Ch5 大数定律和中心极限定理

研究大量的随机现象，常常采用极限形式，由此导致对极限定理进行研究．极限定理的内容很广泛，其中最重要的有两种：

大数定律 与 **中心极限定理**

随机变量序列的两种收敛性

1、依概率收敛

定义 1 设$\{Y_n\}$为一随机变量序列，Y为一随机变量，如果对任意的$\varepsilon > 0$，有

$$\lim_{n\to+\infty} P\left\{|Y_n - Y| < \varepsilon\right\} = 1$$

或

$$\lim_{n\to+\infty} P\left\{|Y_n - Y| \geq \varepsilon\right\} = 0$$

则称$\{Y_n\}$依概率收敛于Y，记作$Y_n \xrightarrow{P} Y$.

定理　若$X_n \xrightarrow{P} X$，$Y_n \xrightarrow{P} Y$，

$g(x, y)$是连续函数，则

$$g(X_n, Y_n) \xrightarrow{P} g(X, Y)$$

2、按分布收敛、弱收敛

例1　设$\{X_n\}$服从退化分布

$$P\left(X_n = \frac{1}{n}\right) = 1, \quad n = 1, 2, \cdots,$$

记X_n的分布函数为$F_n(x)$,则

$$F_n(x) = \begin{cases} 0, & x < \dfrac{1}{n} \\ 1, & x \geq \dfrac{1}{n} \end{cases} \longrightarrow F(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$$

定义2 设随机变量$X, X_1, X_2, \cdots$的分布函数分别为$F(x), F_1(x), F_2(x), \cdots$，若对$F(x)$的任意连续点$x$，都有

$$\lim_{n\to+\infty} F_n(x) = F(x)$$

则称$\{F_n(x)\}$弱收敛于$F(x)$，记作

$$F_n(x) \xrightarrow{W} F(x)$$

也称$\{X_n\}$按分布收敛于X，记作

$$X_n \xrightarrow{L} X.$$

§5.1 大数定律

引例 历史上曾有人做过试验,试图证明抛掷匀质硬币时，出现正反面的机会均等。

实验者	n	n_H	$f_n(H)$
De Morgan	2048	1061	0.5181
Buffon	4040	2048	0.5069
Feller	10000	4979	0.4979
K. Pearson	12000	6019	0.5016
K. Pearson	24000	12012	0.5005

本节要解决的问题

1. 为何能以某事件发生的频率作为该事件的 概率的估计?

2. 为何能以样本均值作为总体期望的估计?

大数定律

大数定律的一般形式

定义1 设随机变量序列$\{X_n\}$,若对任意的$\varepsilon > 0$,有

$$\lim_{n\to+\infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} EX_i\right| < \varepsilon\right\} = 1$$

或

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{n}\sum_{k=1}^{n} EX_k\right| \geq \varepsilon\right) = 0$$

成立，则称该随机变量序列$\{X_n\}$服从大数定律.

大数定律研究的问题是:

随机变量序列 {Xn} 在什么条件下服从大数定律

预备知识

证明大数定律主要的数学工具是切比雪夫不等式.

设随机变量X有期望$E(X)$和方差σ^2，则对于任给 $\varepsilon>0$,

$$P\{|X-E(X)|<\varepsilon\}\geq 1-\frac{\sigma^2}{\varepsilon^2}$$

1、马尔可夫大数定律

定理2 对随机变量序列$\{X_n\}$,若下式成立

$$\frac{1}{n^2}D(\sum_{i=1}^{n}X_i)\to 0$$

马尔可夫条件

则大数定律成立，即对任意的$\varepsilon>0$,有

$$\lim_{n\to+\infty}P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}E(X_i)\right|<\varepsilon\right\}=1$$

设随机变量 $X_1\ X_2\ \cdots, X_n\cdots$ 是一列两两不相关的随机变量，又设它们的方差有界，即存在常数 $c>0$，使

令 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$,　　$E(\overline{X})=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\frac{1}{n}\sum_{i=1}^{n}\mu_i$

$$D(\overline{X})=D\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}D(X_i)\le\frac{c}{n}$$

由切比雪夫不等式得

$$p(|\overline{X}-\frac{1}{n}\sum_{k=1}^{n}E(X_k)|\ge\varepsilon)\le\frac{\sigma^2/n}{\varepsilon^2}$$

$$p(|\frac{1}{n}\sum_{k=1}^{n}X_k-\frac{1}{n}\sum_{k=1}^{n}E(X_k)|\ge\varepsilon)\le\frac{\sigma^2/n}{\varepsilon^2}$$

令 $n->\infty$

$$\lim_{n\to\infty} p\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{n}\sum_{k=1}^{n} E(X_k)\right| \geq \varepsilon\right) = 0$$ 或

$$\lim_{n\to\infty} p\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{n}\sum_{k=1}^{n} E(X_k)\right| < \varepsilon\right) = 1$$

———切比雪夫大数定律

2、切比雪夫Chebyshev 大数定律

定理1 设 r.v. 序列 $X_1, X_2, \cdots, X_n, \cdots$两两不相关，

若每个X_i的方差存在，且有共同的上界，

即$\mathrm{D}(X_i) \le c, \quad i = 1,2,\cdots,$

则 $\forall \varepsilon > 0$ 有

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{n}\sum_{k=1}^{n} EX_k\right| \ge \varepsilon\right) = 0$$

或

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{n}\sum_{k=1}^{n} EX_k\right| < \varepsilon\right) = 1$$

特例：$X_1, X_2, \cdots X_n, \cdots$ 相互独立，且

$$\mathrm{E}(\mathrm{X}_\mathrm{k}) = \mu, \mathrm{D}(\mathrm{X}_\mathrm{k}) = \sigma^2 \ (\mathrm{k} = 1, 2, \cdots)$$

则 $\forall \varepsilon > 0$ 有

$$\lim_{n\to\infty} \mathrm{P}\left(\left| \frac{1}{n} \sum_{k=1}^{n} \mathrm{X}_\mathrm{k} - \mu \right| \geq \varepsilon \right) = 0$$

或

$$\lim_{n\to\infty} \mathrm{P}\left(\left| \frac{1}{n} \sum_{k=1}^{n} \mathrm{X}_\mathrm{k} - \mu \right| < \varepsilon \right) = 1$$

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{P} \mu$$

3、伯努利（Bernoulli）大数定律

定理3 设μ_n是 n 次独立重复伯努利试验中事件 A 发生的次数, p 是每次试验中 A 发生的概率, 则$\forall\, \varepsilon > 0$, 有

$$\lim_{n\to\infty} P\left(\left|\frac{\mu_n}{n} - p\right| \geq \varepsilon\right) = 0$$

或

$$\lim_{n\to\infty} P\left(\left|\frac{\mu_n}{n} - p\right| < \varepsilon\right) = 1$$

注: $\dfrac{\mu_n}{n}$ 称为事件A发生的频率

定理4（辛钦大数定律）

设随机变量序列$X_1, X_2, \ldots$ 相互独立，服从同一分布，具有数学期$E(X_i)=\mu$, $i=1,2,\ldots$， 则对于任意正数ε，有

$$\lim_{n\to\infty} P\{|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| < \varepsilon\} = 1$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} E(X_1) = \mu$$

辛钦

$X_1, X_2, \ldots$ 独立同分布，若$E(X_i)$存在，则

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} E(X_1) = \mu$$

则 $X_1^k, X_2^k, \cdots$ 也独立同分布，若 $E(X_1^k) = \mu_k$ 存在

则 $$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} E(X_1^k) = \mu_k \ (k = 1, 2, \cdots)$$

$$g(A_1, A_2, \cdots, A_k) \xrightarrow{P} g(\mu_1, \mu_2, \cdots, \mu_k)$$

其中 g 为连续函数 .

例2 设随机变量序列$\{X_n\}$独立同分布,且$EX_n^4 < \infty$.

若令$EX_n = \mu, \mathrm{Var}(X_n) = \sigma^2$,考虑

$$Y_n = (X_n - \mu)^2$$

证明:随机变量序列$\{Y_n\}$服从大数定律，即对任意$\varepsilon > 0$,有

$$\lim_{n\to+\infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} Y_i - \sigma^2\right| \geq \varepsilon\right\} = 0$$

例3 设随机变量序列$\{X_n\}$同分布，每个方差存在，且X_n仅与X_{n-1}和X_{n+1}相关，而与其他的X_i不相关，证明$\{X_n\}$服从大数定律.

证明：

$$\frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\left[\sum_{i=1}^{n}\mathrm{Var}\left(X_i\right)+2\sum_{i=1}^{n-1}\mathrm{Cov}\left(X_i,X_{i+1}\right)\right]$$

$$=\frac{1}{n^2}\left[n\cdot\mathrm{Var}\left(X_i\right)+2\sum_{i=1}^{n-1}\mathrm{Cov}\left(X_i,X_{i+1}\right)\right]$$

$$\le\frac{1}{n^2}\left[n\cdot\mathrm{Var}\left(X_i\right)+2(n\text{-}1)\mathrm{Var}\left(X_i\right)\right]\longrightarrow 0$$

$$Cov\left(\mathrm{X_i},\mathrm{X}_{i+1}\right)=\mathrm{E}\left(\mathrm{X_i}-\mathrm{EX_i}\right)\left(\mathrm{X_j}-\mathrm{EX_j}\right)$$

$$\le\sqrt{E\left(X_i-EX_i\right)^2}\sqrt{E\left(X_j-EX_j\right)^2}=\mathrm{Var}\left(X_i\right)$$

§5.2 中心极限定理

本节要解决的问题

1、多个随机变量和的分布是什么？

2、为何正态分布在概率论中占有极其重要的地位？

3、大样本统计推断的理论基础是什么？

中心极限定理

中心极限定理的客观背景

中心极限定理研究的问题是：

随机变量之和 $X_1+X_2+\ldots\ldots+X_n$ 的概率分布

当n无限增大时，这个和的极限分布是什么呢？

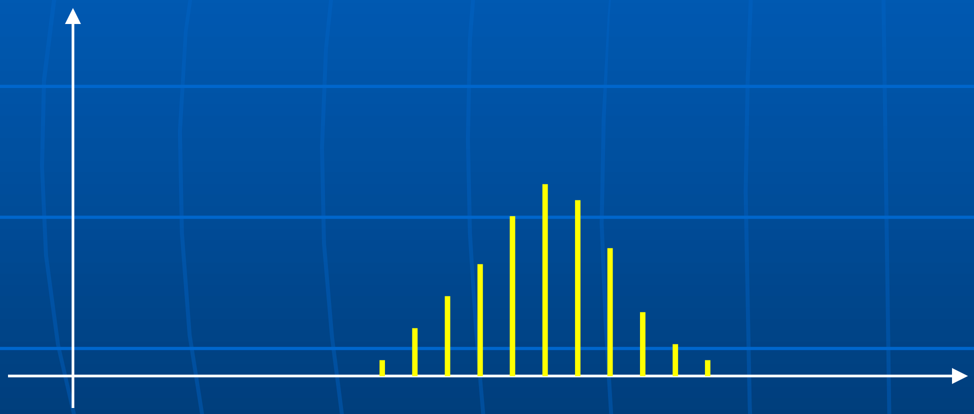
在什么条件下极限分布会是正态的呢？

1、独立随机变量和

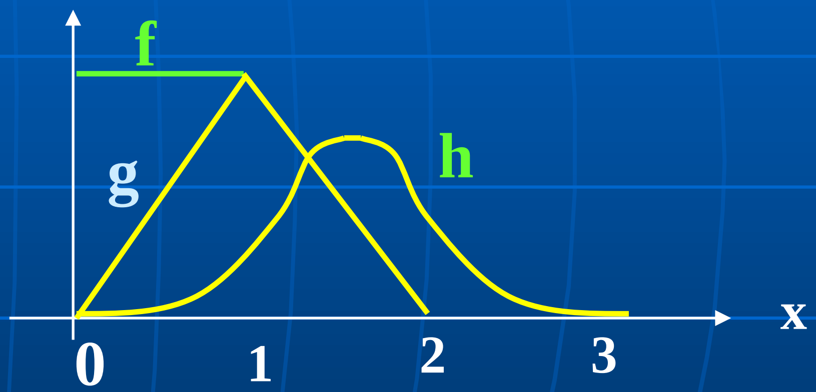
讨论对象 $Y_n=\sum_{i=1}^{n}X_i$

$$Y_n^*=\frac{Y_n-EY_n}{\sqrt{\mathrm{Var}(Y_n)}}$$

讨论问题 当$n\to+\infty$时，Y_n的分布是什么？



例:20个0-1分布的和的分布



几个(0,1)上均匀分布的和的分布

$X_1\sim f(x)$　　$X_1+X_2\sim g(x)$

$X_1+X_2+X_3\sim h(x)$

2、独立同分布的中心极限定理

定理1 林德贝格－勒维中心极限定理

设随机变量序列$\{X_n\}$独立同分布，且有期望和方差：

$$E(X_k)=\mu,\ \mathrm{Var}(X_k)=\sigma^2>0,\ k=1,2,\cdots$$

则对于任意实数 x， $F_n(x)\xrightarrow{W}\Phi(x)$

$$\lim_{n\to\infty}P\left(\frac{\sum_{k=1}^{n}X_k-n\mu}{\sqrt{n}\sigma}\le x\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt=\Phi(x)$$

$$Y_n^*\xrightarrow{L}Y\sim N(0,1).$$

此定理说明，不论$X_1, X_2, \cdots X_n, \cdots$ 原来服从什么分布，只要是独立同分布，当 n 足够大时，总可以近似地认为

$$Y_n^* = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \overset{\text{近似}}{\sim} N(0,1)$$

$$Y_n = \sum_{i=1}^{n} X_i \overset{\text{近似}}{\sim} N(n\mu, n\sigma^2),$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{\text{近似}}{\sim} N(\mu, \frac{\sigma^2}{n})$$

例1：一加法器同时收到20个噪声电压Vi，设它们是相互独立的随机变量，且都在区间 (0,10) 上服从均匀分布，记 $V=\sum_{i=1}^{20}V_i$， 求 $P\{V>105\}$.

解：因为 $V_i\sim U(0,10)$， 易知

$$E(V_i)=\frac{0+10}{2}=5\quad D(V_i)=\frac{1}{12}(10-0)^2=\frac{100}{12}\quad i=1,2,\cdots,20$$

由定理知

$$\xi_{20}=\frac{\sum_{i=1}^{20}V_i-20E(V_i)}{\sqrt{D(\sum_{i=1}^{20}V_i)}}=\frac{V-20\times5}{\sqrt{20}\times\sqrt{\frac{100}{12}}}$$

近似服从 $N(0,1)$ 分布。

于是

$$P(V>105)=P\left\{\frac{V-20\times5}{\sqrt{20}\times\sqrt{\frac{100}{12}}}>\frac{105-20\times5}{\sqrt{20}\times\sqrt{\frac{100}{12}}}\right\}$$

$$=P\left\{\frac{V-100}{10\times\sqrt{\frac{5}{3}}}>0.387\right\}\approx1-\Phi(0.387)\approx0.348$$

所以 $P\{V>105\}\approx0.348$

3、二项分布的正态近似

定理2 **棣莫佛—拉普拉斯中心极限定理**

设 $Y_n \sim b(n, p)$，$0 < p < 1$，$n = 1,2,\ldots$

则对任一实数 x，有

$$\lim_{n\to\infty} P\left(\frac{Y_n - np}{\sqrt{np(1-p)}} \le x\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}\,dt$$

即对任意的 $a < b$,

$$\lim_{n\to\infty} P\left(a < \frac{Y_n - np}{\sqrt{np(1-p)}} \le b\right) = \frac{1}{\sqrt{2\pi}}\int_{a}^{b} e^{-\frac{t^2}{2}}\,dt$$

$Y_n \sim N(np, np(1-p))$ (近似)

中心极限定理的应用

例2 炮火轰击敌方防御工事 100 次, 每次轰击命中的炮弹数服从同一分布, 其数学期望为 2 , 标准差为1.5. 若各次轰击命中的炮弹数是相互独立的, 求100 次轰击

(1) 至少命中180发炮弹的概率;

(2) 命中的炮弹数不到200发的概率.

解 设 X_k 表示第 k 次轰击命中的炮弹数

$$E(X_k)=2,\quad \mathrm{Var}(X_k)=1.5^2,\quad k=1,2,\cdots,100$$

$X_1, X_2, \cdots, X_{100}$ 相互独立,

设 X 表示100次轰击命中的炮弹数, 则

$$X=\sum_{k=1}^{100}X_k,\quad E(X)=200,\quad \mathrm{Var}(X)=225,$$

由独立同分布中心极限定理, 有

$$X\overset{\text{近似}}{\sim}N(200,225)$$

(1) $$P(X\geq 180)\approx 1-\Phi\left(\frac{180-200}{15}\right)$$

$$=1-\Phi(-1.3)=\Phi(1.3)=0.91$$

(2) $$P(0\leq X<200)\approx\Phi\left(\frac{200-200}{15}\right)-\Phi\left(\frac{0-200}{15}\right)$$

$$=\Phi(0)-\Phi(-13.33)=0.5$$

例3 某车间有200台车床，每台独立工作，每台正常工作的概率为0.6. 开工时每台耗电量为 r 千瓦. 问供 电所至少要供给这个车间多少电力, 才能以 99.9% 的概率保证这个车间不会因供电不足而影响生产?

解 设至少要供给这个车间 a 千瓦的电力,

X 为开工的车床数， 则 $X \sim b(200, 0.6)$,

由棣莫弗—拉普拉斯中心极限定理，有

$$X \sim N(120, 48) \text{ (近似)}$$

问题转化为求 a, 使 $P(0 \le rX \le a) = 99.9\%$

$$P(0 \le rX \le a) \approx \Phi\left(\frac{a/r-120}{\sqrt{48}}\right) - \Phi\left(\frac{0-120}{\sqrt{48}}\right)$$

$$= \Phi\left(\frac{a/r-120}{\sqrt{48}}\right)$$

$$\boxed{\begin{aligned} &= \Phi(-17.32) \\ &\approx 0 \end{aligned}}$$

反查标准正态函数分布表, 得

$\Phi(3.09) = 99.9\%$ 令 $\dfrac{\frac{a}{r}-120}{\sqrt{48}} = 3.09$

解得 $a = (3.09\sqrt{48}+120)r$

$\approx 141r$ (千瓦)

例4 设有一批种子，其中良种占1/6. 试估计在任选的6000粒种子中，良种比例与 1/6 比较上下不超过1%的概率.

解 设 X 表示6000粒种子中的良种数，

则 $X \sim b(6000, 1/6)$

由德莫佛—拉普拉斯中心极限定理,

有 $X \overset{\text{近似}}{\sim} N\left(1000, \dfrac{5000}{6}\right)$

$$P\left(\left|\frac{X}{6000}-\frac{1}{6}\right|<0.01\right)=P\left(\left|X-1000\right|<60\right)$$

$$\approx \Phi\left(\frac{1060-1000}{\sqrt{5000/6}}\right)-\Phi\left(\frac{940-1000}{\sqrt{5000/6}}\right)$$

$$=\Phi\left(\frac{60}{\sqrt{5000/6}}\right)-\Phi\left(\frac{-60}{\sqrt{5000/6}}\right)$$

$$=2\Phi\left(\frac{60}{\sqrt{5000/6}}\right)-1\approx 0.9624$$

比较几个近似计算的结果

二项分布(精确结果) $P\left(\left|\frac{X}{6000}-\frac{1}{6}\right|<0.01\right)\approx 0.9590$

中心极限定理 $P\left(\left|\frac{X}{6000}-\frac{1}{6}\right|<0.01\right)\approx 0.9624$

Poisson 分布 $P\left(\left|\frac{X}{6000}-\frac{1}{6}\right|<0.01\right)\approx 0.9379$

Chebyshev 不等式 $P\left(\left|\frac{X}{6000}-\frac{1}{6}\right|<0.01\right)\geq 0.7685$

4、独立不同分布下的中心极限定理

设$\{X_n\}$是相互独立的随机变量序列，且它们具有有限的数学期望和方差：

$$E\left(X_i\right)=\mu_i,\quad \mathrm{Var}\left(X_i\right)=\sigma_i^2,\quad i=1,2,\cdots$$

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

记$B_n=\sqrt{\sum_{i=1}^{n}\sigma_i^2}$

$$EY_n=\sum_{i=1}^{n}\mu_i \qquad \mathrm{Var}(Y_n)=\sum_{i=1}^{n}\sigma_i^2$$

$$Y_n^*=\frac{Y_n-EY_n}{\sqrt{\mathrm{Var}(Y_n)}}=\frac{Y_n-\sum_{i=1}^{n}\mu_i}{B_n}=\sum_{i=1}^{n}\frac{X_i-\mu_i}{B_n}$$

定理4 李雅浦诺夫中心极限定理

设独立随机变量序列$\{X_n\}$，若存在$\delta > 0$,满足

$$\lim_{n\to+\infty}\frac{1}{B_n^{2+\delta}}\sum_{i=1}^{n}E(|X_i-\mu_i|^{2+\delta})=0$$

则对任意的x，有

$$\lim_{n\to+\infty}P\left\{\frac{1}{B_n}\sum_{i=1}^{n}(X_i-\mu_i)\le x\right\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt=\Phi(x)$$