

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？
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？

大数
定律

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

$$\begin{aligned} P(X \geq \varepsilon) &= \int_{\varepsilon}^{+\infty} f(x)dx \leq \int_{\varepsilon}^{+\infty} \frac{x}{\varepsilon} f(x)dx \\ &\leq \frac{1}{\varepsilon} \int_0^{+\infty} xf(x)dx = \frac{E(X)}{\varepsilon} \end{aligned}$$

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：

$$\begin{aligned} & P(|X - E(X)| \geq \varepsilon) \\ &= P(|X - E(X)|^2 \geq \varepsilon^2) \\ &\leq \frac{E(|X - E(X)|^2)}{\varepsilon^2} \\ &= \frac{D(X)}{\varepsilon^2} \end{aligned}$$

6.2 大数定律

1、Chebyshev大数定律（定理）

$X_1, X_2, \dots, X_n, \dots$ 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

$$D(X_k) = \sigma_k^2 \leq c, \quad k = 1, 2, \dots$$

则有

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

证明：

对随机变量 $\frac{1}{n} \sum_{i=1}^n X_i$,

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

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

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

利用契比雪夫不等式，有：

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

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

或者

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

定义 设 $Y_1, Y_2, \dots, Y_n, \dots$ 是一系列随机变量 ,

a 是一常数 , 若 $\forall \varepsilon > 0$ 有

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

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

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

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

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

2、辛钦大数定律

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

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

证明：

在Chebyshev大数定律

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

中

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

代入即得

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

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

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

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

或

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

证 引入随机变量序列 $\{X_k\}$

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

设 $P(X_k = 1) = p$, 则

$$E(X_k) = p, D(X_k) = pq$$

X_1, X_2, \dots, 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

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

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

§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 $X_1, X_2, \dots, X_n, \dots$ 相互独立，服从同一分布，且有期望和方差：

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

则对于任意实数 x ，

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

即：

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

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

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

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

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

设 $Y_n \sim B(n, p)$, $0 < p < 1$, $n = 1, 2, \dots$

则对任一实数 x , 有

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

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

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 $a < b$,

$$\lim_{n \rightarrow \infty} P\left(a < \frac{Y_n - np}{\sqrt{np(1-p)}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{t^2}{2}} dt$$

$$Y_n \sim N(np, np(1-p)) \text{ (近似)}$$

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。

中心极限定理的应用

例1 设有一大批种子，其中良种占 $1/6$. 试估计在任选的6000粒种子中，良种所占比例与 $1/6$ 比较上下不超过 1% 的概率.

解 设 X 表示6000粒种子中的良种数，则

$$X \sim B(6000, 1/6)$$

$$E(X) = 1000, D(X) = \frac{5000}{6}$$

$$X \stackrel{\text{近似}}{\sim} N\left(1000, \frac{5000}{6}\right)$$

$$P\left(\left|\frac{X}{6000} - \frac{1}{6}\right| \leq 0.01\right) = P(|X - 1000| \leq 60) = P(940 \leq X \leq 1060)$$

$$\approx \Phi\left(\frac{1060 - 1000}{\sqrt{5000/6}}\right) - \Phi\left(\frac{940 - 1000}{\sqrt{5000/6}}\right)$$

$$= \Phi\left(\frac{60}{\sqrt{5000/6}}\right) - \Phi\left(\frac{-60}{\sqrt{5000/6}}\right)$$

$$= 2\Phi\left(\frac{60}{\sqrt{5000/6}}\right) - 1 \approx 0.9624$$

比较几个近似计算的结果

用二项分布(精确结果)

$$X \sim B(6000, 1/6)$$

$$\begin{aligned} P\left(\left|\frac{X}{6000} - \frac{1}{6}\right| \leq 0.01\right) &= P(940 \leq X \leq 1060) \\ &= \sum_{k=940}^{1060} C_{6000}^k \left(\frac{1}{6}\right)^k \left(1 - \frac{1}{6}\right)^{6000-k} \\ &\approx 0.9590 \end{aligned}$$

用Poisson 分布

$$\lambda = np = 6000 \cdot \frac{1}{6} = 1000$$

$$\begin{aligned} P\left(\left|\frac{X}{6000} - \frac{1}{6}\right| \leq 0.01\right) &= P(940 \leq X \leq 1060) \\ &= \sum_{k=940}^{1060} \frac{1000^k \cdot e^{-1000}}{k!} \end{aligned}$$

$$\approx 0.9379$$

用Chebyshev 不等式

$$P\left(\left|\frac{X}{6000} - \frac{1}{6}\right| < 0.01\right) \geq 0.7685$$

用中心极限定理

$$P\left(\left|\frac{X}{6000} - \frac{1}{6}\right| < 0.01\right) \approx 0.9624$$