中国科学技术大学
University of Science and Technology of China

数理统计近似定理第二章

2025 年 3 月 17 日

定义与基本性质

- 样本分布函数 $F_n(x)$:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x), \quad -\infty < x < \infty.$$

  - 对每个固定的样本，$F_n(\cdot)$ 是一个分布函数。
  - 对每个固定的 x，$F_n(x)$ 是一个随机变量，服从二项分布 $\text{Binomial}(n, F(x))$。

基本性质:

- 无偏性: $E[F_n(x)] = F(x)$。
- 强相合性: $F_n(x) \xrightarrow{wp1} F(x)$。
- 渐近正态性: 对每个固定的 x, $F_n(x)$ 服从渐近正态分布:

$$F_n(x) \text{ 是} AN\left(F(x), \frac{F(x)(1-F(x))}{n}\right).$$

- 定义：

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|.$$

  - 用于衡量 F_n 与 F 的全局差异。
  - 在假设检验中用于检验 $H_0 : F = F_0$。
- 概率不等式：
  - Dvoretzky-Kiefer-Wolfowitz 不等式：对任意 $d > 0$，

$$P(D_n > d) \leq Ce^{-2nd^2},$$

其中 C 为常数。

- 几乎必然收敛：
  - Glivenko-Cantelli 定理： $D_n \xrightarrow{wp1} 0$。
  - 重对数律（LIL）：

$$\limsup_{n\to\infty} \frac{\sqrt{n}D_n}{\sqrt{2\log\log n}} = \sup_x \sqrt{F(x)(1-F(x))} \quad \text{wp1}.$$

- 渐近分布：
  - Kolmogorov 定理： 若 F 连续，

$$\lim_{n\to\infty} P(\sqrt{n}D_n \le d) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j+1}e^{-2j^2d^2}.$$

- 定义:

$$C_n = n\int_{-\infty}^{\infty}(F_n(x) - F(x))^2 dF(x).$$

  - 用于衡量 F_n 与 F 的加权平方差异。
- 渐近行为:
  - 几乎必然收敛: Finkelstein 定理指出,

$$\lim_{n\to\infty}\frac{C_n}{2\log\log n} = \frac{1}{\pi^2} \quad \text{wp1}.$$

  - 渐近分布: C_n 收敛于独立卡方变量的加权和:

$$C_n \xrightarrow{d} \sum_{j=1}^{\infty}\frac{\chi^2_{1,j}}{j^2\pi^2}.$$

核估计方法:

- 简单差分估计:

$$f_n(x) = \frac{F_n(x+b_n) - F_n(x-b_n)}{2b_n},$$

需选择带宽 $b_n \to 0$ 且 $nb_n \to \infty$。

- 核平滑估计:

$$f_n(x) = \frac{1}{nb_n}\sum_{i=1}^{n} W\left(\frac{x-X_i}{b_n}\right),$$

其中 W 为核函数。

渐近性质:

- 若 f 在 x 处连续且 b_n 选择适当, 则 $f_n(x)$ 是渐近正态的。

- 样本分布函数 $F_n(x)$ 是推断总体分布 $F(x)$ 的基础工具，其全局性质（如 D_n 和 C_n）为分布检验提供理论支持。
- Kolmogorov-Smirnov 统计量通过极值分布刻画经验过程的波动，Cramér-von Mises 统计量通过积分形式衡量差异。
- 样本密度估计需平衡偏差与方差，核方法在光滑性假设下具有渐近最优性。

- **总体矩**:
  - 原点矩: $\alpha_k = E[X^k]$
  - 中心矩: $\mu_k = E[(X-\mu)^k]$
- **样本矩**:
  - 样本原点矩: $a_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$
  - 样本中心矩: $m_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$

- 强相合性:
  - $a_k \xrightarrow{wp1} \alpha_k$, $m_k \xrightarrow{wp1} \mu_k$ (需假设 $\alpha_{2k} < \infty$)。
- 渐近正态性:
  - 样本原点矩向量: 若 $\alpha_{2k} < \infty$, 则:

$$\sqrt{n}\left(a_1 - \alpha_1, \ldots, a_k - \alpha_k\right) \xrightarrow{d} N(0, \Sigma),$$

    协方差矩阵 $\Sigma = [\sigma_{ij}]$ 满足 $\sigma_{ij} = \alpha_{i+j} - \alpha_i\alpha_j$。
  - 样本中心矩向量: 若 $\mu_{2k} < \infty$, 则:

$$\sqrt{n}\left(m_2 - \mu_2, \ldots, m_k - \mu_k\right) \xrightarrow{d} N(0, \Sigma^*),$$

- **偏差修正**:
  - 样本方差: $S^2 = \frac{n}{n-1}s^2$（无偏）
  - 三阶矩: $M_3 = \frac{n^2}{(n-1)(n-2)}m_3$
- **效率对比**:
  - 矩估计 vs MLE: 参数模型中 MLE 更有效
  - 矩估计 vs 分位数估计: 后者对异常值稳健
- **工具定理**:
  - Hoeffding 不等式: $P(|a_k - \alpha_k| > t) \leq 2e^{-2nt^2}$
  - Berry-Esseen 定理: 收敛速率 $O(n^{-1/2})$

核心结论:
无偏修正提升小样本性能，矩方法简单但非参数场景效率受限。

- **渐近联合正态性**：

$$\sqrt{n}\left(\overline{X}-\mu, s^2-\sigma^2\right) \xrightarrow{d} N\left(0, \begin{bmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4-\sigma^4 \end{bmatrix}\right)$$

- 特别地，$\overline{X}$ 和 s^2 渐近独立当且仅当 $\mu_3=0$（如对称分布）。

- **矩估计的效率**:
  - 矩估计量通常是相合的，但在参数族中可能不如极大似然估计（MLE）有效。
  - 若总体分布尾部较重（如柯西分布），高阶矩可能不存在，此时矩方法失效。
- **与分位数估计的对比**:
  - 分位数估计在存在异常值时更稳健，但需分布光滑性假设（如密度连续）。

- **Hoeffding 不等式**：

$$P(|a_k - \alpha_k| > t) \leq Ce^{-cnt^2}$$

- **Berry-Esseen 定理**：

样本均值收敛到正态分布的速率： $O(n^{-1/2})$

- **辅助引理**：

$$E\left|\sum Z_i\right|^{\nu} = O(n^{\nu/2})$$

核心结论

1. 样本矩的渐近正态性：样本原点矩和中心矩在有限矩条件下服从多元正态分布，协方差由高阶矩决定。
2. 无偏修正的必要性：样本方差等中心矩需调整以消除偏差，但高阶修正对渐近性质影响较小。
3. 效率权衡：矩估计简单通用，但在参数模型或重尾分布中可能不如其他方法（如 MLE 或分位数估计）。
4. 工具支持：Hoeffding 不等式和 Berry-Esseen 定理为样本矩的收敛速度和概率控制提供理论基础。

- **定义**:
  - 总体分位数: $\xi_p = \inf\{x : F(x) \geq p\}$
  - 样本分位数: $\xi_{pn} = F_n^{-1}(p)$
- **强相合性**:

$$\xi_{pn} \xrightarrow{\text{wp1}} \xi_p \quad (\text{若}\xi_p\text{唯一})$$

- **指数型概率不等式**:

$$P(|\xi_{pn} - \xi_p| > \epsilon) \leq 2e^{-2n\delta_\epsilon^2}$$

$$\delta_\epsilon = \min\{F(\xi_p + \epsilon) - p, p - F(\xi_p - \epsilon)\}$$

核心结论:
样本分位数强收敛于总体分位数，指数不等式控制偏离概率。

- **渐近正态性**:

$$\sqrt{n}(\xi_{pn} - \xi_p) \xrightarrow{d} N\left(0, \frac{p(1-p)}{f^2(\xi_p)}\right)$$

要求 F 在 ξ_p 可导且 $f(\xi_p) > 0$

- **Bahadur 表示**:

$$\xi_{pn} = \xi_p + \frac{p - F_n(\xi_p)}{f(\xi_p)} + R_n$$

余项 $R_n = O(n^{-3/4}(\log n)^{3/4})$（几乎处处成立）

- **重对数律**:

$$\limsup_{n\to\infty} \frac{\sqrt{n}(\xi_{pn} - \xi_p)}{\sqrt{2\log\log n}} = \frac{\sqrt{p(1-p)}}{f(\xi_p)} \quad \text{wp1}$$

核心结论:
样本分位数渐近正态，Bahadur 表示揭示其线性结构。

- **顺序统计量表示**:

$$\xi_{pn} = \begin{cases} X_{n,\lfloor np \rfloor+1} & (np\text{非整数}) \\ X_{n,np} & (np\text{整数}) \end{cases}$$

- **渐近等价性**:

$$\sqrt{n}(X_{n,k_n} - \xi_{pn}) \xrightarrow{d} N\left(0, \frac{p(1-p)}{f^2(\xi_p)}\right)$$

要求 $k_n/n \to p$ 且 $k_n - np = O(\sqrt{n})$

- **密度函数**:

$$g_n(t) = n\binom{n-1}{\lfloor np \rfloor} F^{\lfloor np \rfloor}(t)(1-F(t))^{n-\lfloor np \rfloor-1} f(t)$$

核心结论:
样本分位数与顺序统计量渐近等价，密度函数收敛于正态分布。

- **稳健性**：
  - 对异常值不敏感
  - 适用于重尾分布（如柯西分布）
- **效率对比**：
  - 正态分布：$\overline{X}$ 比 $\xi_{0.5,n}$ 更有效（相对效率 $2/\pi$）
  - 柯西分布：$\xi_{0.5,n}$ 比 $\overline{X}$ 更有效
- **应用场景**：
  - 稳健估计
  - 非参数推断

- **定义**:
  - 顺序统计量: $X_{n1} \leq X_{n2} \leq \cdots \leq X_{nn}$
  - 极值统计量: 最小值 X_{n1}, 最大值 X_{nn}
- **分布函数**:

$$F_{X_{nk}}(x) = \sum_{j=k}^{n} \binom{n}{j} F^j(x)(1 - F(x))^{n-j}$$

- **密度函数**:

$$f_{X_{nk}}(x) = n\binom{n-1}{k-1} F^{k-1}(x)(1 - F(x))^{n-k} f(x)$$

核心结论:
顺序统计量的分布由总体分布 F 决定，极值统计量反映尾部行为。

- **中心顺序统计量**：

$$\sqrt{n}(X_{nk_n} - \xi_p) \xrightarrow{d} N\left(0, \frac{p(1-p)}{f^2(\xi_p)}\right)$$

要求 $k_n/n \to p$ 且 $0 < p < 1$

- **极值统计量**：最大值 X_{nn}：

$$\frac{X_{nn} - a_n}{b_n} \xrightarrow{d} G$$

G 为极值分布（Gumbel、Fréchet 或 Weibull）

核心结论：
中心顺序统计量渐近正态，极值统计量收敛于极值分布。

- **极值分析**：
  - 洪水、地震等极端事件建模
  - 可靠性分析：最小寿命、最大失效时间
- **分布自由推断**：
  - 置信区间：利用 X_{nk_1}, X_{nk_2} 构造
  - 假设检验：非参数方法
- **样本中位数**：

$$\sqrt{n}(X_{n,\lfloor n/2 \rfloor} - \xi_{0.5}) \xrightarrow{d} N\left(0, \frac{1}{4f^2(\xi_{0.5})}\right)$$

核心结论：
顺序统计量在极值分析、可靠性研究和非参数推断中应用广泛。

- **表示形式**:

$$\xi_{pn} = \xi_p + \frac{p - F_n(\xi_p)}{f(\xi_p)} + R_n$$

余项 $R_n = O(n^{-3/4}(\log n)^{3/4})$（几乎处处成立）

- **重对数律**:

$$\limsup_{n\to\infty} \frac{\sqrt{n}(\xi_{pn} - \xi_p)}{\sqrt{2\log\log n}} = \frac{\sqrt{p(1-p)}}{f(\xi_p)} \quad \text{wp1}$$

- **渐近正态性**:

$$\sqrt{n}(\xi_{pn} - \xi_p) \xrightarrow{d} N\left(0, \frac{p(1-p)}{f^2(\xi_p)}\right)$$

核心结论:
Bahadur 表示将样本分位数分解为线性项和高阶余项，导出渐近正态性。

- **表示形式**:

$$X_{nk_n} = \xi_p + \frac{k_n/n - F_n(\xi_p)}{f(\xi_p)} + \tilde{R}_n$$

余项 $\tilde{R}_n = O(n^{-3/4}(\log n)^{3/4})$（几乎处处成立）

- **渐近等价性**:

$$\sqrt{n}(X_{nk_n} - \xi_{pn}) \xrightarrow{d} N\left(0, \frac{p(1-p)}{f^2(\xi_p)}\right)$$

要求 $k_n/n \to p$ 且 $k_n - np = O(\sqrt{n})$

- **联合表示**:

$$\xi_{p_j n} = \xi_{p_j} + \frac{p_j - F_n(\xi_{p_j})}{f(\xi_{p_j})} + R_{jn}, \quad j = 1, \ldots, k$$

核心结论:
顺序统计量与样本分位数渐近等价，联合表示支持多分位数分析。

- **Kiefer 定理**:

$$R_n = O(n^{-3/4}(\log\log n)^{3/4}) \quad \text{wp1}$$

余项具有非退化极限分布

- **Duttweiler 定理**:

$$E[(n^{3/4}R_n)^2] = \left(\frac{2p(1-p)}{\pi}\right)^{1/2} + o(n^{-1/4+\epsilon})$$

- **极限分布**:

$$n^{3/4}R_n \xrightarrow{d} \text{非退化分布}$$

核心结论:
余项的高阶性质由 Kiefer 和 Duttweiler 定理精确刻画。

- **置信区间**:

$$\left(\xi_{pn} \pm \frac{z_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}f(\xi_p)}\right)$$

- **稳健估计**:
  - 对异常值不敏感
  - 适用于重尾分布
- **高维扩展**:
  - 多维分位数的 Bahadur 表示
  - 联合渐近正态性

核心结论:
Bahadur 表示为分位数推断、稳健估计和高维扩展提供理论基础。

- **构造方法**：
  - 选择顺序统计量 X_{nk_1}, X_{nk_2}
  - 置信区间：(X_{nk_1}, X_{nk_2})
- **置信系数**：

$$P(X_{nk_1} < \xi_p < X_{nk_2}) = P(U_{nk_1} < p < U_{nk_2})$$

$U_{n1} \leq \cdots \leq U_{nn}$ 为均匀分布顺序统计量

- **计算**：

$$P(U_{nk_1} < p < U_{nk_2}) = I_p(k_1, n - k_1 + 1) - I_p(k_2, n - k_2 + 1)$$

I_p 为不完全 Beta 函数

核心结论：
精确方法分布自由，但计算复杂，需查表或数值积分。

- **基于样本分位数**：

$$\left(\xi_{pn} \pm \frac{z_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}f(\xi_p)}\right)$$

需已知 $f(\xi_p)$，可用核密度估计替代

- **基于顺序统计量**：
  - 选择 k_{1n}, k_{2n} 使得：

$$\frac{k_{1n}}{n} = p - \frac{z_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}, \qquad \frac{k_{2n}}{n} = p + \frac{z_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}$$

  - 置信区间：$(X_{nk_{1n}}, X_{nk_{2n}})$

核心结论：
渐近方法基于样本分位数或顺序统计量，后者分布自由且无需估计密度。

- **适用范围**：仅适用于中位数 $\xi_{0.5}$
- **构造方法**：
  - 利用 Wilcoxon 符号秩统计量
  - 置信区间：(W_{na_n}, W_{nb_n})
- **渐近性质**：

$$\sqrt{n}(W_{na_n} - \xi_{0.5}) \xrightarrow{d} N\left(0, \frac{1}{12(\int f^2(x)dx)^2}\right)$$

核心结论：
Wilcoxon 方法适用于中位数，效率依赖于密度函数的积分。

- **适用范围**：仅适用于对称分布的中位数 $\xi_{0.5}$
- **构造方法**：

$$\left(\overline{X} \pm \frac{z_{\alpha/2}s}{\sqrt{n}}\right)$$

$\overline{X}$ 为样本均值，s 为样本标准差

- **渐近性质**：

$$\sqrt{n}(\overline{X} - \xi_{0.5}) \xrightarrow{d} N(0, \sigma^2)$$

核心结论：
样本均值方法仅适用于对称分布，效率依赖于方差。

- **定义**：

$$Y_n(x) = \sqrt{n}(F_n(x) - F(x)), \quad x \in \mathbb{R}$$

标准化形式：$Y_n(t) = \sqrt{n}(F_n(F^{-1}(t)) - t), \quad t \in [0,1]$

- **渐近性质**：

$$Y_n \xrightarrow{d} W^0 \quad \text{在 } D[0,1] \text{ 空间中}$$

W^0 为布朗桥过程，协方差：$\mathrm{Cov}(W^0(s), W^0(t)) = s(1-t)$

- **应用**：

$$\sqrt{n}D_n = \sup_{t\in[0,1]} |Y_n(t)| \xrightarrow{d} \sup_{t\in[0,1]} |W^0(t)|$$

核心结论：
经验过程收敛于布朗桥，为 Kolmogorov-Smirnov 检验提供理论基础。

- **定义**：

$$Z_n(p) = \sqrt{n}(\xi_{pn} - \xi_p), \quad p \in (0,1)$$

标准化形式：$Z_n(p) = \sqrt{n}(F_n^{-1}(p) - F^{-1}(p))$

- **渐近性质**：

$$Z_n(p) \approx -\frac{Y_n(p)}{f(\xi_p)}$$

若 F 在 ξ_p 可导且 $f(\xi_p) > 0$，则：

$$Z_n \xrightarrow{d} -\frac{W^0}{f(F^{-1}(\cdot))}$$

核心结论：
样本分位数过程与经验过程通过密度函数关联，渐近收敛于加权布朗桥。

- **极值过程**：

$$Q_{kn}(t) = \frac{X_{\lfloor nt \rfloor, k} - a_n}{b_n}, \quad t \in [0, 1]$$

渐近性质：收敛于极值分布（如 Gumbel 过程）

- **间隔过程**：

$$S_n(t) = \sqrt{n}(F_n(t) - F(t)), \quad t \in \mathbb{R}$$

渐近性质：收敛于布朗桥过程

核心结论：
极值过程和间隔过程的渐近性质为极端事件和分布的推断提供支持。

- **假设检验**:
  - Kolmogorov-Smirnov 检验
  - Cramér-von Mises 检验
- **置信带构造**:
  - 利用经验过程的极限分布
  - 构造样本分布函数的置信带
- **高维扩展**:
  - 多维经验过程
  - 多维分位数过程

核心结论:
随机过程理论为假设检验、置信带构造和高维扩展提供统一框架。