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

§1 大数定律

- 大数定律的定义
- 切比晓夫大数定律
- 贝努里大数定律
- 辛钦大数定律

问题：测量一个工件时，由于测量具有误差，为什么以各次的平均值来作为测量的结果？而且只要测量的次数足够多，总可以达到要求的精度？

这里反映了什么样的客观统计规律呢？

我们把这问题给出数学表达：如果工件的真值为 a，

第n次测量误差为ξ_n，则$\{\xi_n\}$就是一个独立同分布，

均值为0的随机变量序列，第n次测量值就是$X_n = a + \xi_n$，

$\{X_n\}$就是均值为a的随机变量序列。

测量的经验就是：

当n充分大时，n次的平均测量值

$$\bar{X} = \frac{1}{n}(X_1 + \cdots X_n)$$

应该和真值a很接近。

即大量测量值的算术平均值具有稳定性。

这就是大数定律所阐述的。

一、定义

定义1

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

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

定义2 设$X_1, \cdots X_n \cdots$是随机变量序列，对任意 $\varepsilon > 0$,有

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

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

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

则称$\{X_n\}$服从大数定律。

定理1 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2（切比晓夫大数定律）(Chebyshev 大数定律)

设随机变量 $X_1, \cdots X_n \cdots$ 相互独立,且具有相同的数学期望及方差，$EX_k = \mu, DX_k = \sigma^2, k = 1,2,\cdots,$

则 $\{X_n\}$ 服从大数定律,即对任意的 $\varepsilon > 0$ 有

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

$$\text{或} \lim_{n\to\infty} P\{|\frac{1}{n}\sum_{k=1}^{n} X_k - \mu| \geq \varepsilon\} = 0.$$

证：

$$E\left(\frac{1}{n}\sum_{k=1}^{n}X_k\right)=\frac{1}{n}\left(\sum_{k=1}^{n}EX_k\right)=\frac{1}{n}\left(\sum_{k=1}^{n}\mu\right)=\mu$$

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

由切比晓夫不等式得： $P\{|X-\mu|\geq\varepsilon\}\leq\sigma^2/\varepsilon^2$

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

当$n\to\infty$时， $P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}X_k-\mu\right|\geq\varepsilon\right\}=0$

思考：能否把定理中独立性条件减弱？

定理3（贝努里大数定律）（Bernoulli大数定律）

设n_A为n重贝努里试验中事件A发生的次数，

p是事件A发生的概率，则：对任意的$\varepsilon > 0$有

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

证：令

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

则：$n_A=\sum_{k=1}^{n}X_k,$

$X_1,\cdots,X_n$ 相互独立同服从于两点分布.

且 $EX_k=p$，$DX_k=p(1-p)$，$k=1,2,\cdots,n,$

由定理2有 $\lim\limits_{n\to\infty}P\{|\frac{1}{n}\sum_{i=1}^{n}X_i-p|<\varepsilon\}=1,$

即 $\lim\limits_{n\to\infty}P\{|\frac{n_A}{n}-p|<\varepsilon\}=1.$

该定理给出了频率的稳定性的严格的数学意义。

定理4（辛钦大数定律）

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 $EX_k=\mu,\ \ k=1,2,\cdots,n,\cdots$

则：对任意的 $\varepsilon>0$ 有

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

注：贝努里大数定律是辛钦大数定律的特殊情况。

思考：比较辛钦大数定律与切比晓夫大数定律条件的差别及强弱。

仅知随机变量的期望与方差并不能确定其分布

X	-1	0	1
P	0.1	0.8	0.1

$$E(X)=0, D(X)=0.2$$

与

Y	-2	0	2
P	0.025	0.95	0.025

$$E(Y)=0, D(Y)=0.2$$

有相同的期望方差但是分布却不相同

§2 中心极限定理

- 定义
- 独立同分布的中心极限定理
- 李雅普诺夫定理
- 德莫佛-拉普拉斯定理
- 用频率估计概率时误差的估计

一、定义

设$X_1,\cdots,X_n,\cdots$独立的随机变量序列，且EX_k, DX_k存在，令：

$$Z_n=(\sum_{k=1}^{n}X_k-\sum_{k=1}^{n}EX_k)/\sqrt{\sum_{k=1}^{n}DX_k},$$

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4 （独立同分布的中心极限定理）

设$X_1,\cdots,X_n,\cdots$独立同分布的随机变量序列，且

$$EX_k=\mu,\ \ DX_k=\sigma^2\neq 0,(k=1,2,\cdots)$$

则$\{X_n\}$服从中心极限定理即：

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

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。

定理5 （李雅普诺夫定理）(Liapunov定理)

设$X_1,\cdots,X_n,\cdots$相互独立，且$EX_k=\mu_k$，

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

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：

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

定理6（德莫佛-拉普拉斯定理）(De Moivre--Laplace)

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

对任意 $x \in R_1$ 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**

$EX_k = p$, $DX_k = pq$

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

由定理1有结论成立。

推论：$$\lim_{n\to\infty} P\{\frac{\eta_n - np}{\sqrt{npq}} \le x\} = \Phi(x)$$

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。

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 ~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

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

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

而 $P\{X \le r\}$

$$\approx \Phi(\frac{r-200\times 0.6}{\sqrt{200\times 0.6\times 0.4}}) - \Phi(\frac{-200\times 0.6}{\sqrt{200\times 0.6\times 0.4}})$$

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

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。

用频率估计概率时误差的估计:

由上面的定理知

$$\lim_{n\to\infty} P\{\frac{\eta_n - np}{\sqrt{npq}} \le x\} = \Phi(x)$$

$$P\left\{\left|\frac{\eta_n}{n} - p\right| < \varepsilon\right\} = P\left\{\left|\eta_n - np\right| < n\varepsilon\right\}$$

$$= P\left\{-\varepsilon\sqrt{\frac{n}{pq}} < \frac{\eta_n - np}{\sqrt{npq}} < \varepsilon\sqrt{\frac{n}{pq}}\right\}$$

$$\approx \Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right) - \Phi\left(-\varepsilon\sqrt{\frac{n}{pq}}\right) = 2\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right) - 1$$

用这个关系式可解决许多计算问题。

第一类问题是 已知n, p, ε， 求概率

$$P\left\{\left|\frac{\eta_n}{n} - p\right| < \varepsilon\right\};$$

$$P\left\{\left|\frac{\eta_n}{n} - p\right| < \varepsilon\right\} = 2\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right) - 1$$

第二类问题是要使$\dfrac{\eta_n}{n}$与p的差异不大于定数ε的概率不小于预先给定的数β，问最少应做多少次试验？

这时只需求满足下式的最小的 n,

$$2\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right) - 1 \geq \beta$$

第三类问题是已知n，p 及 β, 求 ε.

例2 今从良种率为1/6的种子中任取6000粒，问能以0.99的概率保证在这6000粒种子中良种所占的比例与1/6的差的绝对值不超过多少？相应的良种粒数在哪个范围内？

解：设良种数为X，则$X \sim B(n,p)$，

其中$n=6000,\ p=1/6.$

设不超过的界限为α，则应有：

$$P\left\{\left|\frac{X}{6000}-\frac{1}{6}\right|\le\alpha\right\}=0.99$$

由德莫佛-拉普拉斯定理

$$P\left\{\left|\frac{X}{6000}-\frac{1}{6}\right|\le\alpha\right\}$$

$n = 6000, p = 1/6.$

$$\lim_{n\to\infty} P\{\frac{\eta_n - np}{\sqrt{npq}} \le x\} = \Phi(x)$$

$$=P\left\{\left|\frac{X-6000\times1/6}{\sqrt{6000\times1/6\times5/6}}\right|\le\frac{6000\alpha}{\sqrt{6000\times1/6\times5/6}}\right\}$$

$$\approx\Phi\left[\frac{6000\alpha}{\sqrt{6000\times1/6\times5/6}}\right]-\Phi\left[-\frac{6000\alpha}{\sqrt{6000\times1/6\times5/6}}\right]$$

$$\approx2\Phi\left[\frac{6000\alpha}{\sqrt{6000\times1/6\times5/6}}\right]-1$$

故近似地有 $2\Phi\left[\frac{6000\alpha}{\sqrt{6000\times1/6\times5/6}}\right]-1=0.99,$

即 $$\Phi\left[\frac{6000\alpha}{\sqrt{6000\times1/6\times5/6}}\right]=0.995,$$

查表得 $$\frac{6000\alpha}{\sqrt{6000\times1/6\times5/6}}=2.58,$$

解得 $\alpha=0.0124.$

$$\left|\frac{X}{6000}-\frac{1}{6}\right|\le\alpha$$

良种粒数X的范围为

$$(1/6-0.0124)\times6000\le X\le(1/6+0.0124)\times6000,$$

即 $925\le X\le1075.$

例3 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*~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4 一加法器同时收到20个噪声电压，$V_k\ (k=1,2,\cdots,20)$ 设它们是互相独立的随机变量，且都在区间(0,10)上服从均匀分布，记

$$V=\sum_{k=1}^{20}V_k$$

求 $P\{V>105\}$的近似值。

解： $EV_k=5,\ DV_k=10^2/12,\ (k=1,2,\cdots,20),$

$$P\{V>105\}=P\left\{\frac{V-20\times5}{\sqrt{20}\times\sqrt{10^2/12}}>\frac{105-20\times5}{\sqrt{20}\times\sqrt{10^2/12}}\right\}$$

$$=P\left\{\frac{V-100}{\sqrt{20}\times(10/\sqrt{12})}>0.387\right\}\approx1-\Phi(0.387)=0.348$$

例5 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，才能保障不超载的概率大于0.977。

解：设最多可装 n 箱，保障不超载的概率大于0.977。

第i箱重量为X_i吨，$i=1,\cdots,n.$

则 $EX_i=50,\quad DX_i=25,\quad i=1,\cdots,n$

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

由中心极限定理有

例5（续）

$$\lim_{n\to\infty} P\{\frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\sigma} \le x\} = \Phi(x)$$

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

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

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

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

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。

第五章 小 结

主要内容：

1) 大数定律的定义，贝努里、辛钦大数定律，切比晓夫大数定律；

2) 中心极限定理的定义，独立同分布的中心极限理和德莫佛-拉普拉斯定理及应用。

要求：

1） 了解大数定律的意义和内容，理解贝努里、辛钦大数定律，了解切比晓夫大数定律。

2） 理解中心极限定理的含义及其客观背景，要掌握独立同分布的中心极限定理和德莫佛-拉普拉斯定理，会利用中心极限定理解决一般实际应用问题。