由于 $i,j\in[n]$, 因此共有 $n(n-1)$ 种不同的 $i\neq j$, 根据布尔不等式有

$$P\left[\exists i\neq j\colon \|\boldsymbol{y}_i-\boldsymbol{y}_j\|_2^2\geqslant(1+\epsilon)\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2 \text{ 或 } \|\boldsymbol{y}_i-\boldsymbol{y}_j\|_2^2\leqslant(1-\epsilon)\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\right]\leqslant 2n^2e^{-k(\epsilon^2-\epsilon^3)/4},$$

设 $2n^2e^{-k(\epsilon^2-\epsilon^3)/4}\leqslant 1/2$, 求解 $k\geqslant 8\log 2n/(\epsilon^2-\epsilon^3)$, 引理得证.

7.5 大数定律

给定一个随机变量序列 $X_1,X_2,\cdots,X_n,\cdots$, 大数定律考虑随机变量均值 $\sum_{i=1}^n X_i/n$ 是否随着 n 的增加而趋于一个稳定点, 即随机变量序列的稳定性.

定义 7.1 (依概率收敛) 给定随机变量序列 $X_1,X_2,\cdots,X_n,\cdots$ 和常数 a, 若对任意 $\epsilon>0$ 有

$$\lim_{n\to\infty}P\left(|X_n-a|<\epsilon\right)=1 \quad 或 \quad \lim_{n\to\infty}P\left(|X_n-a|>\epsilon\right)=0,$$

则称随机变量序列 $X_1,X_2,\cdots,X_n,\cdots$ 依概率收敛于 a, 记为 $X_n\xrightarrow{P}a$.

可以发现随机变量序列收敛于稳定点与数列极限的收敛有本质的不同. 下面给出依概率的一些性质:

- 若 $X_n\xrightarrow{P}a$ 且函数 $g(x)\colon\mathbb{R}\to\mathbb{R}$ 在 $x=a$ 点连续, 则 $g(X_n)\xrightarrow{P}g(a)$;
- 若 $X_n\xrightarrow{P}a$ 和 $Y_n\xrightarrow{P}b$, 以及函数 $g(x,y)\colon\mathbb{R}\times\mathbb{R}\to\mathbb{R}$ 在 $x=a$ 和 $y=b$ 处连续, 则有 $g(X_n,Y_n)\xrightarrow{P}g(a,b)$. 例如, 若 $X_n\xrightarrow{P}a$ 和 $Y_n\xrightarrow{P}b$, 则有 $X_n+Y_n\xrightarrow{P}a+b$ 和 $X_nY_n\xrightarrow{P}ab$.

定理 7.13 (大数定律) 若随机变量序列 $X_1,X_2,\cdots,X_n,\cdots$ 满足

$$\frac{1}{n}\sum_{i=1}^n X_i\xrightarrow{P}\frac{1}{n}\sum_{i=1}^n E[X_i],$$

则称 $\{X_n\}_{n\geqslant 1}$ 服从大数定律.

根据依概率收敛的定义, 可以给出大数定律的等价条件为

$$\lim_{n\to+\infty}P\left(\left|\frac{1}{n}\sum_{i=1}^n X_i-\frac{1}{n}\sum_{i=1}^n E[X_i]\right|>\epsilon\right)=0 \text{ 或 } \lim_{n\to+\infty}P\left(\left|\frac{1}{n}\sum_{i=1}^n X_i-\frac{1}{n}\sum_{i=1}^n E[X_i]\right|<\epsilon\right)=1.$$

大数定理刻画了随机变量的均值依概率收敛于期望的均值 (算术平均值). 下面介绍几种大数定律:

定理 7.14 (马尔可夫 Markov 大数定律) 若随机变量序列 $X_1,X_2,\cdots,X_n,\cdots$ 满足

$$\lim_{n\to+\infty}\frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^n X_i\right)=0,$$

则 $\{X_n\}_{n\geqslant 1}$ 服从大数定理.

马尔可夫大数定律根据 Chebyshev 不等式直接可得, 即

$$\lim_{n\to+\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - E[X_i])\right| \geqslant \epsilon\right) \leqslant \lim_{n\to+\infty} \frac{1}{n^2\epsilon^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = 0.$$

定理 7.15 (切比雪夫 Chebyshev 大数定律) 若随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立, 且存在常数 $c > 0$ 使得 $\mathrm{Var}(X_n) \leqslant c$, 则 $\{X_n\}_{n\geqslant 1}$ 服从大数定律.

此处随机变量的独立性可以修改为'不相关性', 其证明直接也是通过切比雪夫不等式, 即

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - E[X_i])\right| \geqslant \epsilon\right) \leqslant \frac{1}{\epsilon^2 n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \leqslant \frac{c}{n\epsilon^2} \to 0 \qquad (n \to +\infty).$$

定理 7.16 (辛钦 Khintchine 大数定律) 若 $X_1, X_2, \cdots, X_n, \cdots$ 为独立同分布的随机变量, 且每个随机变量的期望 $E[X_i] = \mu$ 存在, 则 $\{X_n\}_{n\geqslant 1}$ 服从大数定律.

辛钦大数定律不要求方差一定存在, 其证明超出了本书范围.

定理 7.17 (Bernoulli大数定律) 设随机变量序列 $X_n \sim B(n, p)$ $(p > 0)$, 对任意 $\epsilon > 0$ 有

$$\lim_{n\to+\infty} P\left(\left|\frac{X_n}{n} - p\right| \geqslant \epsilon\right) = 0, \text{ 即 } X_n/n \xrightarrow{P} p.$$

根据二项分布的性质有 $E[X_n] = np$ 和 $\mathrm{Var}(X_n) = np(1-p)$, 利用 Chebyshev 不等式有

$$\lim_{n\to+\infty} P\left(\left|\frac{X_n}{n} - p\right| \geqslant \epsilon\right) = \lim_{n\to+\infty} P\left(|X_n - np| \geqslant n\epsilon\right) \leqslant \lim_{n\to+\infty} \frac{\mathrm{Var}(X_n)}{\epsilon^2 n^2} = \lim_{n\to+\infty} \frac{p(1-p)}{\epsilon^2 n} \to 0.$$

这里随机变量 $X_n \sim \mathrm{B}(n, p)$ 可以看作独立同分布的随机变量 $Y_1, Y_2, \ldots, Y_n$ 之和, 其中 $Y_i \sim \mathrm{Ber}(p)$.

如何判断随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 满足大数定律:

- 若随机变量独立同分布, 则利用辛钦大数定律查看期望是否存在;
- 对非独立同分布随机变量, 则利用 Markov 大数定律判断方差是否趋于零.

例 7.3 独立的随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 满足 $P(X_n = n^{1/4}) = P(X_n = -n^{1/4}) = 1/2$, 则 $\{X_n\}$ 服从大数定律.

证明 根据题意可得 $E[X_i] = 0$, 以及 $\mathrm{Var}(X_i) = E[X_i^2] = i^{1/2}$, 根据 Chebysheve 不等式和随机变量的独立性有

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geqslant \epsilon\right) \leqslant \frac{1}{n^2\epsilon^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2\epsilon^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{\epsilon^2}\frac{1}{n^2}\sum_{i=1}^{n} i^{1/2} \leqslant \frac{1}{\epsilon^2\sqrt{n}},$$

由此可得 $\lim_{n\to+\infty} P(|\sum_{i=1}^{n} X_i/n| \geqslant \epsilon) = 0$.

7.6 中心极限定理

如何刻画随机变量列 $Y_1, Y_2, \cdots, Y_n, \cdots$ 收敛于随机变量 Y, 可以考虑依分布收敛, 其定义如下:

定义 7.2 设随机变量列 $Y_1, Y_2, \cdots, Y_n, \cdots$ 的分布函数分别为 $F_{Y_n}(y) = P(Y_n \leqslant y)$, 以及随机变量 Y 的分布函数为 $F_Y(y) = P(Y \leqslant y)$, 若

$$\lim_{n\to\infty} P(Y_n \leqslant y) = P(Y \leqslant y), \quad \text{即} \quad \lim_{n\to\infty} F_{Y_n}(y) = F_Y(y),$$

则称随机变量序列 $Y_1, Y_2, \cdots, Y_n, \cdots$ 依分布收敛于 Y, 记 $Y_n \xrightarrow{d} Y$.

对独立的随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$, 考虑标准化后随机变量

$$Y_n = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n E(X_i)}{\sqrt{\mathrm{Var}(\sum_{i=1}^n X_i)}}$$

的极限分布是否服从正态分布? 即中心极限定理. 首先介绍独立同分布的中心极限定理, 又称林德贝格-勒维 (Lindeberg-Lévy) 中心极限定理.

定理 7.18 若独立同分布的随机变量 $X_1, X_2, \cdots, X_n, \cdots$ 的期望为 $E(X_1) = \mu$ 和方差为 $\mathrm{Var}(X_1) = \sigma^2$, 则有

$$Y_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1).$$

根据标准正态分布的分布函数 $\Phi(x)$, 上面的中心极限定理等价于 $\lim_{n\to\infty} P(Y_n \leqslant y) = \Phi(y)$. 当 n 足够大时近似有 $Y_n \sim \mathcal{N}(0,1)$, 可近似有

$$\sum_{i=1}^n X_i \approx \mathcal{N}(n\mu, n\sigma^2) \qquad \text{或} \qquad \frac{1}{n}\sum_{i=1}^n X_i \approx \mathcal{N}(\mu, \sigma^2/n).$$

大数定律给出了随机变量 $\sum_{i=1}^n X_i/n$ 的趋势, 而中心极限定理给出了 $\sum_{i=1}^n X_i/n$ 的近似分布.

例 7.4 设一电压接收器同时接收到 20 个独立同分布的信号电压 X_k ($k \in [20]$), 且 $X_k \sim U(0,10)$, 求电压和大于 105 的概率.

解 根据独立同分布的随机变量 $X_1, X_2, \ldots, X_{20}$ 服从均匀分布 $U(0,10)$, 有 $E(X_k) = 5$ 和 $\mathrm{Var}(X_k) = 100/12 = 25/3$. 设 $X = \sum_{k=1}^{20} X_k$, 则有 $E[X] = 100$ 和 $\mathrm{Var}(X) = 500/3$. 根据中心极限定理近似有

$$\frac{X - E(X)}{\sqrt{\mathrm{Var}(X)}} = \frac{X - 100}{\sqrt{500/3}} \sim \mathcal{N}(0,1).$$

根据标准正态分布的分布函数 $\Phi(x)$ 有

$$P(X \geqslant 105) = P\left(\frac{X-100}{\sqrt{500/3}} \geqslant \frac{105-100}{\sqrt{500/3}}\right) = P\left(\frac{X-100}{\sqrt{500/3}} \geqslant 0.387\right) = 1 - \Phi(0.387).$$

查表完成证明.

例 7.5 某产品装箱, 每箱重量是随机的, 假设其期望是 50 公斤, 标准差为 5 公斤. 若最大载重量为 5 吨, 问每车最多可装多少箱能以 0.997 以上的概率保证不超载?

解 假设最多可装 n 箱能以 0.997 以上的概率保证不超载, 用 X_i 表示第 i 箱重量 ($i \in [n]$), 有 $E[X_i] = 50$ 和 $\mathrm{Var}(X_i) = 25$. 设总重量 $X = \sum_{i=1}^{n} X_i$, 则有 $E[X] = 50n$ 和 $\mathrm{Var}(X) = 25n$. 由中心极限定理近似有

$$(X - 50n)/\sqrt{25n} \sim \mathcal{N}(0, 1).$$

根据标准正态分布的分布函数 $\Phi(x)$ 有

$$P(X \leqslant 5000) = P\left(\frac{X - 50n}{\sqrt{25n}} \leqslant \frac{5000 - 50n}{\sqrt{25n}}\right) = \Phi\left(\frac{5000 - 50n}{\sqrt{25n}}\right) > 0.977 = \Phi(2).$$

根据分布函数的单调性有

$$\frac{1000 - 10n}{\sqrt{n}} > 2 \Longrightarrow 100n^2 - 20000n + 1000^2 > 4n.$$

求解可得 $n > 102.02$ 或 $n < 98.02$, 根据由题意可知 $n = 98$.

接下来介绍另一个中心极限定理: 棣莫弗-拉普拉斯 (De Moivre-Laplace) 中心极限定理:

推论 7.6 设随机变量 $X_n \sim B(n, p)$, 则

$$Y_n = \frac{X_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} \mathcal{N}(0, 1).$$

由此可知当 n 非常大时随机变量 $X_n \sim B(n, p)$ 可近似看成 $X_n \sim \mathcal{N}(np, np(1-p))$, 于是有

$$P(X_n \leqslant y) = P\left(\frac{X_n - np}{\sqrt{np(1-p)}} \leqslant \frac{y - np}{\sqrt{np(1-p)}}\right) \approx \Phi\left(\frac{y - np}{\sqrt{np(1-p)}}\right).$$

这里可以考虑三类问题: i) 已知 n 和 $P[X_n \leqslant y]$, 求 y; ii) 已知 n 和 y, 求 $P[X_n \leqslant y]$; iii)已知 y 和 $P[X_n \leqslant y]$, 求 n.

例 7.6 车间有 200 台独立工作的车床, 每台工作的概率为 0.6, 工作时每台耗电 1 千瓦, 至少供电多少千瓦才能以 99.9% 的概率保证正常生产.

解 设工作的车床数为 X, 则 $X \sim B(200, 0.6)$. 设至少供电 y 千瓦. 根据棣莫弗-拉普拉斯中心定理近似有 $X \sim \mathcal{N}(120, 48)$, 进一步有

$$P(X \leqslant y) \geqslant 0.999 \quad \Rightarrow \quad P\left(\frac{X - 120}{\sqrt{48}} \leqslant \frac{y - 120}{\sqrt{48}}\right) \approx \Phi\left(\frac{y - 120}{\sqrt{48}}\right) \geqslant 0.999 = \Phi(3.1).$$

所以有 $(y-120)/\sqrt{48} \geqslant 3.1$, 求解可得 $y \geqslant 141$.

例 7.7 系统由 100 个相互独立的部件组成, 每部件损坏率为 0.1, 至少 85 个部件正常工作系统才能运行, 求系统运行的概率.

解 用 X 是损坏的部件数, 则 $X \sim B(100, 0.1)$, 以及有 $E[X] = 10$ 和 $\mathrm{Var}(X) = 9$. 根据棣莫弗-拉普拉斯中心定理近似有 $X \sim \mathcal{N}(10, 9)$, 求系统运行的概率为

$$P(X \leqslant 15) = P\left(\frac{X-10}{\sqrt{9}} \leqslant \frac{15-10}{\sqrt{9}}\right) \approx \Phi(5/3).$$

例 7.8 一次电视节目调查中调查 n 人, 其中 k 人观看了电视节目, 因此收看比例 k/n 作为电视节目收视率 p 的估计, 要以至少 90% 的概率有 $|k/n - p| \leqslant 0.05$ 成立, 需要调查多少对象?

解 用 X_n 表示 n 个调查对象中收看节目的人数, 则有 $X_n \sim B(n, p)$. 根据棣莫弗-拉普拉斯中心极限定理近似有 $X_n \sim \mathcal{N}(np, np(1-p))$, 进一步有

$$\begin{aligned} P\left(\left|\frac{X_n}{n} - p\right| \leqslant 0.05\right) &= P\left(\frac{|X_n - np|}{n} \leqslant 0.05\right) = P\left(\frac{|X_n - np|}{\sqrt{np(1-p)}} \leqslant \frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right) \\ &= \Phi\left(\frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right) - \Phi\left(-\frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right) \end{aligned}$$

对于标准正太分布函数有 $\Phi(-\alpha) = 1 - \Phi(\alpha)$ 以及 $p(1-p) \leqslant 1/4$, 于是有

$$P\left(\left|\frac{X_n}{n} - p\right| \leqslant 0.05\right) = 2\Phi\left(\frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right) - 1 > 2\Phi\left(\sqrt{n}/10\right) - 1 > 0.9.$$

所以 $\Phi\left(\sqrt{n}/10\right) \geqslant 0.95$, 查表解得 $n \geqslant 271$.

对独立不同分布的随机变量序列, 有李雅普诺夫 (Lyapunov) 中心极限定理:

定理 7.19 设独立随机变量 $X_1, X_2, \ldots, X_k, \ldots$ 的期望 $E[X_k] = \mu_k$ 和方差 $\mathrm{Var}(X_k) = \sigma_k^2 > 0$. 记 $B_n^2 = \sum_{k=1}^n \sigma_k^2$, 若存在 $\delta > 0$, 当 $n \to \infty$ 时有

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

成立, 则有

$$Y_n = \frac{\sum_{k=1}^n X_k - \sum_{k=1}^n E[X_k]}{\sqrt{\mathrm{Var}(\sum_{k=1}^n X_k)}} \xrightarrow{d} \mathcal{N}(0, 1).$$