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 比较上下小于1%的概率.

解 设 X 表示 6000 粒种子中的良种数 ,

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

$$E(X) = np = 6000 \cdot \frac{1}{6} = 1000,$$

$$D(X) = np(1-p) = 6000 \cdot \frac{1}{6} \cdot \left(1 - \frac{1}{6}\right) = \frac{5000}{6}$$

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

$$= P(|X - 1000| < 60) \quad \geq 1 - \frac{\frac{5000}{6}}{60^2} = \frac{83}{108} = 0.7685$$

实际精确计算:

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

$$=\sum_{k=941}^{1059} C_{6000}^{k}\left(\frac{1}{6}\right)^{k}\left(\frac{5}{6}\right)^{6000-k}=0.959036$$

用Poisson 分布近似计算:

$$\lambda=np=1000$$

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

$$=\sum_{k=941}^{1059}\frac{1000^{k}e^{-1000}}{k!}=0.937934$$

例2 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$0.74\sim0.76$ 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

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

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

$$D(X)=n\cdot0.75\cdot(1-0.75)=0.1875\ n$$

要使 $P\left(0.74<\dfrac{X}{n}<0.76\right)\geq0.90$，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:

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

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

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

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

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

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

辛钦大数定律的意义：

具有相同分布的独立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.

即回答了：为何能以样本均值作为总体期望的估计？

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

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.

即回答了：为何能以某事件发生的频率作为该事件的 概率的估计?