第五章 大数定律及中心极限定理

§1 大数定律

§2 中心极限定理及其应用

§1 大数定律 (law of large number)

- 大数定律的定义
- 辛钦大数定理(弱大数定理)
- 伯努利大数定理

一、定义

定义1 (依概率收敛)

设$Y_1,\cdots,Y_n,\cdots$是随机变量序列， a是一个常数；

若对任意 $\varepsilon > 0$，有

$$\lim_{n\to\infty} P\{|Y_n - a| < \varepsilon\} = 1, \text{ 或 } \lim_{n\to\infty} P\{|Y_n - a| \geq \varepsilon\} = 0,$$

则称$Y_1,\cdots,Y_n,\cdots$依概率收敛于 a,记为

$$Y_n \xrightarrow{P} a.$$

想想：数列的收敛性定义，

比较数列与随机变量序列收敛性的区别。

注意：

$\{X_n\}$依概率收敛于a，意味着对任意给定的$\varepsilon > 0$，当n充分大时，事件$|X_n - a| < \varepsilon$的概率很大，接近于1；

并不排除事件$|X_n - a| \geq \varepsilon$的发生，而只是说它发生的可能性很小.

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.

定义2 （大数定律）

设$X_1,\cdots X_n\cdots$是随机变量序列，对任意 $\varepsilon>0$，有

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

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

记为a_n

即 $$\frac{1}{n}\sum_{k=1}^{n}X_k-a_n\xrightarrow{P}0,$$

则称 $\{X_n\}$服从大数定律。

依概率收敛的序列有以下性质:

若 $X_n \xrightarrow{P} a$, $Y_n \xrightarrow{P} b$,

函数$g(x,y)$在点(a,b)连续,

则 $g(X_n, Y_n) \xrightarrow{P} g(a,b)$.

定理1　辛钦大数定理(弱大数定理)

设随机变量 $X_1,\cdots X_n\cdots$ 相互独立同分布，且具有数学期望 $E(X_k)=\mu,\ \ k=1,2,\cdots,n,\cdots$

则：对任意的 $\varepsilon>0$ 有

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

即:

序列 $\overline{X}=\frac{1}{n}\sum_{k=1}^{n}X_k$ 依概率收敛于 μ,

即　$\overline{X}\xrightarrow{P}\mu.$

证明:

只在方差 $D(X_k)=\sigma^2 \ (k=1,2,\cdots)$存在的条件下证明。

因为$E(\frac{1}{n}\sum_{k=1}^{n}X_k)=\frac{1}{n}\sum_{k=1}^{n}E(X_k)=\frac{1}{n}(n\mu)=\mu,$

又由独立性得

$$D(\frac{1}{n}\sum_{k=1}^{n}X_k)=\frac{1}{n^2}\sum_{k=1}^{n}D(X_k)=\frac{1}{n^2}(n\sigma^2)=\frac{\sigma^2}{n},$$

由切比雪夫不等式得

$$1\geq P\{|\frac{1}{n}\sum_{k=1}^{n}X_k-\mu|<\varepsilon\}\geq 1-\frac{\sigma^2/n}{\varepsilon^2}$$

即得 $\lim\limits_{n\to\infty}P\{|\frac{1}{n}\sum_{k=1}^{n}X_k-\mu|<\varepsilon\}=1.$ 证毕。

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.

例如要估计某地区的平均亩产量，要收割某些有代表性地块，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.

定理2 伯努利大数定理 (*Bernoulli* 大数定理)

雅各布第一·伯努利

设 f_A 为n重伯努利试验中事件 A发生的次数，

p是事件A在每次试验中发生的概率，

则：对任意的 $\varepsilon > 0$ 有

$$\lim_{n\to\infty} P\{|\frac{f_A}{n} - p| < \varepsilon\} = 1, \quad \text{或} \lim_{n\to\infty} P\{|\frac{f_A}{n} - p| \geq \varepsilon\} = 0.$$

证： 令

$$X_k = \begin{cases} 1, & \text{第}k\text{次试验中}A\text{发生}, \\ 0, & \text{第}k\text{次试验中}A\text{不发生}, \end{cases} \qquad k = 1,2,\cdots,n.$$

则：$f_A=\sum_{k=1}^{n}X_k$,

其中$X_1,\cdots,X_n$ 相互独立同服从于两点分布.

且 $E(X_k)=p$，$D(X_k)=p(1-p)$，$k=1,2,\cdots,n$,

由**定理1**有 $\lim_{n\to\infty}P\{|\frac{1}{n}\sum_{k=1}^{n}X_k-p|<\varepsilon\}=1$,

即$\lim_{n\to\infty}P\{|\frac{f_A}{n}-p|<\varepsilon\}=1.$

该定理给出了频率的稳定性的严格的数学意义。

注：伯努利大数定理是辛钦大数定理的特殊情况。

小结

大数定律

伯努利大数定律 $\lim_{n\to\infty} P\{|\frac{f_A}{n} - p| < \varepsilon\} = 1$ $\quad n_A \sim b(n,p)$

辛钦大数定律 $\lim_{n\to\infty} P\{|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| < \varepsilon\} = 1$ $\quad E(X_k) = \mu$

大数定律以严格的数学形式表达了随机现象最根本的性质之一：

平均结果的稳定性

练习

设随机变量序列$X_1,X_2,\cdots,X_n,\cdots$独立同服从参数为$\lambda=\dfrac{1}{2}$的指数分布，

则当$n\to\infty$时，$\dfrac{1}{n}\sum\limits_{i=1}^{n}X_i^2$ 依概率收敛于$\underline{\quad\dfrac{1}{2}\quad}$

§2 中心极限定理 (central limit theorem)

- 定义
- 独立同分布的中心极限定理
- 李雅普诺夫定理
- 棣莫弗-拉普拉斯定理

中心极限定理的客观背景

在实际问题中许多随机变量是由相互独立随机因素的综合（或和)影响所形成的.

例如：炮弹射击的落点与目标的偏差，就受着许多随机因素（如瞄准，空气阻力，炮弹或炮身结构等）综合影响的.

每个随机因素对弹着点（随机变量和）所起的作用都是很小的.那么弹着点服从怎样分布呢 ？

现在我们就来研究独立随机变量之和所特有的规律性问题.

自从高斯指出测量误差服从正态分布之后，人们发现，正态分布在自然界中极为常见.

高斯

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，而每一个个别因素在这种综合影响中所起的作用不大.

这种量(随机变量和)一般都服从或近似服从正态分布

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**.

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$\sum_{k=1}^{n} X_k$

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

讨论Z_n的极限分布是否为标准正态分布.

一、定义

设$X_1,\cdots,X_n,\cdots$相互独立的随机变量序列，且

$E(X_k),D(X_k)$存在，令：

$$Z_n=\left(\sum_{k=1}^{n}X_k-\sum_{k=1}^{n}E(X_k)\right)\Big/\sqrt{\sum_{k=1}^{n}D(X_k)},$$

随机变量之和$\sum_{k=1}^{n}X_k$的标准化变量

若对任意$x\in R$有

$$\lim_{n\to\infty}P\{Z_n\le x\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt=\Phi(x)$$

则称$\{X_n\}$服从中心极限定理。

即当n较大时，Z_n近似服从$N(0,1)$分布.

二、中心极限定理

定理1（独立同分布的中心极限定理）

设随机变量$X_1,\cdots,X_n,\cdots$相互独立,服从同一分布，且

$$E(X_k)=\mu,\quad D(X_k)=\sigma^2\neq 0,(k=1,2,\cdots)$$

则$\{X_n\}$服从中心极限定理，即：$\forall x\in R$

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

中心极限定理说明了正态分布的重要地位，它也是统计学中处理大样本时的重要工具。

第五章 大数定律及中心极限定理

注:

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

2、当n充分大时，随机变量之和与其标准化变量分别有

$$\sum_{k=1}^{n} X_k \overset{\text{近似地}}{\sim} N(n\mu, n\sigma^2) \quad ; \qquad \frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\sigma} \overset{\text{近似地}}{\sim} N(0,1).$$

3、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**$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$

定理2 （李雅普诺夫(Lyapunov)定理）

设$X_1,\cdots,X_n,\cdots$相互独立，且$E(X_k)=\mu_k$,

$D(X_k)=\sigma_k^2\neq 0,(k=1,2,\cdots)$,

设$B_n^2=\sum_{k=1}^{n}\sigma_k^2$,若存在正数$\delta$,

使得当$n\to\infty$时，$\frac{1}{B_n^{2+\delta}}\sum_{k=1}^{n}E\{|X_k-\mu_k|^{2+\delta}\}\to 0$

则$\{X_n\}$服从中心极限定理，即：

$$\lim_{n\to\infty}P\{\sum_{k=1}^{n}(X_k-\mu_k)/\sqrt{\sum_{k=1}^{n}\sigma_k^2}\le x\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt=\Phi(x).$$

定理3（棣莫弗-拉普拉斯(De Moivre-Laplace)定理）

设随机变量 $\eta_n \sim b(n,p)\ (n=1,2,\cdots,\ 0<p<1)$,

则对任意 $x\in R$,有

$$\lim_{n\to\infty} P\{\frac{\eta_n - np}{\sqrt{npq}} \le x\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt = \Phi(x).$$

证明： 由二项分布和两点分布的关系知 $\eta_n = \sum_{k=1}^{n} X_k$,

其中 $X_1,\cdots,X_n$ 相互独立且都服从于两点分布，且

$E(X_k)=p$，$D(X_k)=pq$

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

由定理1有结论成立。

推论：

$$\frac{\eta_n - np}{\sqrt{npq}} \text{ 近似地 } \sim N(0,1)$$

设随机变量 $\eta_n \sim b(n,p)(n=1,2,\cdots)\ (0<p<1)$，

则当 n 充分大时，有

$$P\{a<\eta_n \le b\} = P\{\frac{a-np}{\sqrt{npq}} < \frac{\eta_n - np}{\sqrt{npq}} \le \frac{b-np}{\sqrt{npq}}\}$$

$$\approx \Phi(\frac{b-np}{\sqrt{npq}}) - \Phi(\frac{a-np}{\sqrt{npq}})$$

说明： 这个公式给出了 n 较大时二项分布的概率计算方法。

中心极限定理的应用

例1 车间有200台车床，它们独立地工作着，开工率为0.6,开工时耗电各为1千瓦，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正常生产？

解: 记某时刻工作着的车床数为 X，则 $X \sim b(200,0.6)$.

设至少要供给这个车间 r 千瓦电才能以99.9%的概率保证这个车间正常生产。由题意有

$$P\{X \le r\} \ge 0.999$$

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

由棣莫弗-拉普拉斯定理

$$\frac{X-np}{\sqrt{npq}} \overset{\text{近似地}}{\sim} N(0,1)$$

$$P\{X \le r\}$$

$$\approx \Phi(\frac{r-200\times0.6}{\sqrt{200\times0.6\times0.4}}) - \Phi(\frac{-200\times0.6}{\sqrt{200\times0.6\times0.4}})$$

$$= \Phi(\frac{r-120}{\sqrt{48}}) - \Phi(-17.32) \approx \Phi(\frac{r-120}{\sqrt{48}}) \ge 0.999,$$

查表得 $\frac{r-120}{\sqrt{48}} \ge 3.1$， 所以 $r \ge 141.$

即供给141千瓦电就能以99.9%的概率保证这个车间正常生产。

例2 系统由100个相互独立起作用的部件组成，每个部件的损坏率为0.1。系统要正常工作，至少有85个部件正常工作，求系统正常工作的概率。

解：设 X 是损坏的部件数，则 $X \sim b(100, 0.1)$。

则整个系统能正常工作当且仅当 $X \le 15.$

由德莫佛-拉普拉斯定理有

$$P\{X \le 15\} = P\left\{\frac{X - 100 \times 0.1}{\sqrt{100 \times 0.1 \times 0.9}} \le \frac{15 - 100 \times 0.1}{\sqrt{100 \times 0.1 \times 0.9}}\right\}$$

$$\approx \Phi\left(\frac{15 - 100 \times 0.1}{\sqrt{100 \times 0.1 \times 0.9}}\right) = \Phi\left(\frac{5}{3}\right) = 0.952.$$

例3 一生产线生产的产品成箱包装，每箱的重量是随机的。假设每箱平均重50千克，标准差为5千克。若用最大载重量为5吨的汽车承运，试利用中心极限定理说明每辆车最多可以装多少箱，才能保证不超载的概率大于0.977。

解： 设最多可装 n 箱能保证不超载的概率大于0.977。

第 i 箱重量为 X_i 千克，$i=1,\cdots,n.$

则 $E(X_i)=50,\quad D(X_i)=25,\quad i=1,\cdots,n$

且 $$P\{\sum_{i=1}^{n} X_i \le 5000\} > 0.977$$

由独立同分布的中心极限定理有

例3（续）

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

$\mu = 50, \quad \sigma = 5$

$$P\{\sum_{i=1}^{n} X_i \le 5000\} = P\left\{\frac{\sum_{i=1}^{n} X_i - 50n}{5\sqrt{n}} \le \frac{5000-50n}{5\sqrt{n}}\right\}$$

$$\approx \Phi\left(\frac{5000-50n}{5\sqrt{n}}\right) = \Phi\left(\frac{1000-10n}{\sqrt{n}}\right) > 0.977$$

则 $\frac{1000-10n}{\sqrt{n}} > 2, \quad 100n^2 - 20000n + 1000^2 > 4n,$

解得 $n > 102.02$或$n < 98.02$, 由题意知 $n = 98.$

因此最多可装 98 箱保障不超载的概率大于0.977。

小结

中心极限定理

- 独立同分布中心极限定理
$$\begin{cases} E(X_k)=\mu,\quad D(X_k)=\sigma^2 \\ \Rightarrow \sum_{k=1}^{n} X_k \overset{\text{近似地}}{\sim} N(n\mu, n\sigma^2) \end{cases}$$

- 棣莫弗－拉普拉斯中心极限定理
$$\begin{cases} \eta_n \sim N(n,p) \\ \Rightarrow \eta_n \overset{\text{近似地}}{\sim} N(np, np(1-p)) \end{cases}$$

- 李雅普诺夫中心极限定理
$$\begin{cases} E(X_k)=\mu_k, D(x_k)=\sigma_k^{\ 2} \\ \Rightarrow \sum_{k=1}^{n} X_k \overset{\text{近似地}}{\sim} N(\sum_{k=1}^{n}\mu_k, B_n^{\ 2}) \end{cases}$$

注: 随机变量 $X_1, X_2, \cdots$ 是相互独立的 .

作业:

P126-127:　5，　7，14