南京大学
NANJING UNIVERSITY

第五章 大数定律及中心极限定理

要解决的问题：

1. 为何能以某事件发生的频率作为该事件发生概率的估计?
2. 为何能以样本均值作为总体均值的估计?

答案：

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

主要内容

- 预备：切比雪夫不等式
- 5.1 大数定律
- 5.2 中心极限定理

预备：切比雪夫不等式

切比雪夫(Chebyshev)不等式

设随机变量 X 具有数学期望 $EX = \mu$，方差 $DX = \sigma^2$，则对于任意正数 ε,

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

或

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

该不等式可以粗略估计随机变量取值落在期望左右某个范围内的概率!

例： 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 , 则

$$X \sim b\left(6000,\ 1/6\right)$$

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

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

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

实际精确计算：

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

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

用泊松分布近似计算：

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： 设每次试验中，事件 A 发生的概率为 0.75, 试用**切比雪夫不等式**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 ~ 0.76 之间的概率不低于 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

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

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

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**切比雪夫不等式**可得:

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

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

解得 $n \geq 18750$

5.1 大数定律

频率稳定性的试验

德摩根(Morgan)投币

投掷一枚硬币，观察正面向上出现的次数

$n = 2048,\quad n_H = 1061,\quad f_n(H) = \mathbf{0.5181}$

蒲丰(Buffon)投币

$n = 4040,\quad n_H = 2048,\quad f_n(H) = \mathbf{0.5069}$

皮尔逊(Pearson) 投币

$n = 12000,\quad n_H = 6019,\quad f_n(H) = \mathbf{0.5016}$

$n = 24000,\quad n_H = 12012,\quad f_n(H) = \mathbf{0.5005}$

能否从理论上证明频率收敛到概率？

伯努利(Bernoulli)大数定理

设 f_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

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

或

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

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$E(X_k)=p, D(X_k)=p(1-p)$

$X_1, X_2, \quad , X_n$ 是相互独立的，

则 $f_A=\sum_{k=1}^{n}X_k$，且服从$b(n,p)$

记$Y_n=\dfrac{f_A}{n}$，　则$E(Y_n)=p$，　$D(Y_n)=\dfrac{p(1-p)}{n}$

由**切比雪夫不等式**可得：

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

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

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

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

伯努利(Bernoulli)大数定理的意义

事件 A 发生的频率 $\frac{f_A}{n}$ “稳定于” 事件 A 在一次试验中发生的概率 p 是指：

频率 $\frac{f_A}{n}$ 与 p 有较大偏差 $\left(\left|\frac{f_A}{n}-p\right|\geq\varepsilon\right)$ 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 $Y_1, Y_2, \quad , Y_n, \quad$ 是一个随机变量序列，

a 是一常数，若 $\forall \varepsilon > 0$ 有

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

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

则称随机变量序列 $Y_1, Y_2, \quad , Y_n, \quad$ **依概率收敛于 a ，**

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

故 $Y_n = \dfrac{f_A}{n} \xrightarrow[n\to\infty]{P} p$

在伯努利大数定理的证明过程中，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 X_k 的共同的数学期望 p .

注：该结果同样适用于服从其它分布的独立随机变量序列（弱大数定理）.

弱大数定理（伯努利大数定理的推广）

设随机变量序列 $X_1, X_2, \quad , X_n,$ **独立同分布，**

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

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

则 $\forall \varepsilon > 0$ 有

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

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

弱大数定理的意义

具有相同数学期望和方差的独立同分布随机变量序列的**算术平均值依概率收敛于数学期望.**

当 n 足够大时, 算术平均值几乎是一常数.

可被

近似代替

南京大学
NANJING UNIVERSITY

5.2 中心极限定理

定理一

林德伯格-列维中心极限定理
(Lindeberg-levi)

[独立同分布的中心极限定理]

定理三

棣莫弗-拉普拉斯中心极限定理
(De Moivre-Laplace)

[二项分布以正态分布为极限分布]

独立同分布的中心极限定理

设随机变量序列 $X_1, X_2, \quad , X_n,$ **独立同分布,** 且有期望和方差：

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

则对于任意实数 x ,随机变量之和 $\sum_{k=1}^{n} X_k$ 的标准化变量的分布函数满足：

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

注

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

则 Y_n 为 $\sum_{k=1}^{n} X_k$ 的标准化随机变量，上页结论为：

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

即 n 足够大时，Y_n 的分布函数近似为标准正态随机变量的分布函数，则

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

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

棣莫弗—拉普拉斯中心极限定理

（独立同分布中心极限定理的特例）

设 $Y_n \sim b(n, p)$，$0 < p < 1$，$n = 1,2,\ldots$

则对任一实数 x，有

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

即 $Y_n \sim N(np, np(1-p))$ (近似)

二项分布以正态分布为极限分布！

例： 对于一个学生而言，来参加家长会的家长人数是一个随机变量，设一个学生无家长、1 名家长，2 名家长来参加会议的概率分别为 0.05, 0.8, 0.15. 若学校共有 400 名学生，设各学生参加会议的家长人数相互独立，且服从同一分布。

(1) 求参加会议的家长人数 X 超过 450 概率；(2) 求有 1 名家长来参加会议的学生人数不多于 340 的概率。

解： (1) 以 X_k 记第 k 个学生来参加会议的家长人数，k = 1, 2, ……, 400, X_k 的分布律为

X_k	0	1	2
P	0.05	0.8	0.15

易知$E(X_k)=1.1,\quad D(X_k)=0.19,\quad k=1,2,\cdots,400$

由独立同分布中心极限定理可知：

$$X=\sum_{k=1}^{400}X_k \qquad N(400\times1.1,400\times0.19)$$

即$X \qquad N(440,76)$

进而：$\dfrac{X-440}{\sqrt{76}} \qquad N(0,1)$

$$P(X>450)=1-P(X\le 450)$$

$$=1-\mathrm{P}\left(\frac{X-440}{\sqrt{76}}\le\frac{450-440}{\sqrt{76}}\right)$$

$$=1-\Phi(1.147)$$

$$=0.1251$$

(2) 以Y记有一名家长参加会议的学生人数，则

$Y\sim b(400,\ 0.8)$

由棣莫弗拉普拉斯中心极限定理可知：

Y $N(400\times0.8,\ 400\times0.8\times0.2)$, 即$Y$ $N(320,\ 64)$

进而：$\dfrac{Y-320}{8} \quad N(0,1)$

$$P(Y \le 340) = \mathrm{P}\left(\frac{Y-320}{8} \le \frac{340-320}{8}\right)$$

$$= \Phi(2.5)$$

$$= 0.9938$$

例： 设有一批种子，其中良种占 1/6. 试估计在任选的 6000 粒种子中，良种比例与 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 $X \sim b(6000, 1/6)$

由棣莫佛—拉普拉斯中心极限定理，

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

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

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

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

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

比较几个近似计算的结果

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

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

泊松分布 $P\left(\left|\frac{X}{6000}-\frac{1}{6}\right|<0.01\right)\approx 0.9379$

切比雪夫不等式 $P\left(\left|\frac{X}{6000}-\frac{1}{6}\right|<0.01\right)\geq 0.7685$

THE END