概率论与数理统计

范正平
fanzhp@mail.sysu.edu.cn

中山大学 数据科学与计算机学院

Chapter 5
大数定理及中心极限定理

随机系列的收敛性

定义 1 设 $X_1, X_2, \cdots, X_n, \cdots$ 为一个随机变量序列，记为$\{X_n\}$ ，若对任何$n \geq 2$，随机变量 $X_1, X_2, \cdots, X_n$ 都相互独立，则称 $\{X_n\}$ 是**相互独立的随机变量序列**。

定义2 设 $\{X_n\}$ 为一随机变量序列，X 为一随机变量或常数，若对任意 $\varepsilon > 0$，有

$$\lim_{n\to\infty} P\{|X_n - X| < \varepsilon\} = 1$$

则称 $\{X_n\}$ 依概率收敛于X, 记为 $X_n \xrightarrow{P} X$ 或

$$X_n - X \xrightarrow{P} 0 \qquad n \to \infty$$

请注意：

$\{X_n\}$依概率收敛于X，意味着对任意给定的$\varepsilon > 0$，当n充分大时，事件$|X_n - X| < \varepsilon$的概率很大，接近于1；并不排除事件$|X_n - X| \geq \varepsilon$的发生，而只是说他发生的可能性很小.

依概率收敛比高等数学中的普通意义下的收敛弱些，它具有某种不确定性.

伯努利（Bernoulli） 大数定律

定理：设 n_A 是 n 次独立重复试验中事件 A 发生的次数, p 是每次试验中 A 发生的概率，则

$\forall \varepsilon > 0$ 有

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

或

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

故 $\dfrac{n_A}{n} \xrightarrow[n\to\infty]{P} p$

伯努利（Bernoulli） 大数定律的意义：

在概率的统计定义中，事件 A 发生的频率 $\frac{n_A}{n}$

“稳定于”事件 A 在一次试验中发生的概率是指：

频率 $\frac{n_A}{n}$ 与 p 有较大偏差 $\left(\left|\frac{n_A}{n}-p\right|\geq\varepsilon\right)$ 是

小概率事件, 因而在 n 足够大时, 可以用频率近似代替 p. 这种稳定称为依概率稳定.

伯努利（Bernoulli） 大数定律

$$X_k = \begin{cases} 1, & \text{第}k\text{次试验}A\text{发生} \\ 0, & \text{第}k\text{次试验}\overline{A}\text{发生} \end{cases}$$

则 $$n_A = \sum_{k=1}^{n} X_k$$

$$E(X_k) = p$$

$$E(\sum_{k=1}^{n} X_k) =$$

$$E(X_1) + E(X_2) + \ldots + E(X_n)$$

$$= np$$

伯努利（Bernoulli） 大数定律

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

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

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

切比雪夫(Chebyshev)大数定律

定理：设随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立，

(指任意给定 $n > 1, X_1, X_2, \cdots, X_n$ 相互独立)，且具有相同的数学期望和方差

$$E(X_k) = \mu, D(X_k) = \sigma^2, \quad k = 1,2,\cdots$$

则 $\forall \varepsilon > 0$ 有

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

或

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

切比雪夫(Chebyshev)大数定律(Cont.)

说明

1、定理中$\{|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu|<\varepsilon\}$是指一个随机事件，当$n\to\infty$时，这个事件的概率趋于1.

2、定理以数学形式证明了随机变量$X_1,\cdots X_n$的算术平均$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$接近数学期望$E(X_k)=\mu$（$k=1,2\cdots,n$），这种接近说明其具有的稳定性.

这种稳定性的含义说明算术平均值是依概率收敛的意义下逼近某一常数.

注1： $X_1, X_2, \cdots, X_n, \cdots$ 不一定有相同的数学期望与方差，可设

$$E(X_k)=\mu_k,\ D(X_k)=\sigma_k^2 \boxed{\le \sigma^2},\quad k=1,2,\cdots$$

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

注2： $X_1, X_2, \cdots, X_n, \cdots$ 相互独立的条件可以去掉，代之以

$$\frac{1}{n^2}D\left(\sum_{k=1}^{n} X_k\right)\xrightarrow{n\to\infty}0$$

切比雪夫大数定律中的条件可以弱化:

下面给出的独立同分布下的大数定律，

不要求随机变量的方差存在.

定理（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

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

注　1、辛钦大数定律为寻找随机变量的期望值提供了一条实际可行的途径.

2、伯努利大数定律是辛钦定理的特殊情况.

3、辛钦定理具有广泛的适用性.

要估计某地区的平均亩产量，要收割某些有代表性块，例如n 块地．计算其平均亩产量，则当n 较大时，可用它作为整个地区平均亩产量的一个估计.

中心极限定理(Cont.)

问题: 某个随机变量是由大量相互独立且均匀小的随机变量相加而成的, 研究其概率分布情况.

如果一个随机变量是由大量相互独立的随机因素的综合影响所造成，而每一个别因素对这种综合影响中所起的作用不大. 则这种随机变量一般都服从或近似服从正态分布.

由于无穷个随机变量之和可能趋于∞，故我们不研究n个随机变量之和本身而考虑它的标准化的随机变量.即考虑随机变量$X_k(k=1,\cdots n)$的和$\sum_{k=1}^{n} X_k$

$$Y_n=\frac{\sum_{k=1}^{n} X_k-E(\sum_{k=1}^{n} X_k)}{\sqrt{D(\sum_{k=1}^{n} X_k)}}$$

讨论Y_n的极限分布是否为标准正态分布

在概率论中，习惯于把和的分布收敛于正态分布这一类定理都叫做中心极限定理.

中心极限定理

定理 (林德贝格-列维中心极限定理)

设随机变量$X_1, X_2, \cdots X_n, \cdots$相互独立，服从同一分布，且具有数学期望和方差：$E(X_k)=\mu, D(X_k)=\sigma^2>0$ $(k=1,2,\cdots)$，则随机变量之和$\sum_{k=1}^{n} X_k$的标准化变量

$$Y_n=\frac{\sum_{k=1}^{n} X_k-n\mu}{\sqrt{n}\sigma}$$ 对于任意x,满足

$$\lim_{n\to\infty} P\left\{\frac{\sum_{i=1}^{n} X_i-n\mu}{\sigma\sqrt{n}}\le x\right\}=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt$$

$$=\Phi(x)$$

林德贝格-列维中心极限定理

注 1、定理表明，独立同分布的随机变量之和$\sum_{k=1}^{n} X_k$，当n充分大时，随机变量之和与其标准化变量分别有

$$\sum_{k=1}^{n} X_k \overset{\text{近似地}}{\sim} N(n\mu, n\sigma^2) \quad ;$$

$$\frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\sigma} \overset{\text{近似地}}{\sim} N(0,1).$$

林德贝格-列维中心极限定理(Cont.)

2、独立同分布中心极限定理的另一种形式可写为

$$\overline{X} \overset{\text{近似地}}{\sim} N(\mu,\sigma^2/n) \quad \text{或} \quad \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \overset{\text{近似地}}{\sim} N(0,1)$$

其中$\overline{X}=\dfrac{1}{n}\sum\limits_{k=1}^{n}X_k$

3、虽然在一般情况下，我们很难求出$\sum\limits_{k=1}^{n}X_k$的分布的确切形式，但当n很大时，可以求出近似分布.

当 $X_1, X_2, \cdots, X_n \cdots$相互独立 ,但不服从同一分布时:

定理（李雅普诺夫(Liapounov)定理)

设随机变量$X_1, X_2, \cdots, X_n \cdots$相互独立，它们具有数学期望和方差:

$$E(X_k) = \mu_k, \qquad D(X_k) = \sigma_k^2, (k = 1, 2, \cdots)$$

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

若存在正数δ，使得当$n \to \infty$时，

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

则随机变量之和$\sum\limits_{k=1}^{n} X_k$的标准化变量:

$$\frac{\sum_{k=1}^{n} X_k - E(\sum_{k=1}^{n} X_k)}{\sqrt{D(\sum_{k=1}^{n} X_k)}} = \frac{\sum_{k=1}^{n} X_k - \sum_{k=1}^{n} \mu_k}{B_n}$$

对于任意x，满足

$$\lim_{n\to\infty} P\left\{\frac{\sum_{k=1}^{n} X_k - \sum_{k=1}^{n} \mu_k}{B_n} \le x\right\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2} \mathrm{d}t$$

$$= \Phi(x)$$

$$Z_n = \frac{\sum_{k=1}^{n} X_k - \sum_{k=1}^{n} \mu_k}{B_n}$$

请注意：

1、定理中随机变量之和$\sum_{k=1}^{n} X_k$及其标准化变量Z_n在n很大时,分别近似服从

$$\sum_{k=1}^{n} X_k \overset{\text{近似地}}{\sim} N(\sum_{k=1}^{n} \mu_k, B_n^2)\ ; \qquad Z_n \overset{\text{近似地}}{\sim} N(0,1)$$

2、随机变量X_k无论服从什么分布，只要满足定理条件，随即变量之和$\sum_{k=1}^{n} X_k$，当n很大时，就近似服从正态分布，这就是为什么正态分布在概率论中所占的重要地位的一个基本原因.

3. 棣莫弗——拉普拉斯中心极限定理

设随机变量 $\eta_n (n = 1, 2, \cdots)$ 服从参数为n, p的二项分布，则对任意x，有

$$\lim_{n\to\infty} P\{\frac{\eta_n - np}{\sqrt{np(1-p)}} \le x\}$$

$$= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \Phi(x)$$

定理表明, 当n充分大时, 二项分布$B(n,p)$ 可近似地用正态分布$N(np, np(1-p))$来代替. 因此, 当$X\sim B(n, p)$, 且n充分大时, 有

$$P\{a < X \le b\} \approx \Phi(\frac{b-np}{\sqrt{npq}}) - \Phi(\frac{a-np}{\sqrt{npq}})$$

其中$q=1-p$.

利用微分中值定理可以进一步证明:
对于随机变量$X\sim B(n, p)$, n充分大时,

$$P\{X=k\}\approx\frac{1}{\sqrt{2\pi npq}}e^{-\frac{(k-np)^2}{2npq}}=\frac{1}{\sqrt{npq}}\varphi\left(\frac{k-np}{\sqrt{npq}}\right)$$

注正态分布和泊松分布虽然都是二项分布的

极限分布,泊松定理要求$n\to\infty$，同时

$$p\to 0,\quad np\to\lambda$$

棣莫佛—拉普拉斯定理只要求$n\to\infty$。

例 一加法器同时收到20个噪声电压$V_k\,(k=1,2,\cdots n)$，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$V=\sum_{k=1}^{n}V_k$，求$P\{V>105\}$的近似值.

解 易知$E(V_k)=5,\ D(V_k)=100/12 \quad (k=1,2,\cdots 20).$

由定理知，$V=\sum_{k=1}^{20}V_k \overset{\text{近似地}}{\sim} N(20\times 5, \frac{100}{12}\times 20)$

于是 $P\{V>105\}=p\left\{\frac{V-20\times 5}{(\sqrt{100/12})\sqrt{20}}>\frac{105-20\times 5}{(\sqrt{100/12})\sqrt{20}}\right\}$

$$=p\left\{\frac{V-20\times 5}{(\sqrt{100/12})\sqrt{20}}>0.387\right\}$$

$$= 1 - p\left\{ \frac{V - 20\times 5}{\left(\sqrt{100/12}\right)\sqrt{20}} \le 0.387 \right\}$$

$$\approx 1 - \Phi(0.387) = 0.348$$

即有 $P\{V > 105\} \approx 0.348$

例. (供电问题)某车间有200台车床,在生产期间由于需要检修、调换刀具、变换位置及调换工件等常需停车. 设开工率为0.6, 并设每台车床的工作是独立的，且在开工时需电力1千瓦.

问应供应多少瓦电力就能以99.9%的概率保证该车间不会因供电不足而影响生产?

解：对每台车床的观察作为一次试验，每次试验是观察该台车床在某时刻是否工作，工作的概率0.6，共进行200次独立重复试验.

用X表示在某时刻工作着的车床数，

依题意， $X \sim B(200,0.6)$,

设需N千瓦， 现在的问题是：

求满足 $P(1^*X \leq N) \geq 0.999$ 的最小的N.

（由于每台车床在开工时需电力1千瓦，N台工作所需电力即N千瓦.）

由中心极限定理

$$\frac{X-np}{\sqrt{np(1-p)}}$$ 近似$N(0,1)$,

这里 $np=120$, $np(1-p)=48$

于是 $P(X\le N)=P(0\le X\le N)$

$$\approx \Phi(\frac{N-120}{\sqrt{48}})-\Phi(\frac{-120}{\sqrt{48}})$$

由3σ准则，此项为0。

$$\Phi\left(\frac{0-120}{\sqrt{48}}\right)=\Phi(-17.32)\approx 0$$

$$\approx \Phi(\frac{N-120}{\sqrt{48}})$$

由$\Phi(\frac{N-120}{\sqrt{48}})\ge 0.999$ 查正态分布函数表得

$$\Phi(3.1)=0.999$$

故 $\dfrac{N-120}{\sqrt{48}} \geq 3.1$, 从中解得$N \geq 141.5$,

即所求$N=142$.

也就是说, 应供应142 千瓦电力就能以99.9%的概率保证该车间不会因供电不足而影响生产.

例 每颗炮弹命中目标的概率都为0.01,求500发炮弹至少命中2发的概率.

解 500发炮弹命中目标的炮弹数$X \sim B(n,p)$,
其中：$n=500$, $p=0.01$, $np=5$,

下面用三种方法计算并加以比较

要求的是P{X≥2}

1° 用二项分布公式计算：

$$P\{X\geq 2\}=1-P\{X=0\}-P\{X=1\}$$

$$=1-C_{500}^{0}\times 0.01^{0}\times 0.99^{500}-C_{500}^{1}\times 0.01\times 0.99^{499}\approx 0.96024$$

2° 用泊松分布计算： $\lambda=np=500\times 0.01=5$

$$P\{X\geq 2\}\approx 1-\frac{\lambda^{0}}{0!}e^{-\lambda}-\frac{\lambda^{1}}{1!}e^{-\lambda}=1-e^{-5}(1+5)\approx 0.95957$$

3° 用正态分布近似计算

$$P\{X \geq 2\} = 1 - P\{X < 2\} = 1 - P\{X \leq 1\}$$

$$\approx 1 - \Phi\left(\frac{1 - np}{\sqrt{npq}}\right)$$

$$= 1 - \Phi(-1.7978)$$

$$= 0.96327$$

大数定律与中心极限定理的区别:

设 $\{X_n\}$ 为独立同分布随机变量序列，且

$$EX_i = \mu \qquad DX_i = \sigma^2 > 0$$

则由大数定理，对于任意的 $\varepsilon > 0$ 有

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

大数定律并未给出 $P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| < \varepsilon\right\}$ 的表达式, 但保证了其极限是1.

大数定律与中心极限定理的区别(Cont.)

而在以上同一条件下，中心极限定理(林德伯格—莱维）亦成立，这时，对于任意的$\varepsilon > 0$及某固定的n，有

$$P\left\{\left|\frac{1}{n}\sum X_i - \mu\right| < \varepsilon\right\} = P\left\{\left|\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}\right| < \frac{\sqrt{n}\varepsilon}{\sigma}\right\}$$

由于$2\Phi\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right) - 1 \xrightarrow[n\to\infty]{} 1$，因此，在所给条件下，中心极限定理不仅给出了概率的近似表达式，而且也能保证了其极限是1，可见中心极限定理的结论更为深入。

作业:

Exes.: 3, 13