# 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

### 重要不等式

设非负 r.v. $X$ 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 r.v.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** ——马尔可夫 ( Markov ) 不等式

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 $\varepsilon > 0$,

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

**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 $\varepsilon > 0$,

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

当 $\varepsilon^2 \leq D(X)$ 无实际意义，

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

**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，

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

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

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

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

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

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

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

要使 $P\left(0.74<\frac{X}{n}<0.76\right)\geq 0.90$ ， 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 $n \geq 18750$

## 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 $\forall \varepsilon > 0$ 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 r.v. 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

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

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

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

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 $n_A/n$ “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.

**定义** 设 $Y_1,Y_2,\cdots,Y_n,\cdots$ 是一系列 r.v.

$a$ 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 r.v. 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 r.v. 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 r.v. 序列

## Chebyshev 大数定律

设 r.v. 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 r.v.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.

数学期望 可被 算术均值 近似代替

**注1** $X_1, X_2, \cdots, X_n, \cdots$ 不一定有相同的数学期望与方差，可设

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

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

**注2** $X_1, X_2, \cdots, X_n, \cdots$ 相互独立的条件可以去掉，代之以

$$\frac{1}{n^2}D\left(\sum_{k=1}^{n}X_k\right)\xrightarrow{n\to\infty}0$$

**注3** 设 r.v.序列 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立具有相同的分布，且

$$E(X_i^k) = \mu_k, \quad i = 1,2,\cdots$$

则 $\forall \varepsilon > 0$ 有

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

记

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k = M_k$$

则

$$M_1 \xrightarrow[n\to\infty]{P} \mu_1$$

$$M_2 \xrightarrow[n\to\infty]{P} \mu_2$$

$$\cdots\cdots\cdots\cdots$$

$$M_k \xrightarrow[n\to\infty]{P} \mu_k$$

若 $g(x_1, x_2, \cdots, x_k)$ 连续，则

$$g(M_1, M_2, \cdots, M_k) \xrightarrow[n\to\infty]{P} g(\mu_1, \mu_2, \cdots, \mu_k)$$

作业 P185 习题五

3　4

第12周

## 问 题

电视台需作节目*A*收视率的调查.每天在播电视的同时,随机地向当地居民打电话询问是否在看电视. 若在看电视,再问是否在看节目*A*. 设回答看电视的居民户数为 *n*. 若要保证以 95%的概率使调查误差在10%之内, *n* 应取多大?

每晚节目$A$播出一小时,调查需同时进行,设每小时每人能调查20户,每户居民每晚看电视的概率为70%,电视台需安排多少人作调查.

又，若使调查误差在 1 % 之内,$n$ 应取多大?