第五章　大数定律与中心极限定理

本章要解决的问题

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

3. 为何正态分布在概率论中占有极其重要的地位？
4. 大样本统计推断的理论基础是什么？

中心极限定理

§5.1 大数定律

贝努里（Bernoulli） 大数定律

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

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

证 引入随机变量序列$\{X_k\}$

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

设 $P(X_k = 1) = p$, 则 $E(X_k) = p, D(X_k) = pq$

$X_1, X_2, \cdots, X_n$ 相互独立， $n_A = \sum_{k=1}^{n} X_k$

记 $Y_n = \frac{1}{n}\sum_{k=1}^{n} X_k$， $E(Y_n) = p$， $D(Y_n) = \frac{pq}{n}$

由Chebyshev 不等式

$$0 \leq P\left(\left|\frac{n_A}{n} - p\right| \geq \varepsilon\right)$$

$$= P\left(\left|\frac{\sum_{k=1}^{n} X_k}{n} - E(X_k)\right| \geq \varepsilon\right)$$

$$= P(|Y_n - E(Y_n)| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} \cdot \frac{pq}{n}$$

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

贝努里（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.

定义 设 $Y_1, Y_2, \cdots, Y_n, \cdots$ 是一系列随机变量，a 是一常数，若 $\forall \varepsilon > 0$ 有

$$\lim_{n\to\infty} P(|Y_n - a| \geq \varepsilon) = 0$$

(或 $\lim_{n\to\infty} P(|Y_n - a| < \varepsilon) = 1$)

则称随机变量序列 $Y_1, Y_2, \cdots, Y_n, \cdots$ 依概率收敛于常数 a，记作

$$Y_n \xrightarrow[n\to\infty]{P} a$$

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

在 Bernoulli 定理的证明过程中， Y_n 是相互独立的服从 0-1分布的随机变量序列 $\{X_k\}$ 的算术平均值, Y_n 依概率收敛于其数学期望 p .

结果同样适用于服从其它分布的独立随机变量序列

Chebyshev 大数定律

设随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 两两不相关，(指任意给定 $n > 1$, $X_1, X_2, \cdots, X_n$ 两两不相关)，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 $$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \mu\right| < \varepsilon\right) = 1$$

定理的意义：

具有相同数学期望和方差的不相关随机变量序列的算术平均值依概率收敛于数学期望.

当 n 足够大时，算术平均值几乎就是一个常数，可以用算术平均值近似地代替数学期望.

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

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

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

注2： $X_1, X_2, \cdots, X_n, \cdots$ 两两不相关的条件可以去掉，代之以

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

Markov大数定律

设 $X_1, X_2, \cdots, X_n, \cdots$ 为一随机序列，满足

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

则

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

辛钦大数定律

设 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立，服从同一分布，且具有数学期望 $E(X_k) = \mu$, $k = 1,2,\ldots$, 则对任意正数 $\varepsilon > 0$

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

§5.2 中心极限定理

定理1　独立同分布的中心极限定理

设随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立，服从同一分布，且有期望和方差：

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

则对于任意实数 x，

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

注：　记

$$Y_n = \frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\sigma}$$

则 Y_n 为 $\sum_{k=1}^{n} X_k$ 的标准化随机变量.

$$\lim_{n\to\infty} P(Y_n \le x) = \Phi(x)$$

即 n 足够大时，Y_n 的分布函数近似于标准正态随机变量的分布函数

$$Y_n \overset{\text{近似}}{\sim} N(0,1)$$

$$\sum_{k=1}^{n} X_k = \sqrt{n}\sigma Y_n + n\mu \quad \text{近似服从} \quad N(n\mu, n\sigma^2)$$

定理2　德莫佛 — 拉普拉斯中心极限定理（DeMoivre-Laplace）

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)

注1　n 较小时，例 $n<10$ 时直接用二项分布公式计算.

注2　n 较大而 p 较小 (或 1- p 较小), 近似用poisson分布近似计算. (实际要求np 适中.)

注3　n 较大, $0.1<p<0.9$, 或 $n>100$, $p<0.1$ 时用正态分布近似代替.即

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

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

例　某射手射靶，得十分的概率为0.5，得九分的概率为0.3，得八分的概率0.1得七分的概率为0.05，得六分的概率为0.05. 现独立的射击100次，用切比雪夫不等式和中心极限定理估计总分介于900分与930分之间的概率。

解：　设 X_i 表示射击手第 i 次得分

X_k	10	9	8	7	6
p_k	0.5	0.3	0.1	0.05	0.05

$$E(X_i)=9.15 \quad E(X_i^2)=84.95 \quad D(X_i)=1.2275$$

$$X=\sum_{i=1}^{100} X_i \quad E(X)=915 \quad D(X)=122.75$$

$$X \overset{\text{近似}}{\sim} N(915, 122.75)$$

$$P(900<X<930) \approx \Phi\left(\frac{930-915}{\sqrt{122.75}}\right) - \Phi\left(\frac{900-915}{\sqrt{122.75}}\right) \approx 0.823$$

例　某车间有200台车床，每台独立工作，开工率为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

设 X 为200 台车床的开工数.

$X \sim B(200, 0.6)$，　$X \sim N(120, 48)$ (近似)

问题转化为求 a，使

$$P(0 \le rX \le a) \ge 99.9\%$$

由于将 X 近似地看成正态分布，故

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

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

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

反查标准正态函数分布表，得 $\Phi(3.09) = 99.9\%$

令 $\dfrac{\frac{a}{r}-120}{\sqrt{48}} \ge 3.09$ 解得 $a \ge (3.09\sqrt{48}+120)r \approx 141r$

例 检查员逐个地检查某种产品，每检查一只产品需要用10秒钟．但有的产品需重复检查一次，再用去10秒钟．假设产品需要重复检查的概率为 0.5, 求检验员在 8 小时内检查的产品多于1900个的概率.

解 检验员在 8 小时内检查的产品多于1900个即检查1900个产品所用的时间小于 8 小时.

设 X 为检查1900 个产品所用的时间(单位：秒)

设 X_k 为检查第 k 个产品所用的时间(单位：秒)， k = 1,2,…,1900

X_k	10	20
P	0.5	0.5

$$E(X_k) = 15, \quad D(X_k) = 25$$

$X_1, X_2, \cdots, X_{1900}$ 相互独立,且同分布, $X = \sum_{k=1}^{1900} X_k$

$$E(X) = 1900 \times 15 = 28500$$

$$D(X) = 1900 \times 25 = 47500$$

$$X \overset{\text{近似}}{\sim} N(28500, 47500)$$

$$P(10 \times 1900 \le X \le 3600 \times 8)$$

$$= p(19000 \le X \le 28800)$$

$$\approx \Phi\left(\frac{28800-28500}{\sqrt{47500}}\right) - \Phi\left(\frac{19000-28500}{\sqrt{47500}}\right)$$

$$\approx \Phi(1.376) - \Phi(-43.589)$$

$$\approx 0.9162$$

解法二

$\dfrac{X-19000}{10}$ ——1900个产品中需重复检查的个数

$$\frac{X-19000}{10} \sim B(1900, 0.5) \overset{\text{近似}}{\sim} N(950, 475)$$

$$P(10 \times 1900 \le X \le 3600 \times 8)$$

$$= P(19000 \le X \le 28800)$$

$$= P\left(0 \le \frac{X-19000}{10} \le \frac{28800-19000}{10}\right)$$

$$= P\left(0 \le \frac{X-19000}{10} \le 980\right)$$

$$\approx \Phi\left(\frac{980-950}{\sqrt{475}}\right) - \Phi\left(\frac{0-950}{\sqrt{475}}\right)$$

$$\approx \Phi(1.376) - \Phi(-43.589)$$

$$\approx 0.9162$$

例 **据调查，某小区中一个家庭无车、有1 辆车、有2 辆车的概率分别为0.05，0.8，0.15。若该小区共有400 个家庭，试用中心极限定理计算：**

（1）400 个家庭拥有车辆总数超过450 的概率；

（2）只有1辆车的家庭数不多于340 的概率。

解：（1）设$X_i(i=1,2,\cdots,400)$表示第i个家庭拥有的车辆数，则

$$X_i \sim \begin{pmatrix} 0 & 1 & 2 \\ 0.05 & 0.8 & 0.15 \end{pmatrix},$$

且$E(X_i)=1.1, D(X_i)=0.19. X_1,\cdots,X_{400}$相互独立。

令$X=\sum_{i=1}^{400} X_i$,则X近似服从$N(440,76),$

$$P(X>450)=P(\frac{X-440}{\sqrt{76}}>\frac{10}{\sqrt{76}})\approx 1-\Phi(1.15)=0.1251$$

(2)令Z表示只有一辆车的家庭数，从而$Z \sim B(400,0.8)$

$E(Z)=np=320, D(Z)=np(1-p)=64,$

Z近似服从$N(320,64),$

$$P(Z\le 340)=\Phi(\frac{340-320}{8})=\Phi(2.5)=0.9938$$

中心极限定理的意义

在实际问题中，若某随机变量可以看作是有相互独立的大量随机变量综合作用的结果，每一个因素在总的影响中的作用都很微小，则综合作用的结果服从正态分布.