- 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$F_{Y_n}(y) = \Pr(Y_n \leq y)$和$F_Y(y) = \Pr(Y \leq y)$, 如果

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

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

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

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:

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

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.

例7.3. 假设一个接收器同时接收到 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. 由题意可知$V_1, V_2, \ldots, V_{20}$独立同分布于均匀分布$U(0,10)$, 有$E(V_k) = 5$和$\mathrm{Var}(V_k) = 100/12 = 25/3$. 设$V = \sum_{k=1}^{20} V_k$, 则

$$E(V) = 100 \qquad \mathrm{Var}(V) = 500/3.$$

根据中心极限定理近似有

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

根据$\mathcal{N}(0,1)$的分布函数$\Phi(x)$, 我们得到

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

查表完成证明. □

例7.4. *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, 才能以0.997以上的概率保证不超载?*

解. 假设最多可装n箱不超重, 用X_i表示第i箱重量($i\in[n]$), 易有$E(X_i)=50$, $\mathrm{Var}(X_i)=25$. 设总重量$X=\sum_{i=1}^n X_i$, 则有$E(X)=50n$, $\mathrm{Var}(X)=25n$. 由中心极限定理近似有

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

根据$\mathcal{N}(0,1)$的分布函数$\Phi(x)$, 我们得到

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

有分布函数的单调性有

$$\frac{1000-10n}{\sqrt{n}}>2\Longrightarrow 1000n^2-2000n+1000^2>4n.$$

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

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: 若随机变量$X_n\sim B(n,p)$且n非常大时, 有$X_n\overset{\text{近似}}{\sim}\mathcal{N}(np,np(1-p))$, 从而有如下近似估计:

$$\Pr[X_n\leq y]=\Pr\left[\frac{X_n-np}{\sqrt{np(1-p)}}\leq\frac{y-np}{\sqrt{np(1-p)}}\right]\approx\Phi\left(\frac{y-np}{\sqrt{np(1-p)}}\right).$$

针对上式, 可以考虑利用棣莫弗-拉普拉斯中心极限的三种情况:

- 已知n和$\Pr[X_n\leq y]$, 求y;
- 已知n和y, 求$\Pr[X_n\leq y]$;
- 已知y 和$\Pr[X_n\leq y]$, 求n.

下面看三个例子:

例7.5. 车间有200台独立工作的车床, 每台工作概率为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. 设工作车床数为X, 则有$X \sim B(200, 0.6)$. 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

$$\Pr(X \leq y) \geq 0.999 \quad \Rightarrow \quad \Pr\left(\frac{X-120}{\sqrt{48}} \leq \frac{y-120}{\sqrt{48}}\right) \approx \Phi\left(\frac{y-120}{\sqrt{48}}\right) \geq 0.999 = \Phi(3.1).$$

所以$\frac{y-120}{\sqrt{48}} \geq 3.1$, 即$y \geq 141$. □

例7.6. 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. 设X是损坏的部件数, 则$X \sim B(100, 0.1)$, 则有$E(X) = 10$, $\mathrm{Var}(X) = 9$. 根据棣莫弗-拉普拉斯中心定理近似有$X \sim \mathcal{N}(10, 9)$, 所以

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

□

例7.7. 在一次电视节目调查中, 假设调查了n个人, 其中k个人观看了电视节目, 因此收看比例k/n作为某电视节目收视率p的估计, 要以90%的把握有$|k/n - p| \leq 0.05$成立, 问需要调查多少对象?

解. 设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-拉普拉斯中心定理近似有$X_n \sim \mathcal{N}(0, 1)$, 进一步有

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

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

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

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

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

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

中心极限定理小结:

- 独立同分布中心极限定理: 若$E[X_k] = \mu$, $\mathrm{Var}(X_k) = \sigma^2$, 则$\sum_{k=1}^n X_k \xrightarrow{d} \mathcal{N}(n\mu, n\sigma^2)$;
- 棣莫弗-拉普拉斯中心极限定理: 若$X_k \sim B(k, p)$, 则$X_k \xrightarrow{d} \mathcal{N}(np, np(1-p))$;
- 独立不同分布中心极限定理(李雅普诺夫定理).