第五章　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

5.1 大数定律

一、切比雪夫（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　$$P\{|X - E(X)| \geq \varepsilon\} = \int_{|x-E(X)|\geq\varepsilon} f(x)dx$$

$$\leq \int_{|x-E(X)|\geq\varepsilon} \frac{[x - E(X)]^2}{\varepsilon^2} f(x)dx$$

$$\leq \frac{1}{\varepsilon^2}\int_{-\infty}^{+\infty}[x - E(X)]^2 f(x)dx = \frac{D(X)}{\varepsilon^2}$$

仅证连续型随机变量的情形

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 ,

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

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

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

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

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：

取$\lambda = 1000$

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

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

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 次重复独立试验中, 事件 A 出现的频率在0.74 ~ 0.76 之间的概率大于 0.90?

解 设 X 表示 n 次重复独立试验中事件 A 发生的次数 , 则

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

$$E(X)=0.75n,\ D(X)=0.1875n$$

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 $n\geq 18750$

二、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，则对 $\forall \varepsilon > 0$ 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

$$0 \le P\left\{\left|\frac{n_A}{n} - p\right| \ge \varepsilon\right\}$$

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

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

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

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 $\left\{\left|\frac{n_A}{n} - p\right| \ge \varepsilon\right\}$ 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 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立，

(指任意给定 $n > 1, X_1, X_2, \cdots, X_n$ 相互独立)，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.

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 $\varepsilon > 0$

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

或

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

§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，

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

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

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

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

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

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)

该定理常用的两个公式为：

(1) $$P\{\alpha < Y_n \le \beta\} \approx \Phi\left(\frac{\beta - np}{\sqrt{npq}}\right) - \Phi\left(\frac{\alpha - np}{\sqrt{npq}}\right)$$

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

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3 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

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

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

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

$$P\left\{\left|\frac{X}{6000} - \frac{1}{6}\right| < 0.01\right\} = P\{|X - 1000| < 60\}$$

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

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

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

比较几个近似计算的结果

用二项分布(精确结果) $P\left\{\left|\frac{X}{6000}-\frac{1}{6}\right|<0.01\right\}\approx 0.9590$

用Poisson 分布 $P\left\{\left|\frac{X}{6000}-\frac{1}{6}\right|<0.01\right\}\approx 0.9379$

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

例4 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

$$P\{0\leq rX\leq a\}=99.9\%$$

由于将 X 近似地看成正态分布，故

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

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

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

反查标准正态函数分布表，得

$$\Phi(3.09) = 99.9\%$$

令

$$\frac{\frac{a}{r}-120}{\sqrt{48}} = 3.09$$

解得

$$a = (3.09\sqrt{48}+120)r$$

$$\approx 141r \text{ (千瓦)}$$

例5 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

$\frac{X-19000}{10}$ — 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7 售报员在报摊上卖报，已知每个过路人在报摊上买报的概率为1/3. 令X 是出售了100份报时过路人的数目，求 $P\{280 \le X \le 320\}$.

解 令X_i 为售出了第 $i-1$ 份报纸后到售出第i份报纸时的过路人数，$i = 1,2,\ldots,100$

$$P\{X_i = k\} = p(1-p)^{k-1}\Big|_{p=1/3}, \quad k = 1,2,\cdots$$

(几何分布)

$$E(X_i) = \frac{1}{p}\Big|_{p=1/3} = 3, \quad D(X_i) = \frac{1-p}{p^2}\Big|_{p=1/3} = 6$$

$X_1, X_2, \cdots, X_{100}$ 相互独立，$X = \sum_{k=1}^{100} X_k$

$$E(X) = 300,\ D(X) = 600$$

$X \sim N(300,600)$ (近似)

$$P\{280 \le X \le 320\} \approx \Phi\left(\frac{320-300}{\sqrt{600}}\right) - \Phi\left(\frac{280-300}{\sqrt{600}}\right)$$

$$\approx 2\Phi\left(\frac{20}{\sqrt{600}}\right) - 1$$

$$\approx 2\Phi(0.8165) - 1 \approx 0.5878$$