第六章　大数定律与中心极限定理

本章要解决的问题

答复

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

6.1 契比雪夫不等式

马尔可夫（Markov）不等式

设非负随机变量 X 的期望 $E(X)$存在，则对于任意实数 $\varepsilon > 0$,

$$P(X \geq \varepsilon) \leq \frac{E(X)}{\varepsilon}$$

证 仅证连续型随机变量的情形

$$P(X \geq \varepsilon) = \int_{\varepsilon}^{+\infty} f(x)dx \leq \int_{\varepsilon}^{+\infty} \frac{x}{\varepsilon} f(x)dx$$

$$\leq \frac{1}{\varepsilon}\int_{0}^{+\infty} xf(x)dx = \frac{E(X)}{\varepsilon}$$

推论 1

设随机变量 X 的k阶绝对原点矩 $E(|X|^k)$存在，则对于任意实数 $\varepsilon > 0$,

$$P(|X| \geq \varepsilon) \leq \frac{E(|X|^k)}{\varepsilon^k}$$

证

由马尔可夫（Markov）不等式有

$$P(|X| \geq \varepsilon) = P(|X|^k \geq \varepsilon^k)$$

$$\leq \frac{E(|X|^k)}{\varepsilon^k}$$

推论 2 ——切贝雪夫（ chebyshev ）不等式

设随机变量 X 的方差 $D(X)$ 存在，
则对于任意实数 $\varepsilon > 0$,

$$P(|X - E(X)| \geq \varepsilon) \leq \frac{D(X)}{\varepsilon^2}$$

或

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

证

由马尔可夫（Markov）不等式有：

$$P(|X-E(X)|\geq\varepsilon)$$

$$=P(|X-E(X)|^2\geq\varepsilon^2)$$

$$\leq\frac{E\left(|X-E(X)|^2\right)}{\varepsilon^2}$$

$$=\frac{D(X)}{\varepsilon^2}$$

例 设每次试验中，事件 A 发生的概率为 0.75, 试用 Chebyshev 不等式估计，n 多大时, 才能在 n 次独立重复试验中, 事件 A 出现的频率在$0.74 \sim 0.76$ 之间的概率大于 0.90?

解 设 X 表示 n 次独立重复试验中事件 A 发生的次数 , 则

$$X \sim B(n, 0.75)$$

$$E(X) = 0.75\ n,$$

$$D(X) = n \cdot 0.75 \cdot (1 - 0.75) = 0.1875\ n$$

要使 $P\left(0.74 < \dfrac{X}{n} < 0.76\right) \geq 0.90$ ，求 n

即 $P(0.74n < X < 0.76n) \geq 0.90$

即 $P(|X - 0.75n| < 0.01n) \geq 0.90$

由 Chebyshev 不等式，$\varepsilon = 0.01n$,故

$$P(|X - 0.75n| < 0.01n) \geq 1 - \frac{0.1875n}{(0.01n)^2}$$

令

$$1 - \frac{0.1875n}{(0.01n)^2} \geq 0.90$$

解得 $n \geq 18750$

6.2 大数定律

1、Chebyshev大数定律（定理）

$X_1, X_2, \cdots, X_n, \cdots$ 是相互独立的随机变量序列，

每一 X_k 都有有限的方差，且有公共上界，可设

$$D(X_k) = \sigma_k^2 \le c, \quad k = 1, 2, \cdots$$

则有

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

证明：

对随机变量 $\frac{1}{n}\sum_{i=1}^{n}X_i$ ，

$$E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}E\left(X_i\right)$$

$$D\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}D\left(X_i\right)\ \le\frac{1}{n^2}\cdot nc$$

$$=\frac{c}{n}$$

利用契比雪夫不等式，有：

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i - E(\frac{1}{n}\sum_{i=1}^{n}X_i)\right| \geq \varepsilon\right) \leq \frac{D\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)}{\varepsilon^2} \leq \frac{c}{n\varepsilon^2}$$

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

或者

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

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

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

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

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

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

契比雪夫大数定律即 $\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow[n\to\infty]{P} \frac{1}{n}\sum_{k=1}^{n} E(X_k)$

2、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 $\varepsilon > 0$

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

证明：

在Chebyshev大数定律

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

中

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

代入即得

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

即 $\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow[n\to\infty]{P} \mu$

3、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，

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

在辛钦大数定律

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

中

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

$$\mu=E(X_k)=p$$

代入即得

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

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

§6.3 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，

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

即：

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

即 n 足够大时，$\dfrac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\sigma}$ 的分布函数近似于标准正态随机变量的分布函数。

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

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

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

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

说明：若定理1中 $X_1, X_2, \cdots, X_n, \cdots$

独立同分布为0-1分布，$P(X_i=1) = p$

则：$Y_n = \sum_{i=1}^{n} X_i \sim B(n, p)$

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.

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)，$k = 1,2,\ldots,1900$

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

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

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

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

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

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

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

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

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

$$\approx 0.9162$$