7 大数定律及中心极限定理

7.1 大数定律

给定n个随机变量$X_1, X_2, \cdots, X_n$, 我们考虑n个随机变量的平均值

$$\frac{1}{n}\sum_{i=1}^{n} X_i.$$

大量随机变量的算术平均值具有稳定性问题.

定义7.1 (依概率收敛). 设$Y_1, Y_2, \cdots, Y_n, \cdots$是随机变量序列, a是一常数, 如果对任意$\epsilon > 0$, 有

$$\lim_{n\to\infty} \Pr\{|Y_n - a| < \epsilon\} = 1 \text{或} \lim_{n\to\infty} \Pr\{|Y_n - a| > \epsilon\} = 0$$

成立, 则称随机变量序列$Y_1, Y_2, \cdots, Y_n, \cdots$依概率收敛于$a$, 记$Y_n \xrightarrow{P} a$.

问题: 与数列极限的区别? 下面我们给出依概率的性质:

1) 若$X_n \xrightarrow{P} a$, 且函数$g\colon \mathbb{R} \to \mathbb{R}$在$X = a$点连续, 则$g(X_n) \xrightarrow{P} g(a)$.
2) 若$X_n \xrightarrow{P} a$, $Y_n \xrightarrow{P} b$, 函数$g\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}$在点$(X, Y) = (a, b)$处连续, 则$g(X_n, Y_n) \xrightarrow{P} g(a, b)$.

例如: 如果$X_n \xrightarrow{P} a$和$Y_n \xrightarrow{P} b$, 那么$X_n + Y_n \xrightarrow{P} a + b$, $X_nY_n \xrightarrow{P} ab$.

定理7.1 (大数定律). 设$X_1, X_2, \cdots, X_n, \cdots$是随机变量序列, 如果有

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

则称$\{X_n\}$服从大数定律.

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

定理7.2 (马尔可夫大数定律). 如果随机变量序列$X_1, X_2, \cdots, X_n, \cdots$满足

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

则$\{X_n\}$服从大数定理.

马尔可夫大数定律不要求随机变量序列$X_1, X_2, \cdots, X_n, \cdots$之间相互独立、或同分布, 其证明直接通过Chebyshev不等式有

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - E[X_i])\right| \geq \epsilon\right] \leq \frac{1}{n^2\epsilon^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \to 0 \qquad n \to \infty.$$

定理7.3 (切比雪夫大数定律)**.** 设随机变量序列$X_1, X_2, \cdots, X_n, \cdots$相互独立, 且存在常数$c > 0$, 使得$Var(X_n) \le c$, 则$\{X_n\}$服从大数定律.

此处独立的随机变量可以修改为“不相关随机变量”. 证明直接通过切比雪夫不等式

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

定理7.4 (辛钦大数定律)**.** 设$X_1, X_2, \cdots, X_n, \cdots$为独立同分布随机变量序列, 且期望存在$E[X_i] = \mu$, 则$\{X_n\}$服从大数定律.

证明超出了本书范围, 一般书中只证明方差存在σ^2.

定理7.5 (Bernoulli大数定律)**.** 设X_n为n重*Bernoulli*试验中事件A发生的次数, 记$p = \Pr(A)$, 即$X_n \sim B(n, p)$. 对任意$\epsilon > 0$有$X_n/n \xrightarrow{P} p$, 即

$$\lim_{n\to\infty} \Pr\left[\left|\frac{X_n}{n} - p\right| \ge \epsilon\right] = 0.$$

例7.1. 设$X_1, X_2, \cdots, X_n, \cdots$是独立的随机变量序列, 且满足$\Pr\{X_n = n^{1/4}\} = \Pr\{X_n = -n^{1/4}\} = 1/2$. 求证: $\{X_n\}$服从大数定律.

Proof. 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不等式和独立性有

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \ge \epsilon\right] \le \frac{1}{n^2\epsilon^2}\mathrm{Var}(\sum_{i=1}^{n} X_i) = \frac{1}{n^2\epsilon^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{\epsilon^2}\frac{1}{n^2}\sum_{i=1}^{n} i^{1/2} \le \frac{1}{\epsilon^2\sqrt{n}}$$

当$n \to 0$时上式趋于零. □

下面补充一个Chebysheve不等式的应用例子:

例7.2. 设随机变量X和Y满足$E(X) = -2$, $E(Y) = 2$, $Var(X) = 1$, $Var(Y) = 4$, $\rho_{XY} = -1/2$. 利用*Chebyshev*不等式估计$\Pr(|X + Y| \ge 6)$的上界.

解. 根据期望的线性关系有$E[X + Y] = 0$, 根据相关系数的定义有

$$\rho_{XY} = \frac{E[(X - E(X))(Y - E(Y))]}{\sqrt{Var(X)}\sqrt{Var(Y)}} = -\frac{1}{2}.$$

由此可得$Var(X + Y) = Var(X) + Var(Y) + 2E_{XY}[X - E(X)][Y - E(Y)] = 3$. 根据Chebysheve不等式有$\Pr\{|X + Y| \ge 6\} \le \mathrm{Var}(X + Y)/36 = 1/12$. □

大数定律小结

- 伯努力大数定律: 对二项分布$X_n \sim \mathrm{B}(n, p)$, 有$\Pr[|X_n/n - p| < \epsilon] \to 1$ $(n \to \infty)$;
- 切比雪夫大数定律: 对独立随机变量序列$\{X_i\}$满足$\mathrm{Var}(X_i) \le c$, 有$\Pr[|\sum_{i=1}^{n}(X_i - E[X_i])/n| < \epsilon] \to 1$ $(n \to \infty)$;

- 辛钦大数定律: 对独立同分布随机变量序列$\{X_i\}$, 如果期望存在, 则有$\Pr[|\sum_{i=1}^n(X_i - E[X_i])|/n < \epsilon] \to 1\ (n \to \infty)$;
- 马尔可夫大数定律: 如果随机变量序列$\{X_i\}$满足$Var(\sum_{i=1}^n X_n)/n \to 0$, 则有$\Pr[|\sum_{i=1}^n(X_i - E[X_i])|/n < \epsilon] \to 1\ (n \to \infty)$.

7.2 中心极限定理

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

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

的极限分布是否为是正态分布.

首先介绍一个定义: 依分布收敛.

定义7.2 (依分布收敛). *设$Y_1, Y_2, \cdots, Y_n, \cdots$是随机变量序列, Y是一随机变量, 设分布函数$F_{Y_n}(y) = \Pr(Y_n \le y)$和$F_Y(y) = \Pr(Y \le y)$, 如果*

$$\lim_{n\to\infty} \Pr[Y_n \le y] = \Pr[Y \le y], \quad i.e., \quad \lim_{n\to\infty} F_{Y_n}(y) = F_Y(y),$$

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

定理7.6 (林德贝格-勒维中心极限定理, 又称"独立同分布中心极限定理"). *设$X_1, X_2, \cdots, X_n$是独立同分布(i.i.d)随机变量, $E(X_1) = \mu$, $Var(X_1) = \sigma^2$, 则*

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

Y_n是n个随机变量的标准化, 其极限分布为标准正态分布, 上述中心极限定理可表示为

$$\lim_{n\to\infty} \Pr[Y_n \le y] = \Phi(y).$$

当n足够大, 近似有$Y_n \sim \mathcal{N}(0, 1)$, 上述中心极限定理的变形公式:

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

推论7.1 (棣莫弗-拉普拉斯中心极限定理). *设$X_n \sim B(n, p)$, 则*

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

棣莫弗-拉普拉斯中心极限定理表明当n非常大时, 有$\mu_n \overset{\text{近似}}{\sim} \mathcal{N}(np, np(1-p))$. 其应用为: 设随机变量$X_n$服从二项分布$B(n, p)$, 当$n$充分大时有

$$\begin{aligned}
\Pr[a \le X_n \le b] &= \Pr\left[\frac{a - np}{\sqrt{np(1-p)}} \le \frac{X_n - np}{\sqrt{np(1-p)}} \le \frac{b - np}{\sqrt{np(1-p)}}\right] \\
&\approx \Phi\left(\frac{b - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - np}{\sqrt{np(1-p)}}\right).
\end{aligned}$$

大数定律给出了当$n \to \infty$时n个随机变量平均值$\frac{1}{n}\sum_{i=1}^{n} X_i$的趋势. 而中心极限定理给出了当$n \to \infty$时$\frac{1}{n}\sum_{i=1}^{n} X_i$的具体分布.

定理7.7 (李雅普诺夫中心极限定理, 又称"独立不同分布中心极限定理"). *设$\{X_n\}$为独立随机变量序列, 其期望$E[X_n] = \mu_n$, 方差$Var(X_n) = \sigma_n^2 > 0$. 设$B_n^2 = \sum_{k=1}^{n} \sigma_k^2$, 若存在$\delta > 0$, 当$n \to \infty$ 时: $\frac{1}{B_n^{2+\delta}} \sum_{k=1}^{n} E[|X_k - \mu_k|^{2+\delta}] \to 0$, 则*

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

小结:

$$\text{中心极限定理}\begin{cases} \text{独立同分布:} E[X_k] = \mu, Var(X_k) = \sigma^2, \sum_{k=1}^{n} X_k \xrightarrow{d} \mathcal{N}(n\mu, n\sigma^2). \\ \text{棣莫弗-拉普拉斯中心极限定理:} X_k \sim B(k,p), X_k \xrightarrow{d} \mathcal{N}(np, np(1-p)). \\ \text{独立不同分布的李雅普诺夫中心极限定理:} \begin{cases} E[X_k] = \mu_k, Var(X_k) = \sigma_k^2. \\ \sum_{i=1}^{n} X_k \xrightarrow{d} \mathcal{N}(\sum_{k=1}^{n} \mu_k, \sum_{k=1}^{n} \sigma_k^2). \end{cases} \end{cases}$$

利用中心极限定理求解各种问题:

例7.3. *一接收器同时接收到20个信号电压$V_k (k \in [20])$, 它们独立且均服从$U(0,10)$, 求电压和大于105的概率.*

解. 设$V = \sum_{k=1}^{20} V_k$, $E(V_k) = 5$, $Var(V_k) = \frac{100}{12} = \frac{25}{3}$, 所以$E(V) = 100$, $Var(V) = \frac{500}{3}$, 由中心极限定理: $\frac{V - E(V)}{\sqrt{Var(V)}} = \frac{V-100}{\sqrt{500/3}} \sim \mathcal{N}(0,1)$, 所以

$$\Pr(V \geq 105) = \Pr(\frac{V-100}{\sqrt{500/3}} \geq \frac{105-100}{\sqrt{500/3}}) = \Pr(\frac{V-100}{\sqrt{500/3}} \geq 0.387) = 1 - \Phi(0.387)(\text{查表}).$$

□