty} F(x) = 0, \lim_{x\to\infty} F(x) = 1$.

Theorem 1.20 设 $\{X_k\}_{k=1}^n$ 独立的充分必要条件是对于任意 $B_k \in \mathcal{B}^{(mk)}$ ，均有

$$P\left(\bigcap_{k=1}^n \{X^{(k)} \in B_k\}\right) = \prod_{k=1}^n P\left(X^{(k)} \in B_k\right).$$

§1.4 收敛方式

若不加说明，我们用大写字母 X, Y, Z 表示 $(\Omega, \mathcal{F}, \mathbb{P})$ 上的随机变量.

1.4.1 几乎必然收敛

Definition 1.21 设 $\{X_n\}_{n=1}^\infty$ 为一随机变量序列，X 亦是一随机变量，若存在 $N \in \mathcal{F}$ 满足 $\mathbb{P}(N) = 0$ ，使得

$$X_n(\omega) \to X(\omega) \quad \forall \omega \in \Omega \backslash N$$

则称 X_n 几乎必然收敛到 X .

Proposition 1.22 设 X_n 是随机变量序列，且 $X_n \to X \quad$ a.s. 则

1. 若 X_{n_k} 是 X_n 的子列，则 $X_{n_k} \to X \quad$ a.s.
2. 几乎必然收敛序列的极限函数是几乎必然唯一的
3. 若 $X_n = Y_n, X = Y \quad$ a.s. 则 $Y_n \to Y \quad$ a.s.

Theorem 1.23 若 $X_n^{(k)} \to X^{(k)} \quad$ a.s. $\quad \forall 1 \leqslant k \leqslant m$ ，F 是 $\mathbb{R}^m$ 上的连续函数，则

$$G\left(X_n^{(1)}, \ldots, X_n^{(m)}\right) \to G\left(X^{(1)}, \ldots, X^{(m)}\right) \quad \text{a.s.}$$

Proof. 由条件知存在零测集 $N_k \in \mathcal{F}$，使得对于所有 $\omega \in \Omega \setminus N_k$，有

$$\lim_{n\to\infty} X_n^{(k)}(\omega) = X^{(k)}(\omega).$$

令 $N = \bigcup_{k=1}^m N_k$，则 $\mathbb{P}(N) = 0$，且对于所有 $\omega \in \Omega \setminus N$ 和每个 k，均有 $X_n^{(k)}(\omega) \to X^{(k)}(\omega)$.

现在考虑函数 $G:\mathbb{R}^m \to \mathbb{R}$，已知 G 连续. 对于任意 $\omega \in \Omega \setminus N$，向量序列

$$\left(X_n^{(1)}(\omega),\dots,X_n^{(m)}(\omega)\right) \to \left(X^{(1)}(\omega),\dots,X^{(m)}(\omega)\right)$$

由 G 的连续性，有

$$G\left(X_n^{(1)}(\omega),\dots,X_n^{(m)}(\omega)\right) \to G\left(X^{(1)}(\omega),\dots,X^{(m)}(\omega)\right).$$

由于这对所有 $\omega \in \Omega \setminus N$ 成立，且 $\mathbb{P}(N)=0$，因此

$$G\left(X_n^{(1)},\dots,X_n^{(m)}\right) \to G\left(X^{(1)},\dots,X^{(m)}\right) \quad \text{几乎必然.}$$

□

Corollary 1.24 设 $X_n \to X, Y_n \to Y \quad \text{a.s.}$，则

1. $X_n + Y_n \longrightarrow X + Y \quad \text{a.s.}$
2. $X_n \cdot Y_n \longrightarrow X \cdot Y \quad \text{a.s.}$
3. $cX_n \longrightarrow cX \quad \text{a.s.} \quad c \in \mathbb{R}$

Theorem 1.25 随机变量列 X_n 收敛到某随机变量 X 的充要条件是:

$$X_m - X_n \to 0 \quad \text{a.s.} \quad (m,n\to\infty)$$

Proof. 必要性: 由于 $X_n \to X \quad \text{a.s.}$ 故存在零测集 $N \in \mathcal{F}$ 使得

$$X_n(\omega) \longrightarrow X(\omega) \quad \forall \omega \in \Omega \backslash N$$

利用 Cauchy 准则立刻获得当 $m,n\to\infty$ 时, $X_m(\omega)-X_n(\omega)\to 0, \forall\omega\in\Omega\backslash N$, 即 $X_m - X_n \to 0 \quad \text{a.s.}$

充分性: 由条件知存在零测集 $M \in \mathcal{F}$ 使得

$$X_m(\omega) - X_n(\omega) \to 0 \quad \forall \omega \in \Omega \backslash M$$

由 Cauchy 准则：在 X_n 收敛到某随机变量 $X'(\forall\omega\in\Omega\backslash M)$ 于是可以构造

$$X(\omega) = \begin{cases} X' & \omega \in M^c \\ 0 & \omega \in M \end{cases}$$

则 X 即为所求. □

Theorem 1.26 $X_n \to X \quad \text{a.s.} \iff \forall \varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\left(\bigcup_{m=n} \{|X_m - X| \geqslant \varepsilon\}\right) = 0$$

Proof. 几乎必然收敛的定义为：

$$X_n \to X \quad \text{a.s.} \Longleftrightarrow \mathbb{P}\left(\left\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1.$$

这等价于：对任意 $\varepsilon > 0$，有

$$\mathbb{P}\left(\limsup_{n\to\infty}\{|X_n - X| \geqslant \varepsilon\}\right) = 0,$$

设 $A_m(\varepsilon) = \{|X_m - X| \geqslant \varepsilon\}$，则

$$\limsup_{n\to\infty} A_n(\varepsilon) = \bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty} A_m(\varepsilon).$$

因此，

$$X_n \to X \quad \text{a.s.} \iff \forall\varepsilon > 0, \quad \mathbb{P}\left(\bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty} A_m(\varepsilon)\right) = 0.$$

令 $B_n = \bigcup_{m=n}^{\infty} A_m(\varepsilon)$，则 $\{B_n\}$ 是递减集合列. 根据概率的连续性，有：

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} B_n\right) = \lim_{n\to\infty} \mathbb{P}(B_n).$$

代入 B_n 的定义，得：

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty} A_m(\varepsilon)\right) = \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{m=n}^{\infty} A_m(\varepsilon)\right).$$

结合以上结果，有：

$$X_n \to X \quad \text{a.s.} \Longleftrightarrow \forall\varepsilon > 0, \quad \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{m=n}^{\infty}\{|X_m - X| \geqslant \varepsilon\}\right) = 0.$$

□

Theorem 1.27 (Borel-Cantelli) 设 $\{A_k\}_{k=1}^{\infty}$ 为一可测集列，则

$$\sum_{k=1}^{\infty} P(A_k) < \infty \implies \mathbb{P}\left(\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k\right) = 0$$

Proof. 由于

$$\sum_{k=1}^{\infty} P(A_k) < \infty \implies \sum_{k=n}^{\infty} \mathbb{P}(A_k) = 0$$

故

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k\right) \leqslant \mathbb{P}\left(\bigcup_{k=n}^{\infty} A_k\right) \leqslant \sum_{k=n}^{\infty} \mathbb{P}(A_k) = 0$$

□

1.4.2 依概率收敛

Definition 1.28 设 $\{X_n\}_{n=1}^{\infty}$ 是随机变量序列，X 是随机变量，若对于任意 $\varepsilon > 0$ 有

$$\lim_{n\to\infty} \mathbb{P}(\{|X_n - X| > \varepsilon\}) = 0$$

则称序列 X_n 依概率收敛于 X，记为 $X_n \xrightarrow{p} X$.

Proposition 1.29 设 $X_n \xrightarrow{p} X$ 则

1. 若 X_{n_k} 是 X_n 的子列，则 $X_{n_k} \xrightarrow{p} X$
2. 若另有 $X_n \xrightarrow{p} Y$ 则 $X = Y \quad \text{a.s.}$
3. 若 $X_n \xrightarrow{p} X, X_n = Y_n, X = Y \quad \text{a.s.}$ 则 $Y_n \xrightarrow{p} Y$.

Proof.

1. 显然
2. 注意到：
$$P(|X-Y| \geqslant \varepsilon) \leqslant P\left(|X_n - X| \geqslant \frac{\varepsilon}{2}\right) + P\left(|X_n - Y| \geqslant \frac{\varepsilon}{2}\right)$$
由于 $X_n \xrightarrow{p} X$ 和 $X_n \xrightarrow{p} Y$ 故
$$\lim_{n\to\infty} P(|X-Y| \geqslant \varepsilon) = P(|X-Y| \geqslant \varepsilon) = 0 \quad \forall \varepsilon > 0$$
取 $\varepsilon = \dfrac{1}{m}$，则：
$$P(|X-Y| > 0) = P\left(\bigcup_{m=1}^{\infty}\left(|X-Y| \geqslant \frac{1}{m}\right)\right) \leqslant \sum_{m=1}^{\infty} P\left(|X-Y| \geqslant \frac{1}{m}\right) = 0$$
3. 由题设，存在零测集 A_n 和 B，使得：
$$Y_n = X_n, \forall \omega \in A_n^c \quad X = Y, \forall \omega \in B^c$$
令 $M = B \cup A, A = \bigcup_{n=1}^{\infty} A_n$ 则 $P(M) = 0$. 于是
$$\begin{aligned}\{|Y_n - Y| > \varepsilon\} &= \{|Y_n - Y| > \varepsilon\} \bigcap M^c + \{|Y_n - Y| > \varepsilon\} \bigcap M \\ &\subset \{|X_n - X| > \varepsilon\} \bigcap M^c + M \\ &\subset \{|X_n - X| > \varepsilon\} \bigcup M.\end{aligned}$$
带上概率符号 P 立刻得到结果.

$\square$

Theorem 1.30 本定理描述了概率收敛和几乎必然收敛的关系

1. $X_n \to X \quad \text{a.s.} \implies X_n \xrightarrow{p} X$
2. $X_n \xrightarrow{p} X \implies \exists \left\{X_{n_k}\right\}_{k=1}^{\infty} \subseteq \{X_n\}_{n=1}^{\infty}$ 使得 $X_{n_k} \to X \quad \text{a.s.}$
3. $X_n \xrightarrow{p} X \iff \forall \left\{X_{n_k}\right\}_{k=1}^{\infty} \subseteq \{X_n\}_{n=1}^{\infty}, \exists \left\{X_{n_{k_j}}\right\}_{j=1}^{\infty} \subseteq \left\{X_{n_k}\right\}_{k=1}^{\infty}$ 使得 $X_{n_{k_j}} \to X \quad \text{a.s.}$

Proof.

1. 易得
2. 由于 $X_n \xrightarrow{p} X$，即对任意 $\varepsilon > 0$，有

$$\lim_{n\to\infty} P\left(|X_n - X| > \varepsilon\right) = 0.$$

取 $\varepsilon_k = \dfrac{1}{k}(k \in \mathbb{N})$，则对每个 k，存在 n_k 使得

$$P\left(\left|X_{n_k} - X\right| > \frac{1}{k}\right) < \frac{1}{2^k},$$

且可选取 $n_1 < n_2 < \cdots < n_k < \cdots$，即 $\left\{X_{n_k}\right\}$ 为递增子序列. 定义

$$A_k = \left\{\left|X_{n_k} - X\right| > \frac{1}{k}\right\}$$

由构造可知：

$$\sum_{k=1}^{\infty} P\left(A_k\right) < \sum_{k=1}^{\infty} \frac{1}{2^k} = 1 < \infty \implies P\left(\limsup_{k\to\infty} A_k\right) = 0$$

即就是

$$X_{n_k} \xrightarrow{\text{a.s.}} X$$

3. 必要性显然，下证充分性：假设不成立，则存在 $\varepsilon > 0, \delta > 0$ 及无穷多个 n 使得

$$P\left(|X_n - X| > \varepsilon\right) \geqslant \delta$$

设这些 n 为 $n_1 < n_2 < \cdots$，则 $\left\{X_{n_k}\right\}$ 为子序列且满足

$$P\left(\left|X_{n_k} - X\right| > \varepsilon\right) \geqslant \delta \quad \forall k \geqslant 1$$

由假设，存在子序列 $\left\{X_{n_{k_j}}\right\}$ 使得 $X_{n_{k_j}} \xrightarrow{\text{a.s.}} X$. 但几乎必然收敛蕴含依概率收敛，故对充分大的 j，有

$$P\left(\left|X_{n_{k_j}} - X\right| > \varepsilon\right) < \delta,$$

矛盾. 因此假设错误，必有 $X_n \xrightarrow{p} X$.

□

Theorem 1.31 若 $X_n^{(k)} \xrightarrow{p} X^{(k)} \quad \forall 1 \leqslant k \leqslant m$ ，G 是 $\mathbb{R}^m$ 上的一致连续函数，则

$$G\left(X_n^{(1)}, \ldots, X_n^{(m)}\right) \xrightarrow{p} G\left(X^{(1)}, \ldots, X^{(m)}\right)$$

不妨记

$$Y_n = (X_n^{(1)}, \ldots, X_n^{(m)}), \quad Y = (X^{(1)}, \ldots, X^{(m)})$$

故结论等价于

$$G(Y_n) \xrightarrow{p} G(Y).$$

Proof. 对任意 $\delta > 0$，有：

$$P\left(\|Y_n - Y\|_\infty \geqslant \delta\right) = P\left(\max_{1\leqslant k\leqslant m} |X_n^{(k)} - X^{(k)}| \geqslant \delta\right).$$

注意到：

$$\left\{\max_{1\leqslant k\leqslant m} |X_n^{(k)} - X^{(k)}| \geqslant \delta\right\} = \bigcup_{k=1}^{m} \left\{|X_n^{(k)} - X^{(k)}| \geqslant \delta\right\},$$

由概率的次可加性：

$$P(\|Y_n - Y\|_\infty \geqslant \delta) \leqslant \sum_{k=1}^{m} P\left(|X_n^{(k)} - X^{(k)}| \geqslant \delta\right).$$

由于每个 $X_n^{(k)} \xrightarrow{p} X^{(k)}$，右边每一项都趋于 0，故：

$$\lim_{n\to\infty} P(\|Y_n - Y\|_\infty \geqslant \delta) = 0,$$

即 $Y_n \xrightarrow{p} Y$.

设 $\varepsilon > 0$. 由于 G 在 D 上一致连续，存在 $\delta > 0$，使得对任意 $y, y' \in D$ 满足 $\|y - y'\|_\infty < \delta$，有：

$$|G(y) - G(y')| < \varepsilon.$$

考虑事件：

$$A_n = \{|G(Y_n) - G(Y)| > \varepsilon\}.$$

由于 $Y_n, Y \in D$ 几乎必然，若 $\|Y_n - Y\|_\infty < \delta$，则 $|G(Y_n) - G(Y)| < \epsilon$. 因此：

$$A_n \subseteq \{\|Y_n - Y\|_\infty \geqslant \delta\}.$$

故：

$$P(A_n) \leqslant P(\|Y_n - Y\|_\infty \geqslant \delta) \to 0 \quad (n \to \infty),$$

即 $G(Y_n) \xrightarrow{p} G(Y)$. □

Remark 1.32 我们实际证明了对于子集 $D \in \mathbb{R}^m$ 的情形，此时只需要保证各随机变量都以概率 1 落在 D 中即可.

Theorem 1.33 若 $X_n^{(k)} \xrightarrow{p} X^{(k)} \quad \forall 1 \leqslant k \leqslant m$，$G$ 是 $\mathbb{R}^m$ 上的连续函数，则

$$G\left(X_n^{(1)}, \ldots, X_n^{(m)}\right) \xrightarrow{p} G\left(X^{(1)}, \ldots, X^{(m)}\right)$$

不妨记

$$Y_n = (X_n^{(1)}, \ldots, X_n^{(m)}), \quad Y = (X^{(1)}, \ldots, X^{(m)})$$

故结论等价于

$$G(Y_n) \xrightarrow{p} G(Y).$$

Proof. 要证明 $G(Y_n) \xrightarrow{p} G(Y)$，只需证明 $\forall \{Y_{n_k}\} \subset \{Y_n\}, \exists \{Y_{n_{k_j}}\} \subset \{Y_{n_k}\}$ 使得

$$G(Y_{n_{k_j}}) \xrightarrow{\text{a.s.}} G(Y).$$

设 $\{Y_{n_k}\}$ 是任意子列. 由于 $Y_n \xrightarrow{p} Y$，根据依概率收敛的性质，存在子子列 $\{Y_{n_{k_j}}\}$ 使得

$$Y_{n_{k_j}} \xrightarrow{\text{a.s.}} Y.$$

由于 $Y_{n_{k_j}} \xrightarrow{\text{a.s.}} Y$ 且 G 是连续函数，由连续映射定理（几乎处处收敛版本）可得:

$$G(Y_{n_{k_j}}) \xrightarrow{\text{a.s.}} G(Y).$$

故定理得证. □

1.4.3 依分布收敛

我们记函数 F 的连续点全体为 $C(F)$

Definition 1.34 设 X_n 的分布函数为 F_n，X 的分布函数为 F，若有

$$F_n(x) \to F(x) \quad \forall x \in C(F)$$

则称 X_n 依分布收敛到 X，记为 $X_n \xrightarrow{d} X$.

Theorem 1.35 $X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$

Proof. 只需证: 对所有的 x, 有

$$F(x-0) \leqslant \varliminf_{n\to\infty} F_n(x) \leqslant \varlimsup_{n\to\infty} F_n(x) \leqslant F(x+0). \tag{1.1}$$

令 $x' < x$, 则

$$\begin{aligned}\{X \leqslant x'\} &= \{X_n \leqslant x, X \leqslant x'\} \cup \{X_n > x, X \leqslant x'\} \\ &\subset \{X_n \leqslant x\} \cup \{|X_n - X| \geqslant x - x'\}\end{aligned}$$

$\implies$

$$F(x') \leqslant F_n(x) + P(|X_n - X| \geqslant x - x').$$

$X_n \xrightarrow{P} X$, 蕴含 $P(|X_n - X| \geqslant x - x') \to 0 (n \to \infty)$. 所以有

$$F(x') \leqslant \lim_{n\to\infty} F_n(x).$$

令 $x' \to x$, 即得

$$F(x-0) \leqslant \varliminf_{n\to\infty} F_n(x).$$

同理, 当 $x'' > x$ 时, 有

$$\varlimsup_{n\to\infty} F_n(x) \leqslant F(x'').$$

令 $x'' \to x$, 即得

$$\varlimsup_{n\to\infty} F_n(x) \leqslant F(x+0).$$

这就证明了定理. □

Corollary 1.36 当 $X = c$ 为一常数时,

$$X_n \xrightarrow{p} c \iff X_n \xrightarrow{d} c$$

Proof. 仍然记 X_n 的分布函数为 $F_n(x)$, 常数 c 的分布函数为:

$$F_c(x) = \begin{cases} 0 & x < c, \\ 1 & x \geqslant c. \end{cases}$$

需要证明对任意 $\varepsilon > 0$ ，有

$$\lim_{n\to\infty} P(|X_n - c| > \varepsilon) = 0$$

注意:

$$P(|X_n - c| > \varepsilon) = P(X_n < c - \varepsilon) + P(X_n > c + \varepsilon).$$

由于 $c - \varepsilon, c + \varepsilon \in C(F_c)$，于是

$$F_n(c-\epsilon) = P(X_n \leqslant c - \epsilon) \to F_c(c-\epsilon) = 0$$
$$F_n(c+\epsilon) = P(X_n \leqslant c + \epsilon) \to F_c(c+\epsilon) = 1$$

因此,

$$P(|X_n - c| > \varepsilon) = P(X_n < c - \varepsilon) + P(X_n > c + \varepsilon) \to 0$$

即 $X_n \xrightarrow{p} c$. □

Theorem 1.37 (Slutsky 引理) 设 $X_n \xrightarrow{d} X, Y_n \xrightarrow{p} c$ 则

1. $X_n + Y_n \xrightarrow{d} X + c$
2. $X_n Y_n \xrightarrow{d} X \cdot c$
3. 若 $c \neq 0$，则 $\dfrac{X_n}{Y_n} \xrightarrow{d} \dfrac{X}{c}$

Lemma 1.38 如果 $X_n \xrightarrow{d} X$ 且 $Y_n \xrightarrow{p} c$，那么随机向量序列 (X_n, Y_n) 依分布收敛于 (X, c)，即：

$$(X_n, Y_n) \xrightarrow{d} (X, c)$$

Slutsky 引理的证明. 分别定义函数

$$f_1(x,y) = x + y, \quad f_2(x,y) = x \cdot y, \quad f_3(x,y) = \frac{x}{y}$$

利用连续映射定理和上述引理直接完成证明. □

Lemma 1.39 设 $X_n - X_n' \xrightarrow{p} 0$ 且 $X_n' \xrightarrow{d} X$ ，则 $X_n \xrightarrow{d} X$.

Proof. 设 X 的分布函数为 $F(x) = P(X \leqslant x)$，需要证明：

$$\forall \varepsilon > 0, \lim_{n\to\infty} P(X_n \leqslant x) = F(x) \quad \forall x \in C(F)$$

其中 $C(F)$ 是 F 的连续点集. 我们有：

$$\begin{aligned}
P(X_n \leqslant x) &= P(X_n \leqslant x, |X_n' - X_n| \leqslant \varepsilon) + P(X_n \leqslant x, |X_n' - X_n| > \varepsilon) \\
&\leqslant P(X_n' \leqslant x + X_n' - X_n, |X_n' - X_n| \leqslant \varepsilon) + P(|X_n' - X_n| > \varepsilon) \\
&\leqslant P(X_n' \leqslant x + \varepsilon) + P(|X_n' - X_n| > \varepsilon) \\
&= P(X_n' \leqslant x + \varepsilon) + o(1)
\end{aligned}$$

同时：

$$\begin{aligned}
P(X_n' \leqslant x - \varepsilon) &\leqslant P(X_n' \leqslant x - \varepsilon, |X_n' - X_n| \leqslant \varepsilon) + P(|X_n' - X_n| \geqslant \varepsilon) \\
&\leqslant P(X_n \leqslant x) + P(|X_n' - X_n| \geqslant \varepsilon) \\
&= P(X_n \leqslant x) + o(1)
\end{aligned}$$

因此得到不等式：

$$P(X_n' \leqslant x - \varepsilon) + o(1) \leqslant P(X_n \leqslant x) \leqslant P(X_n' \leqslant x + \varepsilon) + o(1)$$

由于 $X_n' \xrightarrow{d} X$，且 F 的间断点至多可数，可以选择 $\varepsilon_k \downarrow 0$ 使得 $x \pm \varepsilon_k \in C(F)$，则：

$$\lim_{n\to\infty} P(X_n' \leqslant x \pm \varepsilon_k) = F(x \pm \varepsilon_k)$$

取极限得：

$$F(x - \varepsilon_k) \leqslant \liminf_{n\to\infty} P(X_n \leqslant x) \leqslant \limsup_{n\to\infty} P(X_n \leqslant x) \leqslant F(x + \varepsilon_k)$$

令 $\varepsilon_k \to 0$，由 $x \in C(F)$ 得 $F(x \pm \varepsilon_k) \to F(x)$，故：

$$\lim_{n\to\infty} P(X_n \leqslant x) = F(x)$$

因此 $X_n \xrightarrow{d} X$. □

基于引理1.39对 Slutsky 第一条的证明. 注意到：

$$(X_n + Y_n) - (X_n + c) \xrightarrow{d} 0 \iff (X_n + Y_n) - (X_n + c) \xrightarrow{p} 0$$

又由于 $X_n + c \xrightarrow{d} X + c$ 故 $X_n + Y_n \xrightarrow{d} X + c$ □

§1.5 例子

Example 1.40 若 $X_1,\cdots,X_n \sim U[0,1]$，$\hat{\theta}$ 是其极大似然估计，证明

1. $\hat{\theta} \xrightarrow{\text{a.s.}} \theta$
2. $n\left(\theta-\hat{\theta}\right) \xrightarrow{d} \text{Exp}(\frac{1}{\theta})$

Example 1.41 设 $X_1,\ldots,X_n \sim \text{Exp}(\lambda,a)$ 即

$$f_{\lambda,a}(x)=\lambda\exp\left(-\lambda(x-a)\right)\mathbb{I}_{\{x\geqslant a\}}$$

设 $(\hat{\lambda}_n,\hat{a}_n)$ 为其极大似然估计，证明

$$(\hat{\lambda}_n,\hat{a}_n)\xrightarrow{P}(\lambda,a)$$

Solution. 计算其对数似然函数为：

$$\ell(\lambda,a)=n\ln\lambda-\lambda\sum_{i=1}^{n}(x_i-a)+\ln\mathbf{1}\{\min\{x_1,\ldots,x_n\}\geqslant a\}$$

对数似然函数关于 a 单调递增，因此最大似然估计量为：

$$\hat{a}_n=\min\{X_1,\ldots,X_n\}$$

代入并求导

$$\frac{\partial\ell}{\partial\lambda}=\frac{n}{\lambda}-\sum_{i=1}^{n}(X_i-\hat{a}_n)=0 \implies \hat{\lambda}_n=\frac{n}{\sum_{i=1}^{n}(X_i-\hat{a}_n)}$$

由于 $\hat{a}_n=\min\{X_1,\ldots,X_n\}$，且每个 $X_i\geqslant a$，有：

$$\mathbb{P}(\hat{a}_n>a+\varepsilon)=[\mathbb{P}(X_1>a+\varepsilon)]^n=\left(e^{-\lambda\varepsilon}\right)^n=e^{-n\lambda\varepsilon}\longrightarrow 0$$

因此 $\hat{a}_n\xrightarrow{\mathbf{P}}a$.

令：

$$Y_n=\frac{1}{n}\sum_{i=1}^{n}(X_i-a),\quad Z_n=\hat{a}_n-a$$

则：

$$\hat{\lambda}_n=\frac{1}{Y_n-Z_n}$$

由大数定律：

$$Y_n\xrightarrow{\mathbf{P}}\mathbb{E}[X_1-a]=\frac{1}{\lambda}$$

由 $Z_n\xrightarrow{\mathbf{P}}0$，因此：

$$Y_n-Z_n\xrightarrow{\mathbf{P}}\frac{1}{\lambda}$$

由连续映射定理：

$$\hat{\lambda}_n=\frac{1}{Y_n-Z_n}\xrightarrow{\mathbf{P}}\lambda$$

□

Chapter 2

积分理论

§2.1 o_p, O_p 记号

Definition 2.1 对于两个确定性序列 $\{a_n\}_{n=1}^{\infty}$ 和 $\{b_n\}_{n=1}^{\infty}$：

- 若存在常数 $c > 0$ 使得 $|\frac{a_n}{b_n}| \leqslant c$ 对所有 n 成立，则称 $\frac{a_n}{b_n}$ 有界，记作 $a_n = O(b_n)$；
- 若 $\{b_n\}_{n=1}^{\infty}$ 非负且 $\frac{a_n}{b_n} \to 0$，则记作 $a_n = o(b_n)$
- 若 $a_n = O(b_n)$ 且 $b_n = O(a_n)$，则称它们同阶，记作 $a_n \asymp b_n$；
- 若 $\frac{a_n}{b_n} \to 1$，则称它们渐近等价，记作 $a_n \sim b_n$

Definition 2.2 (矩阵范数) 设 $\mathbf{A} = (a_{ij})_{p\times q}$，常见矩阵范数定义如下：

1. 谱范数：
$$\|\mathbf{A}\|_2 = \sup_{\mathbf{u}\in\mathbb{R}^q\setminus\{\mathbf{0}\}} \frac{\|\mathbf{A}\mathbf{u}\|_2}{\|\mathbf{u}\|_2} = \sqrt{\lambda_{\max}(\mathbf{A}^{\mathrm{T}}\mathbf{A})},$$
其中 $\lambda_{\max}(\cdot)$ 表示最大特征值.

2. l_1 范数（最大列和范数）：
$$\|\mathbf{A}\|_1 = \sup_{\mathbf{u}\in\mathbb{R}^q\setminus\{\mathbf{0}\}} \frac{\|\mathbf{A}\mathbf{u}\|_1}{\|\mathbf{u}\|_1} = \max_{1\leqslant j\leqslant q} \sum_{i=1}^{p} |a_{ij}|.$$

3. l_∞ 范数（最大行和范数）：
$$\|\mathbf{A}\|_\infty = \sup_{\mathbf{u}\in\mathbb{R}^q\setminus\{\mathbf{0}\}} \frac{\|\mathbf{A}\mathbf{u}\|_\infty}{\|\mathbf{u}\|_\infty} = \max_{1\leqslant i\leqslant p} \sum_{j=1}^{q} |a_{ij}|.$$

Theorem 2.3 对向量 $\mathbf{v} \in \mathbb{R}^q$、矩阵 $\mathbf{A} \in \mathbb{R}^{p\times q}$ 和 $\mathbf{B} \in \mathbb{R}^{q\times k}$，其范数满足以下性质（$\alpha \in \{2, 1, \infty\}$）：

1. $\|\mathbf{Av}\|_\alpha \leqslant \|\mathbf{A}\|_\alpha \|\mathbf{v}\|_\alpha$；
2. $\|\mathbf{AB}\|_\alpha \leqslant \|\mathbf{A}\|_\alpha \|\mathbf{B}\|_\alpha$；
3. $\|\mathbf{A}+\mathbf{B}\|_\alpha \leqslant \|\mathbf{A}\|_\alpha + \|\mathbf{B}\|_\alpha$

Proof. 我们分别对 $\alpha = 2, 1, \infty$ 三种范数证明这三个性质. 以下记

$$\boldsymbol{A} = (a_{ij})_{p\times q} \quad \boldsymbol{B} = (b_{ij})_{q\times k} \quad \boldsymbol{v} = (v_1, \ldots, v_q)^\top$$

.

性质 1: $\|\boldsymbol{Av}\|_\alpha \leqslant \|\boldsymbol{A}\|_\alpha \|\boldsymbol{v}\|_\alpha$

1. 当 $\alpha = 2$ 时，由矩阵 2-范数的定义：

$$\|\boldsymbol{A}\|_2 = \sup_{\boldsymbol{x}\neq \boldsymbol{0}} \frac{\|\boldsymbol{Ax}\|_2}{\|\boldsymbol{x}\|_2},$$

   因此对任意 $\boldsymbol{v} \neq \boldsymbol{0}$，有 $\|\boldsymbol{Av}\|_2 \leqslant \|\boldsymbol{A}\|_2 \|\boldsymbol{v}\|_2$；当 $\boldsymbol{v} = \boldsymbol{0}$ 时等式自然成立.

2. 当 $\alpha = 1$ 时：

$$\begin{aligned}
\|\boldsymbol{Av}\|_1 &= \sum_{i=1}^{p} \left| \sum_{j=1}^{q} a_{ij} v_j \right| \leqslant \sum_{i=1}^{p} \sum_{j=1}^{q} |a_{ij}| |v_j| \\
&= \sum_{j=1}^{q} |v_j| \left(\sum_{i=1}^{p} |a_{ij}| \right) \leqslant \left(\max_{1\leqslant j\leqslant q} \sum_{i=1}^{p} |a_{ij}| \right) \sum_{j=1}^{q} |v_j| \\
&= \|\boldsymbol{A}\|_1 \|\boldsymbol{v}\|_1.
\end{aligned}$$

3. 当 $\alpha = \infty$ 时：

$$\begin{aligned}
\|\boldsymbol{Av}\|_\infty &= \max_{1\leqslant i\leqslant p} \left| \sum_{j=1}^{q} a_{ij} v_j \right| \leqslant \max_{1\leqslant i\leqslant p} \sum_{j=1}^{q} |a_{ij}| |v_j| \\
&\leqslant \left(\max_{1\leqslant i\leqslant p} \sum_{j=1}^{q} |a_{ij}| \right) \left(\max_{1\leqslant j\leqslant q} |v_j| \right) = \|\boldsymbol{A}\|_\infty \|\boldsymbol{v}\|_\infty.
\end{aligned}$$

性质 2: $\|\boldsymbol{AB}\|_\alpha \leqslant \|\boldsymbol{A}\|_\alpha \|\boldsymbol{B}\|_\alpha$

1. 当 $\alpha = 2$ 时，对任意非零向量 $\boldsymbol{x}$：

$$\|\boldsymbol{ABx}\|_2 \leqslant \|\boldsymbol{A}\|_2 \|\boldsymbol{Bx}\|_2 \leqslant \|\boldsymbol{A}\|_2 \|\boldsymbol{B}\|_2 \|\boldsymbol{x}\|_2,$$

   因此 $\|\boldsymbol{AB}\|_2 \leqslant \|\boldsymbol{A}\|_2 \|\boldsymbol{B}\|_2$。

2. 当 $\alpha = 1$ 时：

$$
\begin{aligned}
\|\boldsymbol{AB}\|_1 &= \max_{1\leqslant j\leqslant k}\sum_{i=1}^{p}\left|\sum_{l=1}^{q} a_{il}b_{lj}\right| \leqslant \max_{1\leqslant j\leqslant k}\sum_{i=1}^{p}\sum_{l=1}^{q}|a_{il}||b_{lj}| \\
&= \max_{1\leqslant j\leqslant k}\sum_{l=1}^{q}|b_{lj}|\left(\sum_{i=1}^{p}|a_{il}|\right) \leqslant \left(\max_{1\leqslant l\leqslant q}\sum_{i=1}^{p}|a_{il}|\right)\left(\max_{1\leqslant j\leqslant k}\sum_{l=1}^{q}|b_{lj}|\right) \\
&= \|\boldsymbol{A}\|_1\|\boldsymbol{B}\|_1.
\end{aligned}
$$

3. 当 $\alpha = \infty$ 时：

$$
\begin{aligned}
\|\boldsymbol{AB}\|_\infty &= \max_{1\leqslant i\leqslant p}\sum_{j=1}^{k}\left|\sum_{l=1}^{q} a_{il}b_{lj}\right| \leqslant \max_{1\leqslant i\leqslant p}\sum_{j=1}^{k}\sum_{l=1}^{q}|a_{il}||b_{lj}| \\
&= \max_{1\leqslant i\leqslant p}\sum_{l=1}^{q}|a_{il}|\left(\sum_{j=1}^{k}|b_{lj}|\right) \leqslant \left(\max_{1\leqslant i\leqslant p}\sum_{l=1}^{q}|a_{il}|\right)\left(\max_{1\leqslant l\leqslant q}\sum_{j=1}^{k}|b_{lj}|\right) \\
&= \|\boldsymbol{A}\|_\infty\|\boldsymbol{B}\|_\infty.
\end{aligned}
$$

性质 3: $\|\boldsymbol{A}+\boldsymbol{B}\|_\alpha \leqslant \|\boldsymbol{A}\|_\alpha + \|\boldsymbol{B}\|_\alpha$

1. 当 $\alpha = 2$ 时，对任意非零向量 $\boldsymbol{x}$：

$$
\|(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{x}\|_2 = \|\boldsymbol{Ax}+\boldsymbol{Bx}\|_2 \leqslant \|\boldsymbol{Ax}\|_2 + \|\boldsymbol{Bx}\|_2 \leqslant (\|\boldsymbol{A}\|_2 + \|\boldsymbol{B}\|_2)\|\boldsymbol{x}\|_2,
$$

因此 $\|\boldsymbol{A}+\boldsymbol{B}\|_2 \leqslant \|\boldsymbol{A}\|_2 + \|\boldsymbol{B}\|_2$。

2. 当 $\alpha = 1$ 时：

$$
\begin{aligned}
\|\boldsymbol{A}+\boldsymbol{B}\|_1 &= \max_{1\leqslant j\leqslant q}\sum_{i=1}^{p}|a_{ij}+b_{ij}| \leqslant \max_{1\leqslant j\leqslant q}\sum_{i=1}^{p}(|a_{ij}|+|b_{ij}|) \\
&\leqslant \max_{1\leqslant j\leqslant q}\sum_{i=1}^{p}|a_{ij}| + \max_{1\leqslant j\leqslant q}\sum_{i=1}^{p}|b_{ij}| \\
&= \|\boldsymbol{A}\|_1 + \|\boldsymbol{B}\|_1.
\end{aligned}
$$

3. 当 $\alpha = \infty$ 时：

$$
\begin{aligned}
\|\boldsymbol{A}+\boldsymbol{B}\|_\infty &= \max_{1\leqslant i\leqslant p}\sum_{j=1}^{q}|a_{ij}+b_{ij}| \leqslant \max_{1\leqslant i\leqslant p}\sum_{j=1}^{q}(|a_{ij}|+|b_{ij}|) \\
&\leqslant \max_{1\leqslant i\leqslant p}\sum_{j=1}^{q}|a_{ij}| + \max_{1\leqslant i\leqslant p}\sum_{j=1}^{q}|b_{ij}| \\
&= \|\boldsymbol{A}\|_\infty + \|\boldsymbol{B}\|_\infty.
\end{aligned}
$$

□

Theorem 2.4 (谱范数的紧性) 对于对称矩阵 $\boldsymbol{A}=(a_{ij})_{p\times p}$，其谱范数是最" 紧" 的，即

$$\|\boldsymbol{A}\|_2 \leqslant \|\boldsymbol{A}\|_\alpha \quad \alpha\in\{1,\infty\}.$$

Proof. 首先注意到

$$\max_{1\leqslant i\leqslant p}|\lambda_i(\boldsymbol{A})| = \max_{1\leqslant i\leqslant p}\left|\lambda_i^{1/2}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})\right| = \|\boldsymbol{A}\|_2.$$

对任意特征值 $\lambda_i(\boldsymbol{A})$ 及其对应的特征向量 $\boldsymbol{v}_i$，有

$$|\lambda_i(\boldsymbol{A})|\cdot\|\boldsymbol{v}_i\|_\alpha = \|\lambda_i(\boldsymbol{A})\boldsymbol{v}_i\|_\alpha = \|\boldsymbol{A}\boldsymbol{v}_i\|_\alpha \leqslant \|\boldsymbol{A}\|_\alpha\|\boldsymbol{v}_i\|_\alpha.$$

因此，

$$\|\boldsymbol{A}\|_2 = \max_{1\leqslant i\leqslant p}|\lambda_i(\boldsymbol{A})| \leqslant \|\boldsymbol{A}\|_\alpha.$$

□

2.1.1

Definition 2.5 称随机变量序列 $\{X_n\}_{n\geqslant 1}$ 依概率有界，如果对于任意给定的 $\varepsilon>0$，存在 $M>0$ 使得

$$\sup_n P(|X_n|>M)<\varepsilon$$

此时记 $X_n=O_p(1)$

Definition 2.6 称 $\{\boldsymbol{X}_\alpha,\alpha\in A\}$ 一致紧，如果对于任意给定的 $\varepsilon>0$，存在 $M>0$，使得

$$\sup_{\alpha\in A}P(\|\boldsymbol{X}_\alpha\|>M)<\varepsilon$$

Theorem 2.7 (Portmanteau) 对于任意的随机向量 X_n 和 X，下列表述等价：

1. $X_n \xrightarrow{d} X$;
2. 对一切有界实值连续函数 f，均有 $\mathbb{E}\left\{f(X_n)\right\}\to\mathbb{E}\{f(X)\}$;
3. 对一切有界实值一致连续函数 f，均有 $\mathbb{E}\left\{f(X_n)\right\}\to\mathbb{E}\{f(X)\}$;
4. 对一切闭集 C，$\varlimsup_{n\to\infty}P(X_n\in C)\leqslant P(X\in C)$,
5. 对一切开集 G，$\varliminf_{n\to\infty}P(X_n\in G)\geqslant P(X\in G)$;
6. 若 A 是 **Borel** 集，$P(\partial A)=0$，则 $\lim_{n\to\infty}P(X_n\in A)=P(X\in A)$.

Theorem 2.8 (Prohorov) 设 $\{X_n\}_{n=1}^\infty$ 是随机向量序列，且依分布收敛到一随机向量 X，则

$$\|X_n\|=O_p(1)$$

Proof. 证明 $\|X_n\| = O_p(1)$，即对任意 $\varepsilon > 0$，存在 $M > 0$，使得

$$\sup_n \mathbb{P}(\|X_n\| > M) < \varepsilon.$$

固定 $\varepsilon > 0$. 由于 X 是随机向量，存在 $M_0 > 0$，使得

$$\mathbb{P}(\|X\| \geqslant M_0) < \frac{\varepsilon}{2}.$$

令闭集 $F = \{x : \|x\| \geqslant M_0\}$，则

$$\mathbb{P}(X \in F) = \mathbb{P}(\|X\| \geqslant M_0) < \frac{\varepsilon}{2}.$$

由于 $X_n \xrightarrow{d} X$，根据 Portmanteau 定理

$$\limsup_{n\to\infty} \mathbb{P}(\|X_n\| \geqslant M_0) = \limsup_{n\to\infty} \mathbb{P}(X_n \in F) \leqslant \mathbb{P}(X \in F) < \frac{\varepsilon}{2}.$$

因此，存在正整数 N，使得对任意 $n \geqslant N$，

$$\mathbb{P}(\|X_n\| \geqslant M_0) < \varepsilon.$$

对 $n = 1, 2, \ldots, N-1$，由于每个 X_n 是随机向量，存在 $M_n > 0$，使得

$$\mathbb{P}(\|X_n\| > M_n) < \varepsilon.$$

令

$$M = \max\{M_0, M_1, \ldots, M_{N-1}\}.$$

则

$$\sup_n \mathbb{P}(\|X_n\| > M) < \varepsilon.$$

□

Theorem 2.9 设 $f : \mathbb{R}^m \to \mathbb{R}^1$ 在 **Borel** 集 C 内连续且 $P(X \in C) = 1$，那么

1. $X_n \xrightarrow{p} X \implies f(X_n) \xrightarrow{p} f(X)$
2. $X_n \xrightarrow{\text{a.s.}} X \implies f(X_n) \xrightarrow{\text{a.s.}} f(X)$
3. $X_n \xrightarrow{d} X \implies f(X_n) \xrightarrow{d} f(X)$

仅证 $X_n \xrightarrow{d} X$. 设 $F \subset \mathbb{R}$ 为任意闭集. 注意到:

$$f^{-1}(F) \subset \overline{f^{-1}(F)} \subset f^{-1}(F) \cup C^c$$

其中第二个包含关系证明如下：任取 $x \in \overline{f^{-1}(F)}$，存在序列 $\{x_m\} \subset f^{-1}(F)$ 使得 $x_m \to x$. 分两种情况:

1. 若存在子列 $\{x_{m_k}\} \subset C$，则由 f 在 C 上连续性和 F 的闭性得 $f(x) \in F$，即 $x \in f^{-1}(F)$

2. 若存在 N 使当 $m \geqslant N$ 时 $x_m \in C^c$，则 $x \in C^c$

故 $x \in f^{-1}(F) \cup C^c$. 由 Portmanteau 定理可得：

$$\begin{aligned}\limsup_{n\to\infty} P(f(X_n) \in F) &= \limsup_{n\to\infty} P(X_n \in f^{-1}(F)) \\ &\leqslant \limsup_{n\to\infty} P(X_n \in \overline{f^{-1}(F)}) \\ &\leqslant P(X \in \overline{f^{-1}(F)})\end{aligned}$$

由 $P(X \in C^c) = 0$ 得：

$$\begin{aligned}P(X \in \overline{f^{-1}(F)}) &\leqslant P(X \in f^{-1}(F) \cup C^c) \\ &\leqslant P(X \in f^{-1}(F)) + P(X \in C^c) \\ &= P(X \in f^{-1}(F)) = P(f(X) \in F)\end{aligned}$$

综合上述不等式：

$$\limsup_{n\to\infty} P(f(X_n) \in F) \leqslant P(f(X) \in F)$$

根据 Portmanteau 定理中 (1) $\Leftrightarrow$ (2) 的等价性，这正好表明：

$$f(X_n) \xrightarrow{d} f(X)$$

□

Theorem 2.10 我们有

1. 如果 $X_n \xrightarrow{d} X$ 且 $d(X_n, Y_n) \xrightarrow{p} 0$，则 $Y_n \xrightarrow{d} X$；
2. 如果 $X_n \xrightarrow{d} X$ 且 $Y_n \xrightarrow{p} \mathbf{a}$，则 $(X_n, Y_n^{\mathrm{T}})^{\mathrm{T}} \xrightarrow{d} (X, \mathbf{a}^{\mathrm{T}})^{\mathrm{T}}$

§2.2 例——线性模型的收敛速率

Example 2.11 考虑线性回归模型：

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

其中 $\mathbf{X}$ 是 $n \times p$ 维列满秩随机协变量矩阵，$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ 是误差项. 基本假设：

1. $\mathbb{E}(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}_{n\times 1}$
2. $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T|\mathbf{X}) = \sigma^2 \mathbf{I}_n$，其中 $\sigma^2 > 0$ 是常数
3. 对于充分大的 n，几乎必然成立：

$$\infty > c_1 \geqslant \lambda_{\max}(\mathbf{X}^T\mathbf{X}/n) \geqslant \lambda_{\min}(\mathbf{X}^T\mathbf{X}/n) \geqslant c_0 > 0$$

Example 2.12 证明最小二乘估计 $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ 满足：

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 = O_p(p^{1/2}n^{-1/2})$$

Proof. 我们有

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 = \|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\|_2$$

由假设 3，$\mathbf{X}^T\mathbf{X}/n$ 的特征值在 $[c_0, c_1]$ 之间，因此：

$$\|(\mathbf{X}^T\mathbf{X}/n)^{-1}\|_2 \leqslant \frac{1}{c_0}$$

于是：

$$\begin{aligned}
\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 &= \|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\|_2 \\
&= \left\|\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1}\frac{1}{n}\mathbf{X}^T\boldsymbol{\varepsilon}\right\|_2 \\
&\leqslant \left\|\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1}\right\|_2 \cdot \left\|\frac{1}{n}\mathbf{X}^T\boldsymbol{\varepsilon}\right\|_2 \\
&\leqslant \frac{1}{c_0}\left\|\frac{1}{n}\mathbf{X}^T\boldsymbol{\varepsilon}\right\|_2
\end{aligned}$$

现在考虑 $\left\|\frac{1}{n}\mathbf{X}^T\boldsymbol{\varepsilon}\right\|_2$. 在给定 $\mathbf{X}$ 的条件下：

$$\mathbb{E}\left[\frac{1}{n}\mathbf{X}^T\boldsymbol{\varepsilon}\middle|\mathbf{X}\right] = \mathbf{0}$$

$$\mathrm{Var}\left[\frac{1}{n}\mathbf{X}^T\boldsymbol{\varepsilon}\middle|\mathbf{X}\right] = \frac{1}{n^2}\mathbf{X}^T\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T|\mathbf{X}]\mathbf{X} = \frac{\sigma^2}{n^2}\mathbf{X}^T\mathbf{X}$$

因此：

$$\mathbb{E}\left[\left\|\frac{1}{n}\mathbf{X}^T\boldsymbol{\varepsilon}\right\|_2^2\middle|\mathbf{X}\right] = \mathrm{tr}\left(\frac{\sigma^2}{n^2}\mathbf{X}^T\mathbf{X}\right) = \frac{\sigma^2}{n^2}\mathrm{tr}(\mathbf{X}^T\mathbf{X})$$

由假设 3，$\mathrm{tr}(\mathbf{X}^T\mathbf{X}) \leqslant p\lambda_{\max}(\mathbf{X}^T\mathbf{X}) \leqslant pc_1n$，所以：

$$\mathbb{E}\left[\left\|\frac{1}{n}\mathbf{X}^T\boldsymbol{\varepsilon}\right\|_2^2\middle|\mathbf{X}\right] \leqslant \frac{\sigma^2pc_1}{n}$$

由马尔可夫不等式，对任意 $M > 0$：

$$\mathbb{P}\left(\left\|\frac{1}{n}\mathbf{X}^T\boldsymbol{\varepsilon}\right\|_2 > M\sqrt{\frac{p}{n}}\middle|\mathbf{X}\right) \leqslant \frac{\mathbb{E}\left[\left\|\frac{1}{n}\mathbf{X}^T\boldsymbol{\varepsilon}\right\|_2^2\middle|\mathbf{X}\right]}{M^2p/n} \leqslant \frac{\sigma^2c_1}{M^2}$$

因此：

$$\left\|\frac{1}{n}\mathbf{X}^T\boldsymbol{\varepsilon}\right\|_2 = O_p\left(\sqrt{\frac{p}{n}}\right)$$

最终得到：

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \leqslant \frac{1}{c_0} \cdot O_p\left(\sqrt{\frac{p}{n}}\right) = O_p\left(\sqrt{\frac{p}{n}}\right)$$

□

Example 2.13 考虑岭估计:

$$\hat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y}$$

1. 若 p 固定不变, $\|\boldsymbol{\beta}\|_2 = O(1)$, 证明对于任意 λ 满足 $\lambda = o(n^{1/2})$, 有:

$$\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\|_2 = O_p(n^{-1/2})$$

2. 若去掉"p 固定不变" 的假定, 建立 $\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\|_2$ 的收敛速率.

Proof. p 固定情形岭估计的偏差为:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta} &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y} - \boldsymbol{\beta} \\
&= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} \\
&= [(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X} - \mathbf{I}_p]\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}
\end{aligned}$$

注意恒等式:

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{I}_p - \lambda(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}$$

因此:

$$\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta} = -\lambda(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}$$

取范数:

$$\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\|_2 \leqslant \lambda\|(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\boldsymbol{\beta}\|_2 + \|(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\|_2$$

分析第一项:

$$\begin{aligned}
\lambda\|(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\boldsymbol{\beta}\|_2 &\leqslant \lambda\|(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\|_2\|\boldsymbol{\beta}\|_2 \\
&\leqslant \lambda \cdot \frac{1}{\lambda_{\min}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)} \cdot O(1) \\
&\leqslant \lambda \cdot \frac{1}{nc_0 + \lambda} \cdot O(1) \\
&= O\left(\frac{\lambda}{n}\right) \\
&= o(n^{-1/2})
\end{aligned}$$

分析第二项:

$$\begin{aligned}
\|(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\|_2 &\leqslant \|(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\|_2\|\mathbf{X}^T\boldsymbol{\varepsilon}\|_2 \\
&\leqslant \frac{1}{nc_0 + \lambda}\|\mathbf{X}^T\boldsymbol{\varepsilon}\|_2 \\
&\leqslant \frac{1}{nc_0}\|\mathbf{X}^T\boldsymbol{\varepsilon}\|_2
\end{aligned}$$

由问题一的证明, $\|\mathbf{X}^T\boldsymbol{\varepsilon}\|_2 = O_p(\sqrt{np})$, 当 p 固定时:

$$\frac{1}{nc_0}\|\mathbf{X}^T\boldsymbol{\varepsilon}\|_2 = O_p\left(\frac{\sqrt{n}}{n}\right) = O_p(n^{-1/2})$$

因此：

$$\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\|_2 \leqslant o(n^{-1/2}) + O_p(n^{-1/2}) = O_p(n^{-1/2})$$

在 p 不固定的情况下，我们需要更强的正则性条件. 假设：

1. $\|\boldsymbol{\beta}\|_2 = O(1)$ 或者 $\|\boldsymbol{\beta}\|_2 = O(\sqrt{p})$
2. $p = o(n)$ 或 $p/n \to \gamma \in [0, 1)$
3. λ 的选择满足 $\lambda = O(\sqrt{np})$

沿用 (i) 中的分解：

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\|_2 &\leqslant \lambda\|(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\boldsymbol{\beta}\|_2 + \|(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\|_2 \\ &= O\left(\frac{\lambda}{n}\right) + O_p\left(\sqrt{\frac{p}{n}}\right) \quad (\text{或者} O_p\left(\frac{\lambda\sqrt{p}}{n}\right) + O_p\left(\sqrt{\frac{p}{n}}\right))\end{aligned}$$

特别地，如果 p 固定，则回到 $O_p(n^{-1/2})$ 的结论. □

Chapter 3

渐进理论

§3.1 引入

Example 3.1 记 $\mathbf{x}_i = (1, x_i)^{\mathrm{T}}$，其中 x_i 是控制 y_i 取值为"0" 或"1" 的重要特征变量. 条件概率为 $P(y_i = 1 \mid \mathbf{x}_i) = p_i = 1 - P(y_i = 0 \mid \mathbf{x}_i)$，且

$$p_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)},$$

其中 $\eta_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$. 数理统计中，$\boldsymbol{\beta}$ 的对数似然函数为：

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i\eta_i - \log\left\{1 + \exp(\eta_i)\right\}\right].$$

需要回答两个核心问题：

1. $\widehat{\boldsymbol{\beta}} = \operatorname{argmax} l(\boldsymbol{\beta})$ 的含义和求解方法.
2. $\widehat{p}_i = \exp(\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) / \left\{1 + \exp(\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\right\}$ 的统计性质，以及如何构造区间 $\widehat{C}$ 使得 $P(\widehat{p}_i \in \widehat{C}) \geqslant 95\%$.

§3.2 Taylor 展开

此处从有限增量定理开始叙述 (参考了笔者的数学分析笔记).

Definition 3.2 ($\mathbb{R}^n$ 中的范数) 设 $L \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$，定义 L 的范数 $\|L\|$ 为

$$\|L\| = \sup_{|\boldsymbol{h}|=1} |L\boldsymbol{h}|.$$

Lemma 3.3 (有限增量定理) 设 E 是 $\mathbb{R}^n$ 中的凸开集，$f : E \longrightarrow \mathbb{R}^m$ 在 E 上可微，且存在 $M > 0$ 使得对任意的 $\boldsymbol{x} \in E$ 均有 $\left\|f'(\boldsymbol{x})\right\| \leqslant M$. 那么对任意的 $\boldsymbol{a}, \boldsymbol{b} \in E$ 有

$$|f(\boldsymbol{b}) - f(\boldsymbol{a})| \leqslant M|\boldsymbol{b} - \boldsymbol{a}|.$$

Lemma 3.4 设 E 是 $\mathbb{R}^n$ 中的区域，$f \in C(E)$ 具有 k 阶导数，$\mathbf{a}, \mathbf{h} \in X$ 满足 $[\mathbf{a}, \mathbf{a}+\mathbf{h}] \subseteq E$. 定义函数 $g(t) = f(\mathbf{a}+t\mathbf{h})$ ，则 g 也 k 阶可导，且有

$$g^{(k)}(t) = f^{(k)}(\mathbf{x}_0 + t\mathbf{h})\,\mathbf{h}^{\otimes k}$$

(其中 $(\overbrace{\mathbf{h}, \ldots, \mathbf{h}}^{k\text{ 个}})$ 记为 $\mathbf{h}^{\otimes k}$)

Proof. 对 k 用归纳法，$k=0$ 的情形平凡地成立，假设 $f \in C^{\ell}(E)(\ell < k)$ 时结论成立，考虑 $\ell+1$ 的情况. 因为 $f^{(\ell)}$ 可导，所以有如下方向导数

$$f^{(\ell+1)}(\mathbf{a}+t\mathbf{h})(\mathbf{h}) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\left(f^{(\ell)}(\mathbf{a}+t\mathbf{h}+\varepsilon\mathbf{h}) - f^{(\ell)}(\mathbf{a}+t\mathbf{h})\right)$$

将左右两边作用在 $\mathbf{h}^{\otimes \ell}$ 上，并利用多重线性映射的连续性可得

$$f^{(\ell+1)}(\mathbf{a}+t\mathbf{h})(\mathbf{h})\mathbf{h}^{\otimes \ell} = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\left(f^{(\ell)}(\mathbf{a}+t\mathbf{h}+\varepsilon\mathbf{h})\,\mathbf{h}^{\otimes \ell} - f^{(\ell)}(\mathbf{a}+t\mathbf{h})\,\mathbf{h}^{\otimes \ell}\right).$$

□

Remark 3.5 条件 $[\mathbf{a}, \mathbf{a}+\mathbf{h}] \subseteq E$ 也可以直接替换为 E 是凸集

Theorem 3.6 (Taylor 公式 (Lagrange 余项) (标量版)) 设 E, $\mathbf{a}$, $\mathbf{h}$ 同前，若 $f \in C(E)$ 具有 $k+1$ 阶导数，则存在一个 $\theta \in (0,1)$ ，使得

$$f(\mathbf{a}+\mathbf{h}) = \sum_{\ell=0}^{k} \frac{1}{\ell!} f^{(\ell)}(\mathbf{a})\,\mathbf{h}^{\otimes \ell} + R_k(\mathbf{x}; f, \mathbf{a}), R_k(\mathbf{x}; f, \mathbf{a}) = \frac{1}{(k+1)!} f^{(k+1)}(\mathbf{a}+\theta\mathbf{h})\,\mathbf{h}^{\otimes (k+1)}.$$

Remark 3.7 Taylor 多项式的前三项是很重要的，它们是

$$f(\boldsymbol{a}) + \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\boldsymbol{a}) h_j + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{a}) h_i h_j = f(\boldsymbol{a}) + f'(\boldsymbol{a})\boldsymbol{h} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{a}) h_i h_j.$$

若记

$$H_f(\boldsymbol{a}) = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2}(\boldsymbol{a}) & \dfrac{\partial^2 f}{\partial x_1 \partial x_2}(\boldsymbol{a}) & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n}(\boldsymbol{a}) \\ \dfrac{\partial^2 f}{\partial x_2 \partial x_1}(\boldsymbol{a}) & \dfrac{\partial^2 f}{\partial x_2^2}(\boldsymbol{a}) & \cdots & \dfrac{\partial^2 f}{\partial x_2 \partial x_n}(\boldsymbol{a}) \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial^2 f}{\partial x_n \partial x_1}(\boldsymbol{a}) & \dfrac{\partial^2 f}{\partial x_n \partial x_2}(\boldsymbol{a}) & \cdots & \dfrac{\partial^2 f}{\partial x_n^2}(\boldsymbol{a}) \end{bmatrix}.$$

则前三项可表示为

$$f(\boldsymbol{a}) + f'(\boldsymbol{a})\boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^{\mathrm{T}} H_f(\boldsymbol{a})\boldsymbol{h}.$$

我们称 $H_f(\boldsymbol{a})$ 为 f 在 $\boldsymbol{a}$ 处的 Hesse 矩阵.

Remark 3.8 此处我们引进了多重线性映射的符号，若采用常规符号，还有下列表示方法

1. 设 E 是 $\mathbb{R}^n$ 中的凸开集，$f : E \longrightarrow \mathbb{R}$. 如果 $f \in C^{r+1}(E)$，则对 E 中任意两点 $\boldsymbol{a}$ 和 $\boldsymbol{a}+\boldsymbol{h}$，存在 $\theta \in (0,1)$ 使得

$$f(\boldsymbol{a}+\boldsymbol{h}) = f(\boldsymbol{a}) + \sum_{1\leqslant|\boldsymbol{\alpha}|\leqslant r} \frac{1}{\boldsymbol{\alpha}!}\frac{\partial^{\boldsymbol{\alpha}} f}{\partial \boldsymbol{x}^{\boldsymbol{\alpha}}}(\boldsymbol{a})\boldsymbol{h}^{\alpha} + \sum_{|\boldsymbol{\alpha}|=r+1} \frac{1}{\boldsymbol{\alpha}!}\frac{\partial^{\boldsymbol{\alpha}} f}{\partial \boldsymbol{x}^{\boldsymbol{\alpha}}}(\boldsymbol{a}+\theta\boldsymbol{h})\boldsymbol{h}^{\boldsymbol{\alpha}}.$$

2. 假设 E 是 $\mathbb{R}^n$ 中的凸开集，$f : E \longrightarrow \mathbb{R}$. 如果 $f \in C^{r}(E)$，则对任意的 $\boldsymbol{a} \in E$，当 $\boldsymbol{h} \to \boldsymbol{0}$ 时有

$$f(\boldsymbol{a}+\boldsymbol{h}) = f(\boldsymbol{a}) + \sum_{1\leqslant|\boldsymbol{\alpha}|\leqslant r} \frac{1}{\boldsymbol{\alpha}!}\frac{\partial^{\alpha} f}{\partial \boldsymbol{x}^{\boldsymbol{\alpha}}}(\boldsymbol{a})\boldsymbol{h}^{\boldsymbol{\alpha}} + o\left(|\boldsymbol{h}|^r\right).$$

其中对 $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)$ 记

$$|\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_n, \quad \boldsymbol{\alpha}! = \alpha_1! \cdots \alpha_n!$$

以及

$$\frac{\partial^{\alpha} f}{\partial \boldsymbol{x}^{\alpha}} = \frac{\partial^{|\boldsymbol{\alpha}|} f}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}},$$

并对 $\boldsymbol{h} = (h_1, \cdots, h_n)^{\mathrm{T}} \in \mathbb{R}^n$ 记 $\boldsymbol{h}^{\alpha} = h_1^{\alpha_1} \cdots h_n^{\alpha_n}$.

例 (3.1) 第一问. 回忆逻辑回归的对数似然函数

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \mathbf{x}_i^T \boldsymbol{\beta} - \log(1+\exp(\mathbf{x}_i^T\boldsymbol{\beta}))\right]$$

分别计算其梯度和 Hessian 矩阵：

$$\nabla \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \mathbf{x}_i - p_i \mathbf{x}_i\right] = \mathbf{X}^T(\mathbf{y}-\mathbf{p})$$

$$\nabla^2 \ell(\boldsymbol{\beta}) = -\sum_{i=1}^{n} p_i(1-p_i)\mathbf{x}_i\mathbf{x}_i^T = -\mathbf{X}^T\mathbf{W}\mathbf{X}$$

其中 $p_i = \dfrac{\exp(\mathbf{x}_i^T\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^T\boldsymbol{\beta})}$，$\mathbf{W} = \mathrm{diag}(p_i(1-p_i))$ 利用 Newton-Raphson 算法：

$$\boldsymbol{\beta} = \boldsymbol{\beta}^{(t)} - \left[\mathbf{H}\left(\boldsymbol{\beta}^{(t)}\right)\right]^{-1} \mathbf{g}\left(\boldsymbol{\beta}^{(t)}\right)$$

代入得

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left(\mathbf{X}^{\mathrm{T}}\mathbf{W}^{(t)}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\left(\mathbf{y}-\mathbf{p}^{(t)}\right)$$

□

§3.3 Delta 方法及其应用

3.3.1 基本定理

Theorem 3.9 (一阶 Delta 方法) 设 $\{T_n\}$ 是一列随机变量，$\{\alpha_n\}$ 是一列实数且 $\alpha_n \to \infty$(通常 $\alpha_n = \sqrt{n}$). 如果存在参数 θ 和常数 $\sigma^2 > 0$，使得

$$\alpha_n(T_n - \theta) \xrightarrow{d} N(0, \sigma^2)$$

且函数 $g : \mathbb{R} \to \mathbb{R}$ 在点 θ 处可导，那么有：

$$\alpha_n[g(T_n) - g(\theta)] \xrightarrow{d} N(0, [g'(\theta)]^2\sigma^2)$$

Proof. 由于 g 在 θ 处可导，根据 Taylor 定理：

$$g(T_n) = g(\theta) + g'(\theta)(T_n - \theta) + R_n$$

其中余项 $R_n = o_p(|T_n - \theta|)$. 重整化并整理得：

$$\alpha_n[g(T_n) - g(\theta)] = g'(\theta) \cdot [\alpha_n(T_n - \theta)] + \alpha_n R_n$$

由于 $\alpha_n(T_n - \theta) = O_p(1)$，有 $T_n - \theta = O_p(1/\alpha_n)$，因此：

$$\alpha_n R_n = \alpha_n \cdot o_p(1/\alpha_n) = o_p(1) \xrightarrow{p} 0$$

应用 Slutsky 定理：

$$\alpha_n[g(T_n) - g(\theta)] = g'(\theta) \cdot \underbrace{[\alpha_n(T_n - \theta)]}_{\xrightarrow{d} N(0,\sigma^2)} + \underbrace{\alpha_n R_n}_{\xrightarrow{p} 0} \xrightarrow{d} N(0, [g'(\theta)]^2\sigma^2)$$

□

Theorem 3.10 (二阶 Delta 方法) 在相同的前提条件下，如果 $g'(\theta) = 0$ 且 $g''(\theta)$ 存在且连续，那么：

$$\alpha_n^2[g(T_n) - g(\theta)] \xrightarrow{d} \frac{1}{2}g''(\theta)\sigma^2\chi_1^2$$

其中 χ_1^2 是自由度为 1 的卡方分布.

Proof. 由二阶泰勒展开：

$$g(T_n) = g(\theta) + \underbrace{g'(\theta)}_{=0}(T_n - \theta) + \frac{1}{2}g''(\theta)(T_n - \theta)^2 + o_p((T_n - \theta)^2)$$

重整化得：

$$\alpha_n^2[g(T_n) - g(\theta)] = \frac{1}{2}g''(\theta)[\alpha_n(T_n - \theta)]^2 + \alpha_n^2 o_p((T_n - \theta)^2)$$

由于 $\alpha_n(T_n-\theta)\xrightarrow{d} Z\sim N(0,\sigma^2)$，根据连续映射定理：

$$[\alpha_n(T_n-\theta)]^2\xrightarrow{d} Z^2=\sigma^2\chi_1^2$$

余项 $\alpha_n^2 o_p((T_n-\theta)^2)=o_p(1)\xrightarrow{p} 0$. 应用 Slutsky 定理即得结论. □

Theorem 3.11 (多元 Delta 方法) 设 $\mathbf{T}_n$ 是一个 k 维随机向量，满足：

$$\alpha_n(\mathbf{T}_n-\boldsymbol{\theta})\xrightarrow{d} N(\mathbf{0},\Sigma)$$

且函数 $g:\mathbb{R}^k\to\mathbb{R}^m$ 在 $\boldsymbol{\theta}$ 处可微，其雅可比矩阵为 $Dg(\boldsymbol{\theta})(m\times k$ 矩阵). 那么：

$$\alpha_n[g(\mathbf{T}_n)-g(\boldsymbol{\theta})]\xrightarrow{d} N(\mathbf{0},Dg(\boldsymbol{\theta})\Sigma[Dg(\boldsymbol{\theta})]^T)$$

Proof. 由多元泰勒展开：

$$g(\mathbf{T}_n)=g(\boldsymbol{\theta})+Dg(\boldsymbol{\theta})(\mathbf{T}_n-\boldsymbol{\theta})+\mathbf{R}_n$$

其中 $\mathbf{R}_n=o_p(\|\mathbf{T}_n-\boldsymbol{\theta}\|)$. 重整化得：

$$\alpha_n[g(\mathbf{T}_n)-g(\boldsymbol{\theta})]=Dg(\boldsymbol{\theta})\cdot[\alpha_n(\mathbf{T}_n-\boldsymbol{\theta})]+\alpha_n\mathbf{R}_n$$

由于 $\alpha_n\mathbf{R}_n=o_p(1)\xrightarrow{p} 0$，应用 Slutsky 定理和多元正态分布的性质即得结论. □

3.3.2 例子

例 (3.1) 第二问. 由极大似然估计理论（Cramér 定理），对于 logistic 回归模型，有：

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\xrightarrow{d} N(\mathbf{0},\boldsymbol{\Sigma})$$

其中 $\boldsymbol{\Sigma}=I^{-1}(\boldsymbol{\beta}_0)$ 是 Fisher 信息矩阵的逆矩阵. 对于 logistic 回归，Fisher 信息矩阵为：

$$I(\boldsymbol{\beta}_0)=\mathbb{E}\left[-\frac{\partial^2\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\right]_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\implies \boldsymbol{\Sigma}=\left[\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}p_{0i}(1-p_{0i})\right]^{-1}$$

定义函数 $g:\mathbb{R}^2\to\mathbb{R}$ 为：

$$g(\boldsymbol{\beta})=\frac{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}$$

则 $\widehat{p}_i=g(\widehat{\boldsymbol{\beta}})$，$p_{0i}=g(\boldsymbol{\beta}_0)$. 计算 g 的梯度：

$$\nabla g(\boldsymbol{\beta})=\frac{\partial g(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=g(\boldsymbol{\beta})[1-g(\boldsymbol{\beta})]\mathbf{x}_i$$

特别地，

$$\nabla g(\boldsymbol{\beta}_0)=p_{0i}(1-p_{0i})\mathbf{x}_i$$

由 Delta 方法（一阶展开）:

$$\widehat{p}_i - p_{0i} = g(\widehat{\boldsymbol{\beta}}) - g(\boldsymbol{\beta}_0) \approx [\nabla g(\boldsymbol{\beta}_0)]^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

进而

$$\sqrt{n}(\widehat{p}_i - p_{0i}) \approx [\nabla g(\boldsymbol{\beta}_0)]^{\mathrm{T}}\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

由 Slutsky 定理和连续映射定理:

$$\sqrt{n}(\widehat{p}_i - p_{0i}) \xrightarrow{d} N(0, \sigma^2)$$

其中

$$\sigma^2 = [\nabla g(\boldsymbol{\beta}_0)]^{\mathrm{T}}\boldsymbol{\Sigma}[\nabla g(\boldsymbol{\beta}_0)] = p_{0i}^2(1 - p_{0i})^2\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{x}_i$$

由渐近正态性:

$$\frac{\sqrt{n}(\widehat{p}_i - p_{0i})}{\sigma} \xrightarrow{d} N(0, 1)$$

因此,

$$P\left(\left|\frac{\sqrt{n}(\widehat{p}_i - p_{0i})}{\sigma}\right| \leqslant z_{0.975}\right) \to 0.95$$

其中 $z_{0.975} \approx 1.96$ 是标准正态分布的 0.975 分位数. 等价地,

$$P\left(p_{0i} \in \left[\widehat{p}_i - 1.96\frac{\sigma}{\sqrt{n}}, \widehat{p}_i + 1.96\frac{\sigma}{\sqrt{n}}\right]\right) \to 0.95$$

由于 σ^2 依赖于未知参数 $\boldsymbol{\beta}_0$，用一致估计量代替:

- 用 $\widehat{p}_i$ 估计 p_{0i}
- 用 $\widehat{\boldsymbol{\Sigma}} = \left[\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\widehat{p}_i(1 - \widehat{p}_i)\right]^{-1}$ 估计 $\boldsymbol{\Sigma}$

得到方差估计:

$$\widehat{\sigma}^2 = \widehat{p}_i^2(1 - \widehat{p}_i)^2\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\mathbf{x}_i$$

最终置信区间为:

$$\widehat{C} = \left[\widehat{p}_i - 1.96\sqrt{\frac{\widehat{\sigma}^2}{n}}, \widehat{p}_i + 1.96\sqrt{\frac{\widehat{\sigma}^2}{n}}\right]$$

由 Slutsky 定理，替换一致估计量不影响渐近分布，因此:

$$P(p_{0i} \in \widehat{C}) \to 0.95$$

□

§3.4 逆矩阵方法

Definition 3.12 (Neumann 级数) 设 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 是 $p\times p$ 矩阵，且 $\boldsymbol{A}$ 可逆. 若矩阵级数

$$\sum_{k=0}^{\infty}[\boldsymbol{I}-\boldsymbol{A}^{-1}\boldsymbol{B}]^k$$

收敛，则 $\boldsymbol{B}^{-1}$ 存在且满足

$$\boldsymbol{B}^{-1}=\left(\sum_{k=0}^{\infty}[\boldsymbol{I}-\boldsymbol{A}^{-1}\boldsymbol{B}]^k\right)\boldsymbol{A}^{-1}$$

特别地，当 $\|\boldsymbol{I}-\boldsymbol{A}^{-1}\boldsymbol{B}\|<1$（在某种矩阵范数下）时，该级数绝对收敛.

Theorem 3.13 (矩阵逆的扰动展开) 设 $\boldsymbol{A}$ 和 $\boldsymbol{B}=\boldsymbol{A}+\boldsymbol{\Delta}$ 是 $p\times p$ 矩阵，$\boldsymbol{A}$ 可逆，且 $\|\boldsymbol{A}^{-1}\boldsymbol{\Delta}\|<1$. 则 $\boldsymbol{B}$ 可逆，且

$$\boldsymbol{B}^{-1}=\boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\boldsymbol{\Delta}\boldsymbol{A}^{-1}+\boldsymbol{A}^{-1}\boldsymbol{\Delta}\boldsymbol{A}^{-1}\boldsymbol{\Delta}\boldsymbol{A}^{-1}-\cdots$$

等价地，可以写成

$$\boldsymbol{B}^{-1}=\sum_{k=0}^{l}(-1)^k(\boldsymbol{A}^{-1}\boldsymbol{\Delta})^k\boldsymbol{A}^{-1}+\boldsymbol{R}_{l+1}$$

其中余项 $\boldsymbol{R}_{l+1}=(-1)^{l+1}(\boldsymbol{A}^{-1}\boldsymbol{\Delta})^{l+1}\boldsymbol{B}^{-1}$.

Proof. 由条件 $\|\boldsymbol{A}^{-1}\boldsymbol{\Delta}\|<1$，Neumann 级数收敛. 注意到

$$\boldsymbol{B}=\boldsymbol{A}(\boldsymbol{I}+\boldsymbol{A}^{-1}\boldsymbol{\Delta})$$

因此

$$\boldsymbol{B}^{-1}=(\boldsymbol{I}+\boldsymbol{A}^{-1}\boldsymbol{\Delta})^{-1}\boldsymbol{A}^{-1}=\left(\sum_{k=0}^{\infty}(-1)^k(\boldsymbol{A}^{-1}\boldsymbol{\Delta})^k\right)\boldsymbol{A}^{-1}$$

截断到第 l 项即得所需结果. □

§3.5 Laplace 方法

Definition 3.14 (Laplace 方法) 设 $f:[a,b]\to\mathbb{R}$ 是一个充分光滑的函数，$M>0$ 是一个大参数. 假设存在唯一的 $x_0\in(a,b)$ 使得:

1. $f'(x_0)=0$（临界点条件）
2. $f''(x_0)<0$（极大值条件）
3. 对任意 $\delta>0$，$\sup\limits_{|x-x_0|\geqslant\delta}f(x)<f(x_0)$（唯一最大值条件）

考虑积分：

$$I(M)=\int_a^b e^{Mf(x)}dx$$

则当 $M\to\infty$ 时，有如下渐近展开：

$$I(M)=e^{Mf(x_0)}\sqrt{\frac{2\pi}{-Mf''(x_0)}}\left[1+O\left(\frac{1}{M}\right)\right]$$

Definition 3.15 设 $\mathbf{x}\in\mathbb{R}^d$，$f:\mathbb{R}^d\to\mathbb{R}$ 充分光滑，存在唯一的 $\mathbf{x}_0\in\mathbb{R}^d$ 使得：

1. $\nabla f(\mathbf{x}_0)=0$
2. Hessian 矩阵 $H(\mathbf{x}_0)$ 负定
3. 对任意 $\delta>0$，$\sup\limits_{\|\mathbf{x}-\mathbf{x}_0\|\geqslant\delta} f(\mathbf{x})<f(\mathbf{x}_0)$

考虑积分：

$$I(M)=\int_{\mathbb{R}^d} e^{Mf(\mathbf{x})}d\mathbf{x}$$

则当 $M\to\infty$ 时，有如下渐近展开：

$$I(M)=e^{Mf(\mathbf{x}_0)}\frac{(2\pi)^{d/2}}{M^{d/2}\sqrt{|\det(H(\mathbf{x}_0))|}}\left[1+O\left(\frac{1}{M}\right)\right]$$

Proof. 步骤 **1**：局部化：对于任意 $\delta>0$，将积分分解为：

$$I(M)=\int_{x_0-\delta}^{x_0+\delta} e^{Mf(x)}dx+\int_{|x-x_0|>\delta} e^{Mf(x)}dx$$

由于 $f(x)$ 在 x_0 处取得唯一最大值，存在 $\eta>0$ 使得：

$$\sup_{|x-x_0|>\delta} f(x)\leq f(x_0)-\eta$$

因此第二项满足：

$$\int_{|x-x_0|>\delta} e^{Mf(x)}dx=O\left(e^{M(f(x_0)-\eta)}\right)$$

步骤 **2**：泰勒展开与变量替换：在 x_0 的邻域内，$f(x)$ 的泰勒展开为：

$$f(x)=f(x_0)+\frac{1}{2}f''(x_0)(x-x_0)^2+R(x)$$

令 $t=\sqrt{-Mf''(x_0)}(x-x_0)$，则：

$$x-x_0=\frac{t}{\sqrt{-Mf''(x_0)}},\quad dx=\frac{dt}{\sqrt{-Mf''(x_0)}}$$

且

$$\frac{M}{2}f''(x_0)(x-x_0)^2=-\frac{t^2}{2}$$

步骤 3：主要项的估计：主要积分项变为：

$$I_0(M) = \frac{e^{Mf(x_0)}}{\sqrt{-Mf''(x_0)}} \int_{-\delta\sqrt{-Mf''(x_0)}}^{\delta\sqrt{-Mf''(x_0)}} \exp\left\{-\frac{t^2}{2} + MR\left(x_0 + \frac{t}{\sqrt{-Mf''(x_0)}}\right)\right\} dt$$

由于 $R(x) = o((x - x_0)^2)$，当 $M \to \infty$ 时：

$$MR\left(x_0 + \frac{t}{\sqrt{-Mf''(x_0)}}\right) = o(1)$$

因此：

$$e^{MR\left(x_0 + \frac{t}{\sqrt{-Mf''(x_0)}}\right)} = 1 + o(1)$$

步骤 4：扩展到整个实轴当 $M \to \infty$ 时，积分限趋于无穷，且误差项一致收敛到 0，因此：

$$I_0(M) = \frac{e^{Mf(x_0)}}{\sqrt{-Mf''(x_0)}} \left[\int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt + o(1)\right]$$

利用高斯积分公式 $\int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt = \sqrt{2\pi}$，得到：

$$I_0(M) = e^{Mf(x_0)} \sqrt{\frac{2\pi}{-Mf''(x_0)}} (1 + o(1))$$

步骤 5：总积分估计结合局部化步骤的结果：

$$I(M) = I_0(M) + O\left(e^{M(f(x_0)-\eta)}\right) = e^{Mf(x_0)} \sqrt{\frac{2\pi}{-Mf''(x_0)}} (1 + o(1))$$

更精确地，可以证明误差项为 $O(M^{-1})$：

$$I(M) = e^{Mf(x_0)} \sqrt{\frac{2\pi}{-Mf''(x_0)}} \left[1 + O\left(\frac{1}{M}\right)\right]$$

□

Chapter 4

估计理论

§4.1 M 估计

4.1.1 基本定义

Definition 4.1 (Kullback-Leibler 离差) 考虑两个概率密度函数 $f(x)$ 和 $g(x)$. 由 $g(x)$ 到 $f(x)$ 的 Kullback-Leibler (KL) 离差或者相对熵被定义为

$$I_{\mathrm{KL}} = \int f(x)\log\left\{\frac{f(x)}{g(x)}\right\} d\mu(x) = \mathbb{E}_X\left[\log\left\{\frac{f(X)}{g(X)}\right\}\right]$$

其中 $\mu(\cdot)$ 是 $\mathbb{R}$ 上的固定测度.

Definition 4.2 (M 估计) 设 $X_1, \ldots, X_n$ 为来自分布 P 的独立同分布随机变量.M 估计量 $\hat{\boldsymbol{\theta}}_n$ 定义为:

$$\hat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\Theta} M_n(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} m(\boldsymbol{\theta}, X_i)$$

其中 $\Theta \subseteq \mathbb{R}^p$ 为参数空间. 等价地, $\hat{\boldsymbol{\theta}}_n$ 满足:

$$\boldsymbol{\Psi}_n(\hat{\boldsymbol{\theta}}_n) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(\boldsymbol{\theta}, X_i) = \mathbf{0}$$

其中 $\boldsymbol{\psi}(\boldsymbol{\theta}, X) = \nabla_{\boldsymbol{\theta}} m(\boldsymbol{\theta}, X)$ 为影响函数.

Definition 4.3 在最大化框架中:

- 个体目标函数: $m(\boldsymbol{\theta}, X)$
- 样本目标函数: $M_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} m(\boldsymbol{\theta}, X_i) = \mathbb{P}_n m_{\boldsymbol{\theta}}$
- 总体目标函数: $M(\boldsymbol{\theta}) = \mathbb{E}[m(\boldsymbol{\theta}, X)] = P m_{\boldsymbol{\theta}}$

M 估计量定义为：$\hat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\Theta} M_n(\boldsymbol{\theta})$

Definition 4.4 在估计方程框架中：

- 个体估计函数: $\boldsymbol{\psi}(\boldsymbol{\theta}, X) = \nabla_{\boldsymbol{\theta}} m(\boldsymbol{\theta}, X)$
- 样本估计方程: $\boldsymbol{\Psi}_n(\boldsymbol{\theta}) = \dfrac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(\boldsymbol{\theta}, X_i) = \mathbb{P}_n \boldsymbol{\psi}_{\boldsymbol{\theta}}$
- 总体估计方程: $\boldsymbol{\Psi}(\boldsymbol{\theta}) = \mathbb{E}[\boldsymbol{\psi}(\boldsymbol{\theta}, X)] = P\boldsymbol{\psi}_{\boldsymbol{\theta}}$

M 估计量满足：$\boldsymbol{\Psi}_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$

Definition 4.5

$$\begin{aligned}
&\text{个体海森函数：} && \nabla_{\boldsymbol{\theta}} \boldsymbol{\psi}(\boldsymbol{\theta}, X)^T = \nabla^2_{\boldsymbol{\theta}} m(\boldsymbol{\theta}, X) \\
&\text{经验海森矩阵：} && \mathbf{H}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \boldsymbol{\psi}(\boldsymbol{\theta}, X_i)^T = \mathbb{P}_n \nabla_{\boldsymbol{\theta}} \boldsymbol{\psi}_{\boldsymbol{\theta}} \\
&\text{总体海森矩阵：} && \mathbf{H}(\boldsymbol{\theta}) = \mathbb{E}[\nabla_{\boldsymbol{\theta}} \boldsymbol{\psi}(\boldsymbol{\theta}, X)^T] = P\nabla_{\boldsymbol{\theta}} \boldsymbol{\psi}_{\boldsymbol{\theta}}
\end{aligned}$$

Definition 4.6 影响函数协方差矩阵：

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathbb{E}[\boldsymbol{\psi}(\boldsymbol{\theta}, X)\boldsymbol{\psi}(\boldsymbol{\theta}, X)^T] = P[\boldsymbol{\psi}_{\boldsymbol{\theta}} \boldsymbol{\psi}_{\boldsymbol{\theta}}^T]$$

4.1.2 相合性

Theorem 4.7 若 $M_n(\cdot)$ 为一随机函数，$M(\cdot)$ 为一非随机函数，满足

$$\sup_{\boldsymbol{\theta}\in\Theta} |M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})| = o_p(1)$$

另外，$\boldsymbol{\theta}_0$ 为 $M(\boldsymbol{\theta})$ 的可分离的最大值点，即对于任意固定的 $\delta > 0$，

$$\sup_{\boldsymbol{\theta}\in\{\boldsymbol{\theta}: d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) > \delta\}} M(\boldsymbol{\theta}) < M(\boldsymbol{\theta}_0),$$

则任意满足 $M_n(\widetilde{\boldsymbol{\theta}}) \geqslant \sup_{\boldsymbol{\theta}\in\Theta} M_n(\boldsymbol{\theta}) - o_p(1)$ 的 $\widetilde{\boldsymbol{\theta}}$ 均是 $\boldsymbol{\theta}_0$ 的弱相合估计.

Proof. 由 $\widetilde{\boldsymbol{\theta}}$ 的定义以及一致收敛条件知

$$\begin{aligned}
M_n(\widetilde{\boldsymbol{\theta}}) &\geqslant M_n(\boldsymbol{\theta}_0) - o_p(1) \\
&= M(\boldsymbol{\theta}_0) + [M_n(\boldsymbol{\theta}_0) - M(\boldsymbol{\theta}_0)] - o_p(1) \\
&= M(\boldsymbol{\theta}_0) - o_p(1).
\end{aligned}$$

因此

$$\begin{aligned}
0 &\leqslant M(\boldsymbol{\theta}_0) - M(\widetilde{\boldsymbol{\theta}}) \\
&\leqslant [M(\boldsymbol{\theta}_0) - M_n(\widetilde{\boldsymbol{\theta}})] + [M_n(\widetilde{\boldsymbol{\theta}}) - M(\widetilde{\boldsymbol{\theta}})] \\
&\leqslant o_p(1) + |M_n(\widetilde{\boldsymbol{\theta}}) - M(\widetilde{\boldsymbol{\theta}})| \\
&\leqslant \sup_{\boldsymbol{\theta}\in\Theta} |M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})| + o_p(1) \\
&= o_p(1).
\end{aligned}$$

另外，在条件 (2) 下，对于任意的 $\delta > 0$，总存在 $\eta > 0$ 使得

$$\{d(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) > \delta\} \subseteq \{M(\boldsymbol{\theta}_0) - M(\widetilde{\boldsymbol{\theta}}) > \eta\}$$

因此，对于任意固定的 $\delta > 0$，

$$P(d(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) > \delta) \leqslant P(M(\boldsymbol{\theta}_0) - M(\widetilde{\boldsymbol{\theta}}) > \eta) \to 0.$$

□

Definition 4.8 设 Θ 为参数空间，$\{M_n(\cdot)\}_{n=1}^{\infty} \subseteq \Theta$. 对任意的正数 ε 与 η，总存在一个正的随机变量 $\Delta_n(\varepsilon, \eta)$ 和一个不依赖于随机因素的正整数 $N_{\varepsilon,\eta}$，满足以下两个条件：

1. 当 $n \geqslant N_{\varepsilon,\eta}$ 时，随机变量 $\Delta_n(\varepsilon, \eta)$ 超过 ε 的概率小于 η，即
$$P\left(\Delta_n(\varepsilon, \eta) > \varepsilon\right) < \eta;$$
2. 对任意的参数 $\boldsymbol{\theta} \in \Theta$，存在一个包含 $\boldsymbol{\theta}$ 的开集 $E(\boldsymbol{\theta}, \varepsilon, \eta)$，使得对该开集中的任意参数 $\boldsymbol{\theta}'$，均有
$$\sup_{\boldsymbol{\theta}' \in E(\boldsymbol{\theta}, \varepsilon, \eta)} \left|M_n\left(\boldsymbol{\theta}'\right) - M_n(\boldsymbol{\theta})\right| \leqslant \Delta_n(\varepsilon, \eta),$$

则称函数列 $\{M_n(\cdot)\}_{n=1}^{\infty}$ 是随机等度连续的.

Theorem 4.9 (Newey) 若参数空间 Θ 是紧的，并且

$$M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta}) = o_p(1) \quad \forall \boldsymbol{\theta} \in \Theta$$

若 $M_n(\cdot)$ 是随机等度连续的，则

$$\sup_{\boldsymbol{\theta} \in \Theta} |M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})| = o_p(1)$$

Theorem 4.10 若参数空间 Θ 是紧的，并且

$$M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta}) = o_p(1) \quad \forall \boldsymbol{\theta} \in \Theta$$

若存在 $K_n = O_p(1)$ 且

$$\left|M_n\left(\boldsymbol{\theta}'\right) - M_n(\boldsymbol{\theta})\right| \leqslant K_n d\left(\boldsymbol{\theta}', \boldsymbol{\theta}\right)$$

则

$$\sup_{\boldsymbol{\theta} \in \Theta} |M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})| = o_p(1)$$

Proof. 由于 Θ 是紧集，对于任意 $\varepsilon > 0$，存在有限个点 $1, 2, \cdots, m(\varepsilon)$ 使得：

$$\Theta \subset \bigcup_{i=1}^{m(\varepsilon)} B\left(\theta_i, \varepsilon\right)$$

由点态收敛条件，对每个中心点 θ_i：

$$\left|M_n\left(\theta_i\right) - M\left(\theta_i\right)\right| = o_p(1)$$

由于只有有限个点，可得：

$$\max_{1\leqslant i\leqslant m(\varepsilon)} \left|M_n\left(\theta_i\right)-M\left(\theta_i\right)\right| = o_p(1)$$

对于任意 $\theta\in\Theta$ ，存在某个 $i\in\{1,2,\ldots,m(\varepsilon)\}$ 使得 $\theta\in B\left(\theta_i,\varepsilon\right)$. 于是有：

$$\begin{aligned}|M_n(\theta)-M(\theta)| &\leqslant |M_n(\theta)-M_n\left(\theta_i\right)|+|M_n\left(\theta_i\right)-M\left(\theta_i\right)|+|M\left(\theta_i\right)-M(\theta)| \\ &\leqslant \underbrace{K_n d\left(\theta,\theta_i\right)}_{\leqslant K_n\varepsilon}+\underbrace{|M_n\left(\theta_i\right)-M\left(\theta_i\right)|}_{o_p(1)}+\underbrace{\sup_{(\theta,\theta_i)<\varepsilon}|M\left(\theta_i\right)-M(\theta)|}_{\to 0}\end{aligned}$$

完成证明.

□

4.1.3 渐进正态性

Theorem 4.11 (经典泰勒展开方法) 假设：

1. $\hat{\boldsymbol{\theta}}_n\xrightarrow{p}\boldsymbol{\theta}_0$ (相合性)
2. $\boldsymbol{\psi}_{\boldsymbol{\theta}}(X)$ 在 $\boldsymbol{\theta}_0$ 的邻域内二阶连续可微
3. $\mathbb{E}[\|\boldsymbol{\psi}_{\boldsymbol{\theta}_0}(X)\|^2]<\infty$
4. $\mathbb{E}[\nabla_{\boldsymbol{\theta}}\boldsymbol{\psi}_{\boldsymbol{\theta}_0}(X)]$ 非奇异

则：

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)\xrightarrow{d}N\left(\mathbf{0},\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1})^T\right)$$

其中 $\mathbf{V}_{\boldsymbol{\theta}_0}=\mathbb{E}[\nabla_{\boldsymbol{\theta}}\boldsymbol{\psi}_{\boldsymbol{\theta}_0}(X)]$，$\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}=\mathbb{E}[\boldsymbol{\psi}_{\boldsymbol{\theta}_0}(X)\boldsymbol{\psi}_{\boldsymbol{\theta}_0}(X)^T]$

Proof. 在 $\boldsymbol{\theta}_0$ 处对 $\boldsymbol{\Psi}_n(\hat{\boldsymbol{\theta}}_n)$ 进行 Taylor 展开：

$$\mathbf{0}=\boldsymbol{\Psi}_n(\hat{\boldsymbol{\theta}}_n)=\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0)+\nabla_{\boldsymbol{\theta}}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)+\boldsymbol{R}_n$$

其中余项 $\boldsymbol{R}_n$ 满足：

$$\boldsymbol{R}_n=\frac{1}{2}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)^T\nabla_{\boldsymbol{\theta}}^2\boldsymbol{\Psi}_n(\tilde{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)$$

这里 $\tilde{\boldsymbol{\theta}}_n$ 位于 $\hat{\boldsymbol{\theta}}_n$ 和 $\boldsymbol{\theta}_0$ 之间，整理上式子得到：

$$\sqrt{n}\nabla_{\boldsymbol{\theta}}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)=-\sqrt{n}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0)-\sqrt{n}\boldsymbol{R}_n$$

由于 $\nabla_{\boldsymbol{\theta}}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0)=\mathbf{H}_n(\boldsymbol{\theta}_0)$，我们有：

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)=-\mathbf{H}_n(\boldsymbol{\theta}_0)^{-1}\left[\sqrt{n}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0)+\sqrt{n}\boldsymbol{R}_n\right]$$

由定义：

$$\sqrt{n}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\boldsymbol{\psi}_{\boldsymbol{\theta}_0}(X_i)$$

根据条件 3，$\mathbb{E}[\|\boldsymbol{\psi}_{\boldsymbol{\theta}_0}(X)\|^2] < \infty$，且 $\mathbb{E}[\boldsymbol{\psi}_{\boldsymbol{\theta}_0}(X)] = \boldsymbol{\Psi}(\boldsymbol{\theta}_0) = \mathbf{0}$. 由多元中心极限定理：

$$\sqrt{n}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0})$$

其中 $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0} = \mathbb{E}[\boldsymbol{\psi}_{\boldsymbol{\theta}_0}(X)\boldsymbol{\psi}_{\boldsymbol{\theta}_0}(X)^T]$. 由定义：

$$\mathbf{H}_n(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}}\boldsymbol{\psi}_{\boldsymbol{\theta}_0}(X_i)^T$$

根据大数定律：

$$\mathbf{H}_n(\boldsymbol{\theta}_0) \xrightarrow{p} \mathbb{E}[\nabla_{\boldsymbol{\theta}}\boldsymbol{\psi}_{\boldsymbol{\theta}_0}(X)^T] = \mathbf{V}_{\boldsymbol{\theta}_0}^T$$

由于矩阵求逆是连续函数，由连续映射定理：

$$\mathbf{H}_n(\boldsymbol{\theta}_0)^{-1} \xrightarrow{p} (\mathbf{V}_{\boldsymbol{\theta}_0}^T)^{-1}$$

余项可以写为：

$$\boldsymbol{R}_n = \frac{1}{2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T \nabla_{\boldsymbol{\theta}}^2 \boldsymbol{\Psi}_n(\tilde{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$$

由于 $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = O_p(n^{-1/2})$（这可以从相合性推导出来），我们有：

$$(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = O_p(n^{-1})$$

如果二阶导数 $\nabla_{\boldsymbol{\theta}}^2 \boldsymbol{\Psi}_n(\tilde{\boldsymbol{\theta}}_n)$ 有概率界（这由条件 2 和相合性保证），那么：

$$\boldsymbol{R}_n = O_p(n^{-1})$$

因此：

$$\sqrt{n}\boldsymbol{R}_n = \sqrt{n} \cdot O_p(n^{-1}) = O_p(n^{-1/2}) = o_p(1)$$

将上述结果组合：

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -\mathbf{H}_n(\boldsymbol{\theta}_0)^{-1}\left[\sqrt{n}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0) + o_p(1)\right]$$

由 Slutsky 定理：

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) &\xrightarrow{d} -(\mathbf{V}_{\boldsymbol{\theta}_0}^T)^{-1} \cdot N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}) \\
&= N\left(\mathbf{0}, (\mathbf{V}_{\boldsymbol{\theta}_0}^T)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}[(\mathbf{V}_{\boldsymbol{\theta}_0}^T)^{-1}]^T\right) \\
&= N\left(\mathbf{0}, \mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1})^T\right)
\end{aligned}$$

其中最后一个等式是因为 $(\mathbf{V}_{\boldsymbol{\theta}_0}^T)^{-1} = (\mathbf{V}_{\boldsymbol{\theta}_0}^{-1})^T$. □

Theorem 4.12 当样本量 $n \to \infty$ 且参数维度 p 固定时，如果：

1. $n^{1/2}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0) \xrightarrow{d} \mathbf{Z}_0$
2. 存在 $\widehat{K}_n = O_P(1)$ 使得 $\|\mathbf{H}_n(\boldsymbol{\theta}_0 + \mathbf{u}) - \mathbf{H}_n(\boldsymbol{\theta}_0)\| \leqslant \widehat{K}_n\|\mathbf{u}\|$
3. $-\mathbf{H}_n(\boldsymbol{\theta}_0) \xrightarrow{p} \mathbf{J}_0$ (正定)

则：

1. $\|n^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\| = O_P(1)$

2. $n^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \xrightarrow{d} \mathbf{J}_0^{-1}\mathbf{Z}_0$

<u>**Proof.**</u> 在 $\boldsymbol{\theta}_0$ 处进行泰勒展开：

$$
\begin{aligned}
&0 = \boldsymbol{\Psi}_n(\hat{\boldsymbol{\theta}}_n) = \boldsymbol{\Psi}_n(\boldsymbol{\theta}_0) + \mathbf{H}_n(\tilde{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\
\Longrightarrow\ &\mathbf{H}_n(\tilde{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0) \\
\Longrightarrow\ &\sqrt{n}\mathbf{H}_n(\tilde{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -\sqrt{n}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0) \\
\Longrightarrow\ &\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -\mathbf{H}_n(\tilde{\boldsymbol{\theta}}_n)^{-1}\sqrt{n}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0)
\end{aligned}
$$

其中 $\tilde{\boldsymbol{\theta}}_n$ 在 $\hat{\boldsymbol{\theta}}_n$ 和 $\boldsymbol{\theta}_0$ 之间. 由条件 2: $\|\mathbf{H}_n(\boldsymbol{\theta}_0+\mathbf{u}) - \mathbf{H}_n(\boldsymbol{\theta}_0)\| \leqslant \widehat{K}_n\|\mathbf{u}\|$，且 $\widehat{K}_n = O_P(1)$，可得：

$$
\mathbf{H}_n(\tilde{\boldsymbol{\theta}}_n) = \mathbf{H}_n(\boldsymbol{\theta}_0) + O_P(\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|) \implies \mathbf{H}_n(\tilde{\boldsymbol{\theta}}_n) = \mathbf{H}_n(\boldsymbol{\theta}_0) + o_P(1) \quad (\text{相合性})
$$

由 Slutsky 定理：

$$
\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -\underbrace{\mathbf{H}_n(\tilde{\boldsymbol{\theta}}_n)^{-1}}_{\xrightarrow{p}\mathbf{H}_n(\boldsymbol{\theta}_0)^{-1}\xrightarrow{p}\mathbf{J}_0^{-1}}\ \underbrace{\sqrt{n}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0)}_{\xrightarrow{d}\mathbf{Z}_0} \xrightarrow{d} -\mathbf{J}_0^{-1}\mathbf{Z}_0
$$

由上述表达式，显然有：

$$
\|n^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\| = O_P(1)
$$

□

§4.2 经验分布

4.2.1 经验分布函数

Definition 4.13 设 $X_1,\cdots,X_n$ 是由实数轴上分布函数 F 产生的随机样本，则经验分布函数被定义为

$$
F_n(x) = \frac{\sum_{i=1}^n \mathbb{I}[x_i,+\infty)(x)}{n} = \frac{\sum_{i=1}^n \mathbb{I}(X_i \leqslant x)}{n}
$$

这里 $\mathbb{I}_A(x)$ 表示集合 A 对应的示性函数.

Theorem 4.14 经验分布函数具有很多良好的性质

1. 分布函数：$F_n(x)$ 非递减，右连续.

2. 无偏性：$F_n(x)$ 是 $F(x)$ 的无偏估计：$E[F_n(x)] = F(x)$

3. 相合性：$F_n(x) \xrightarrow{P} F(x) \quad n \to \infty$

4. 渐进正态性：$\sqrt{n}(F_n(x) - F(x)) \xrightarrow{d} N(0, F(x)(1 - F(x)))$

5. 分布性质：$nF_n(x) \sim \text{Binomial}(n, F(x))$

Proof. 仅证 (3)(4)(5)

1. 由强大数定律：

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leqslant x) \xrightarrow{a.s.} E[I(X_1 \leqslant x)] = F(x)$$

几乎必然收敛蕴含依概率收敛.

2. 令 $Y_i = I(X_i \leqslant x)$，则 $Y_1, \ldots, Y_n$ 是独立同分布的 Bernoulli 随机变量，均值为 $F(x)$，方差为 $F(x)(1 - F(x))$. 由中心极限定理：

$$\frac{\sum_{i=1}^{n} Y_i - nF(x)}{\sqrt{nF(x)(1 - F(x))}} \xrightarrow{d} N(0, 1)$$

即：

$$\sqrt{n}(F_n(x) - F(x)) \xrightarrow{d} N(0, F(x)(1 - F(x)))$$

3.

$$nF_n(x) = \sum_{i=1}^{n} I(X_i \leqslant x)$$

每个 $I(X_i \leqslant x)$ 是独立的 Bernoulli 试验，成功概率为 $F(x)$，因此和为二项分布.

□

Theorem 4.15 (Glivenko-Cantelli) 设 $X_1, \cdots, X_n$ 是取自总体分布函数为 $F(x)$ 的独立同分布样本，$\mathbb{F}_n(x)$ 是其经验分布函数. 则当 $n \to \infty$ 时，有

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow{\text{a.s.}} 0$$

Proof. 由 Kolmogorov 强大数定律，对任意固定的 $t \in \mathbb{R}$：

$$\mathbb{F}_n(t) \xrightarrow{\text{a.s.}} F(t), \quad \mathbb{F}_n(t-) \xrightarrow{\text{a.s.}} F(t-)$$

由于 F 是单调非降函数，值域在 $[0, 1]$ 中. 给定 $\varepsilon > 0$，构造分割：

$$-\infty = t_0 < t_1 < \cdots < t_k = \infty$$

使得对每个 $i = 1, \ldots, k$：

$$F(t_i-) - F(t_{i-1}) < \varepsilon$$

对于任意 $t \in [t_{i-1}, t_i)$，利用单调性：

$$\mathbb{F}_n(t) - F(t) \leqslant \mathbb{F}_n(t_i-) - F(t_{i-1}) \leqslant [\mathbb{F}_n(t_i-) - F(t_i-)] + \varepsilon$$
$$\mathbb{F}_n(t) - F(t) \geqslant \mathbb{F}_n(t_{i-1}) - F(t_i-) \geqslant [\mathbb{F}_n(t_{i-1}) - F(t_{i-1})] - \varepsilon$$

上述论证表明，$\mathbb{F}_n(t)$ 和 $\mathbb{F}_n(t-)$ 在 $(t_{i-1}, t_i]$ 内一致收敛. 再利用 k 的有限性得到结论. □

Corollary 4.16 定理表明经验分布函数 F_n 是总体分布函数 F 的强相合一致估计. 这意味着:

$$\lim_{n\to\infty} P\left(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > \varepsilon\right) = 0 \quad \forall\varepsilon > 0$$

4.2.2 经验过程

Definition 4.17 (经验过程) 设 $F_n(x)$ 是经验分布函数, $F(x)$ 是总体分布函数. 定义经验过程 $\mathbb{G}_n$ 为:

$$\mathbb{G}_n(x) = \sqrt{n}(F_n(x) - F(x))$$

Proposition 4.18 (数字特征) 对任意的 $x, s, t \in \mathbb{R}$ 我们有

1. $E[\mathbb{G}_n(x)] = 0$
2. $\mathrm{Cov}[\mathbb{G}_n(s), \mathbb{G}_n(t)] = F(s \wedge t) - F(s)F(t)$

协方差结构的推导. 对于 $s, t \in \mathbb{R}$:

$$\begin{aligned}
\mathrm{Cov}[G_n(s), G_n(t)] &= n \cdot \mathrm{Cov}[F_n(s), F_n(t)] \\
&= n \cdot \mathrm{Cov}\left[\frac{1}{n}\sum I(X_i \leqslant s), \frac{1}{n}\sum I(X_j \leqslant t)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{Cov}[I(X_i \leqslant s), I(X_i \leqslant t)] \\
&= E[I(X \leqslant s)I(X \leqslant t)] - E[I(X \leqslant s)]E[I(X \leqslant t)] \\
&= F(s \wedge t) - F(s)F(t)
\end{aligned}$$

□

Theorem 4.19 对于任意有限个点 $x_1, x_2, \ldots, x_k$, 随机向量:

$$(\mathbb{G}_n(x_1), \mathbb{G}_n(x_2), \ldots, \mathbb{G}_n(x_k)) \xrightarrow{d} N(0, \Sigma)$$

其中 $\Sigma_{ij} = F(x_i \wedge x_j) - F(x_i)F(x_j)$.

Proof. 这是多元中心极限定理的直接推论. □

4.2.3 经验分布

Definition 4.20 (经验分布) 在概率空间 $(\Omega, \mathcal{A}, P)$ 上, 设 $X_1, \ldots, X_n$ 是独立同分布的随机变量. 定义经验分布为:

$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

其中 δ_x 是在点 x 的 **Dirac** 测度（退化分布）.

Remark 4.21 Dirac 测度 δ_x 的定义为：

$$\delta_x(A) = \begin{cases} 1 & \text{如果} x \in A \\ 0 & \text{如果} x \notin A \end{cases}$$

对于任何可测集合 A.

Definition 4.22 (经验过程的一般定义) 对于任意可测函数 f，定义经验过程为：

$$G_n f = n^{1/2}(\mathbb{P}_n f - Pf)$$

其中：

$$\mathbb{P}_n f = \int f d\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

$$Pf = \int f dP = E[f(X)]$$

Example 4.23 取 $f(x) = I(x \leqslant t)$，则：

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leqslant t) = F_n(t)$$

$$Pf = E[I(X \leqslant t)] = F(t)$$

此时 $G_n f = \sqrt{n}(F_n(t) - F(t))$ 就是经典的经验过程.

§4.3 例子——得分检验

Example 4.24 我们有一个统计模型，数据来自分布 $f(x|\theta_{01}, \theta_{02})$，参数为：

$$\theta = (\theta_1^T, \theta_2^T)^T \in \mathbb{R}^p, \quad \theta_1 \in \mathbb{R}^{p-q}, \quad \theta_2 \in \mathbb{R}^q$$

要检验的假设：

$$H_0: \quad \theta_{02} = \mathbf{0}_{q\times 1}$$
$$H_1: \quad \theta_{02} \neq \mathbf{0}_{q\times 1}$$

Definition 4.25 (得分函数) 得分函数是对数似然函数的梯度：

$$U(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = \begin{pmatrix} U_1(\theta) \\ U_2(\theta) \end{pmatrix} = \begin{pmatrix} \dfrac{\partial \ell(\theta)}{\partial \theta_1} \\ \dfrac{\partial \ell(\theta)}{\partial \theta_2} \end{pmatrix}$$

Definition 4.26 (观测矩阵) 观测矩阵是对数似然函数的二阶导数矩阵：

$$H(\theta) = \frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^T} = \begin{pmatrix} H_{11}(\theta) & H_{12}(\theta) \\ H_{21}(\theta) & H_{22}(\theta) \end{pmatrix}$$

Definition 4.27 (Fisher 信息矩阵) Fisher 信息矩阵是观测矩阵的负期望值:

$$I(\theta) = -\mathbb{E}[H(\theta)] = -\mathbb{E}\left[\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^T}\right]$$

分块形式为:

$$I(\theta) = \begin{pmatrix} I_{11}(\theta) & I_{12}(\theta) \\ I_{21}(\theta) & I_{22}(\theta) \end{pmatrix}$$

分析. 在 H_0 下，$\theta_2 = 0$，我们求解约束方程:

$$U_1(\theta_1, 0) = \left.\frac{\partial \ell(\theta)}{\partial \theta_1}\right|_{\theta_2=0} = 0$$

得到约束估计量 $\tilde{\theta}_1$. 检验统计量基于:

$$U_2(\tilde{\theta}_1, 0) = \left.\frac{\partial \ell(\theta)}{\partial \theta_2}\right|_{\theta_2=0,\theta_1=\tilde{\theta}_1}$$

如果 H_0 为真，在真实参数处，关于 θ_2 的得分应该接近零. 如果 $U_2(\tilde{\theta}_1, 0)$ 显著不为零，就拒绝 H_0. □

Solution. 在 H_0 为真的假设下 $(\theta_{02} = 0)$，对 $U_2(\widetilde{\theta}_1, \mathbf{0})$ 进行泰勒展开:

$$\begin{aligned}
\frac{1}{\sqrt{n}}U_2(\widetilde{\theta}_1, \mathbf{0}) &= \frac{1}{\sqrt{n}}U_2(\theta_{01}, \mathbf{0}) + H_{21}(\overline{\theta}_{01}, \mathbf{0})\sqrt{n}(\widetilde{\theta}_1 - \theta_{01}) \\
&= \frac{1}{\sqrt{n}}U_2(\theta_{01}, \mathbf{0}) + H_{21}(\theta_{01}, \mathbf{0})\sqrt{n}(\widetilde{\theta}_1 - \theta_{01}) + R + o_p(1)
\end{aligned}$$

其中 $R = \left(H_{21}(\overline{\theta}_{01}, \mathbf{0}) - H_{21}(\theta_{01}, \mathbf{0})\right)\sqrt{n}(\widetilde{\theta}_1 - \theta_{01})$，$\overline{\theta}_{01}$ 位于 $\widetilde{\theta}_1$ 与 θ_{01} 的连线上. 具体地:

$$\begin{aligned}
\|R\| &\leqslant K_n\|\overline{\theta}_{01} - \theta_{01}\| \cdot \left\|\sqrt{n}(\widetilde{\theta}_1 - \theta_{01})\right\| \\
&\leqslant K_n \cdot \frac{1}{\sqrt{n}}\left\|\sqrt{n}(\widetilde{\theta}_1 - \theta_{01})\right\|^2 \\
&= O_p(n^{-1/2})
\end{aligned}$$

于是我们有

$$\frac{1}{\sqrt{n}}U_2(\widetilde{\theta}_1, \mathbf{0}) = \frac{1}{\sqrt{n}}U_2(\theta_{01}, \mathbf{0}) + H_{21}(\theta_{01}, \mathbf{0})\sqrt{n}(\widetilde{\theta}_1 - \theta_{01}) + o_p(1) \tag{1}$$

由约束方程 $U_1(\widetilde{\theta}_1, \mathbf{0}) = \mathbf{0}$ 出发:

$$\mathbf{0} = U_1(\widetilde{\theta}_1, \mathbf{0}) = U_1(\theta_{01}, \mathbf{0}) + H_{11}(\theta^{**}, \mathbf{0})(\widetilde{\theta}_1 - \theta_{01}) + o_p(1)$$

解得:

$$\sqrt{n}(\widetilde{\theta}_1 - \theta_{01}) = -H_{11}^{-1}(\theta^{**}, \mathbf{0}) \cdot \frac{1}{\sqrt{n}}U_1(\theta_{01}, \mathbf{0}) + o_p(1) \tag{2}$$

其中 θ^{**} 在 $\widetilde{\theta}_1$ 与 θ_{01} 之间. 将 (2) 式代入 (1) 式:

$$\frac{1}{\sqrt{n}}U_2(\widetilde{\theta}_1,\mathbf{0})=\frac{1}{\sqrt{n}}U_2(\theta_{01},\mathbf{0})-H_{21}(\theta_{01},\mathbf{0})H_{11}^{-1}(\theta^{**},\mathbf{0})\cdot\frac{1}{\sqrt{n}}U_1(\theta_{01},\mathbf{0})+o_p(1)$$

不容易得到 (详细见下面的引理):

$$\frac{1}{\sqrt{n}}U_2(\widetilde{\theta}_1,\mathbf{0})\xrightarrow{d}N\left(\mathbf{0},I_{22}-I_{21}I_{11}^{-1}I_{12}\right)$$

□

Lemma 4.28 在正则条件下:

$$\frac{1}{\sqrt{n}}\begin{pmatrix}U_1(\theta_{01},\mathbf{0})\\U_2(\theta_{01},\mathbf{0})\end{pmatrix}\xrightarrow{d}N\left(\mathbf{0},I(\theta_0)\right)\quad H_{ij}(\theta)\xrightarrow{p}-I_{ij}(\theta_0)$$

Lemma 4.29 若

$$\frac{1}{\sqrt{n}}U_2(\widetilde{\theta}_1,\mathbf{0})=\frac{1}{\sqrt{n}}U_2(\theta_{01},\mathbf{0})-H_{21}(\theta_{01},\mathbf{0})H_{11}^{-1}(\theta^{**},\mathbf{0})\cdot\frac{1}{\sqrt{n}}U_1(\theta_{01},\mathbf{0})+o_p(1)$$

则

$$\frac{1}{\sqrt{n}}U_2(\widetilde{\theta}_1,\mathbf{0})\xrightarrow{d}N\left(\mathbf{0},I_{22}-I_{21}I_{11}^{-1}I_{12}\right)$$

Proof. 我们记

$$A_n=\frac{1}{\sqrt{n}}U_2\left(\theta_{01},\mathbf{0}\right),\quad B_n=\frac{1}{\sqrt{n}}U_1\left(\theta_{01},\mathbf{0}\right),\quad C_n=H_{21}\left(\theta_{01},0\right)H_{11}^{-1}\left(\theta^{**}\right)$$

则原式可写为:

$$\frac{1}{\sqrt{n}}U_2\left(\tilde{\theta}_1,\mathbf{0}\right)=A_n-C_nB_n+o_p(1)$$

由假设:

$$\begin{pmatrix}B_n\\A_n\end{pmatrix}\xrightarrow{d}N\left(\mathbf{0},\begin{pmatrix}I_{11}&I_{12}\\I_{21}&I_{22}\end{pmatrix}\right)$$

因此 B_n 和 A_n 是联合渐近正态的. 先分析 C_n 的渐进形式, 由正则条件:

$$H_{21}\left(\theta_{01},0\right)\xrightarrow{p}-I_{21}\quad H_{11}\left(\theta^{**},0\right)\xrightarrow{p}-I_{11}\quad\left(\theta^{**}\xrightarrow{p}\theta_{01}\right)$$

因此 (利用连续映射定理和 Slustky 引理):

$$C_n=H_{21}\left(\theta_{01},\mathbf{0}\right)H_{11}^{-1}\left(\theta^{**},0\right)\xrightarrow{p}\left(-I_{21}\right)\cdot\left(-I_{11}^{-1}\right)=I_{21}I_{11}^{-1}$$

将 C_n 代入:

$$\frac{1}{\sqrt{n}}U_2\left(\tilde{\theta}_1,\mathbf{0}\right)=A_n-C_nB_n+o_p(1)=A_n-I_{21}I_{11}^{-1}B_n+\left(I_{21}I_{11}^{-1}-C_n\right)B_n+o_p(1)$$

由于 $C_n \xrightarrow{p} I_{21}I_{11}^{-1}$，且 $B_n = O_p(1)$(因为渐近正态)，有：

$$\left(I_{21}I_{11}^{-1} - C_n\right) B_n = o_p(1)$$

因此：

$$\frac{1}{\sqrt{n}} U_2\left(\tilde{\theta}_1, 0\right) = A_n - I_{21}I_{11}^{-1}B_n + o_p(1)$$

令：$Z_n = A_n - I_{21}I_{11}^{-1}B_n$，则 Z_n 是 A_n 和 B_n 的线性组合，且：

$$\frac{1}{\sqrt{n}} U_2\left(\tilde{\theta}_1, \mathbf{0}\right) = Z_n + o_p(1) \xrightarrow{d} Z$$

其中 Z 是 Z_n 的极限分布. □

Definition 4.30 (统计量的构造) 构造二次型：

$$LM = \left[\frac{1}{\sqrt{n}} U_2(\tilde{\theta}_1, 0)\right]^T \left(I_{22} - I_{21}I_{11}^{-1}I_{12}\right)^{-1} \left[\frac{1}{\sqrt{n}} U_2(\tilde{\theta}_1, 0)\right] \xrightarrow{d} \chi^2(q)$$

Remark 4.31 用约束估计量 $\tilde{\theta}$ 处的信息矩阵估计：

$$LM = \left[\frac{1}{\sqrt{n}} U_2(\tilde{\theta}_1, 0)\right]^T \left[I_{22}(\tilde{\theta}) - I_{21}(\tilde{\theta})I_{11}^{-1}(\tilde{\theta})I_{12}(\tilde{\theta})\right]^{-1} \left[\frac{1}{\sqrt{n}} U_2(\tilde{\theta}_1, 0)\right]$$

利用连续映射和 Slutsky 引理可以得到这个估计也依分布收敛于 $\chi^2(q)$，称其为 C.R.Rao 统计量.

进一步当期望信息矩阵 $I(\theta)$ 难以计算时，用观测矩阵：

$$LM = \left[\frac{1}{\sqrt{n}} U_2(\tilde{\theta}_1, 0)\right]^T \left[H_{22}(\tilde{\theta}_1) - H_{21}(\tilde{\theta}_1)H_{11}^{-1}(\tilde{\theta}_1)H_{12}(\tilde{\theta}_1)\right]^{-1} \left[\frac{1}{\sqrt{n}} U_2(\tilde{\theta}_1, 0)\right]$$

代替，由于我们假设了矩阵的随机等度连续性，即

$$H(\tilde{\theta}) - H(\theta_0)) = o_p(1)$$

再次使用连续映射和 Slutsky 引理可以得到结果.