4: 大数定律和中心极限定理

张伟平

第四章大数定律和中心极限定理

极限定理是概率论的重要内容, 也是数理统计学的基石之一. 大数定律, 是概率论中讨论随机变量和的平均值的收敛情况, 是数理统计学中参数估计的理论基础. **中心极限定理**, 是概率论中讨论随机变量和的分布以正态分布为极限的一组定理, 这组定理是数理统计学和误差分析的理论基础, 指出了大量随机变量近似服从正态分布的条件.

4.1 大数定律

Definition

如果对任何 $\varepsilon > 0$，都有

$$\lim_{n\to\infty} P(|\xi_n - \xi| \geq \varepsilon) = 0,$$

那么我们就称随机变量序列 $\{\xi_n,\ n \in \mathbb{N}\}$ 依概率收敛到随机变量 ξ，记为 $\xi_n \xrightarrow{p} \xi$.

定理 1. 设 $\{X_n\}$ 是一列独立同分布 *(i.i.d.)* 的随机变量序列，具有公共的数学期望 μ 和方差 σ^2. 则

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{p} \mu,$$

即 $\{X_n\}$ 服从 *(弱)* 大数定律。

[**注**]: 实际上，我们只需要均值存在即有大数定律成立，上述定理中加上了方差存在的条件，只是为了证明的方便。

作为上述定理的一个特例，我们有

如果以 ζ_n 表示 n 重 Bernoulli 试验中的成功次数, 则有

$$\frac{\zeta_n}{n} \xrightarrow{p} p.$$

↑Example

↓Example

如果用 $f_n = \zeta_n/n$ 表示成功出现的频率, 则上例说明 $f_n \xrightarrow{p} p$, 即频率 (依概率) 收敛到概率.

为证明定理1, 我们需要如下的 Chebyshev 不等式:

引理 1 (**Chebyshev 不等式**). *设随机变量 X 的方差存在, 则*

$$P(|X - EX| \geq \varepsilon) \leq \frac{Var(X)}{\varepsilon^2}, \quad \forall\ \varepsilon > 0.$$

我们可以用 Chebyshev 不等式来估计 X 与 EX 的偏差, 但是 Chebyshev 不等式作为一个理论工具比作为估计的实际方法要恰当

一些, 其重要性在于它的应用普遍性, 但是不能希望很普通的命题对一些个别情况给了深刻的结果. 如令 X 为掷一个均匀的骰子所得到的点数, 则 $\mu = EX = 7/2$, $\sigma^2 = \mathrm{Var}(X) = 35/12$. X 与 μ 的最大偏差为 $2.5 \approx 3\sigma/2$. $|X - \mu|$ 大于这个偏差的概率为 0, 然而利用 Chebyshev 不等式仅仅断定这个概率少于 0.47. 这时就需要找更精确的估计.

定理1的证明. 利用 Chebyshev 不等式, 并注意到 $E\overline{X} = \mu$, $\mathrm{Var}\overline{X} = \sigma^2/n$, 我们有,

$$P(|\overline{X} - \mu| \geq \varepsilon) \leq \sigma^2/(n\varepsilon^2) \to 0, \quad n \to \infty \quad \forall \varepsilon > 0.$$

定理得证.

4.2 中心极限定理

中心极限定理是概率论中讨论随机变量序列的分布收敛于正态分布的一类定理. 它是概率论中最重要的一类定理, 有广泛的实际应用背景.

定理 2. 设 $\{X_n\}$ 为 $i.i.d$ 的随机变量序列, 具有公共的数学期望 μ 和方差 σ^2. 则 $X_1+\cdots+X_n$ 的标准化形式 $\frac{1}{\sqrt{n}\sigma}(X_1+\cdots+X_n-n\mu)$ 满足中心极限定理. 即对任意 $x\in\mathbb{R}$, 有

$$\lim_{n\to\infty} F_n(x)=\Phi(x),$$

其中 $F_n(x)$ 为 $\frac{1}{\sqrt{n}\sigma}(X_1+\cdots+X_n-n\mu)$ 的分布函数, 而 $\Phi(x)$ 为标准正态分布 $N(0,1)$ 的分布函数. 记为

$$\frac{1}{\sqrt{n}\sigma}(X_1+\cdots+X_n-n\mu)\xrightarrow{d}N(0,1).$$

定理2的令人吃惊之处就是任何独立同分布的随机变量序列, 不论它的分布是什么, 只要存在有限的方差, 那么它们的标准化部分和都渐近于标准正态分布. 这也说明了正态分布的普遍性.

由定理2, 我们很容易得到如下推论

定理 3. 设 $X_1, \cdots, X_n$ 相互独立且具有相同的分布

$$P(X_1 = 1) = 1 - P(X_1 = 0) = p, \ 0 < p < 1.$$

则有

$$\frac{X_1 + \cdots + X_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} N(0,1).$$

即

$$\lim_{n\to\infty} P\Big(\frac{X_1 + \cdots + X_n - np}{\sqrt{np(1-p)}} \le x\Big) = \Phi(x), \quad \forall\ x \in \mathbb{R}.$$

定理2称为棣莫弗 -拉普拉斯定理, 是历史上最早的中心极限定理.

因为定理2中随机变量 $X_1, \cdots, X_n$ 的和 $X_1 + \cdots + X_n \sim B(n, p)$, 我们利用正态分布近似地估计二项分布.

设 $t_1 < t_2$ 是两个正整数, 则当 n 相当大时, 由定理2, 近似地有

$$P(t_1 \leq X_1 + \cdots + X_n \leq t_2) \approx \Phi(y_2) - \Phi(y_1),$$

其中

$$y_i = (t_i - np)/\sqrt{np(1-p)}, \ i = 1, 2.$$

为提高精度, 我们可把 y_1, y_2 修正为

$$y_1 = (t_1 - 1/2 - np)/\sqrt{np(1-p)}, \quad y_2 = (t_2 + 1/2 - np)/\sqrt{np(1-p)}.$$

设一考生参加 100 道题的英语标准化考试 (每道题均为有两个备选答案的选择题，有且仅有一个答案是正确的)，每道题他都随机地选择一个答案，假设评分标准为：选对得一分，选错或不选不得分。试给出该考生最终得分大于等于 50 的概率.

↑Example

↓Example

每天有 1000 个旅客需要乘坐火车从芝加哥到洛杉矶, 这两个城市之间有两条竞争的铁路, 它们的火车同时开出同时到达并且具有同样的设备. 设这 1000 个人乘坐那一条铁路的火车是相互独立而且又是任意的, 于是每列火车的乘客数目可视为概率为 1/2 的 1000 重 Bernoulli 试验中成功的次数. 如果一列火车设置 $s < n$ 个座位, 那么一旦有多于 s 个旅客来乘车就容纳不下了, 令这个事件发生的概率为 $f(s)$. 利用中心极限定理, 有

$$f(s) \approx 1 - \Phi\Big(\frac{2s - 1000}{\sqrt{1000}}\Big).$$

要求 s 使得 $f(s) < 0.01$, 即在 100 次中有 99 次是有足够的座位的. 查表容易求出 $s = 537$. 这样, 两列火车所有的座位数为 1074, 其中只有 74 个空位, 可见由于竞争而带来的损失是很小的.

↑Example

↓Example

求极限 $\lim\limits_{n\to\infty}\sum\limits_{k=1}^{n}\frac{n^k}{k!}e^{-n}$.

↑Example
↓Example

定理 4. 设 $X\sim B(n,p)$，则有

$$\lim_{n\to\infty}P\left(\frac{X-np}{\sqrt{npq}}\le x\right)=\Phi(x),\quad \forall\ x\in\mathbb{R}$$

即

$$\frac{X-np}{\sqrt{npq}}\overset{asy.}{\sim}N(0,1).$$

Proof. 由二项分布随机变量和 0-1 分布随机变量之间的关系及中心极限定理易证。 □

在仅有独立性和二阶矩有限场合下，我们有

定理 5. 设 $\{X_n\}$ 为独立的随机变量序列，而且具有数学期望 $EX_k = \mu_k$ 和方差 $D(X_k) = \sigma_k^2 < \infty$, $k = 1, 2, \cdots$. 记

$$B_n^2 = \sum_{k=1}^{n} \sigma_k^2$$

若存在正数 δ，使得当 $n \to \infty$ 时

$$\frac{1}{B_n^{2+\delta}} \sum_{k=1}^{n} E|X_k - EX_k|^{2+\delta} \to 0$$

则有

$$\lim_{n\to\infty} P\left(\sum_{k=1}^{n} \frac{X_k - \mu_k}{B_n} \le x\right) = \Phi(x) \qquad \forall x \in \mathbb{R} \qquad (4.1)$$

例题参考课本.

如果独立随机变量序列 $\{X_n,\ n \in \mathbb{N}\}$ 同上述定理，并且对任何 $\tau > 0$，都有

$$\lim_{n\to\infty} \frac{1}{B_n^2} \sum_{k=1}^{n} E\left\{(X_k - a_k)^2 I(|X_k - a_k| \geq \tau B_n)\right\} = 0, \tag{4.2}$$

则称该随机变量序列满足 **Linderberg** 条件.

Definition

定理 6. 设随机变量序列 $\{X_n\}$ 满足 *Linderberg* 条件 *(4.2)*, 则 $\{X_n\}$ 满足中心极限定理, 即 *(4.1)* 式成立.