第五章　大数定律及中心极限定理

第五章 大数定律及中心极限定理

- 大数定理
- 中心极限定理

第五章 大数定律及中心极限定理

研究大量的随机现象，常常采用极限形式，由此导致对极限定理进行研究。极限定理的内容很广泛，其中最重要的有两种:**大数定律**和**中心极限定理**

大数定理
- 频率为何可以代替概率
- 样本均值为何可以代替总体期望均值

中心极限定理
- 正态分布为何占有极其重要的地位
- 为何可进行大样本统计推断

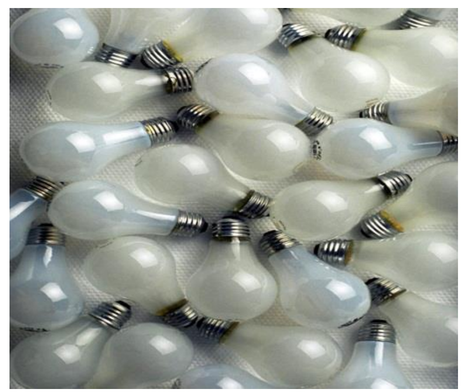
§1 大数定律

一、大数定律的客观背景

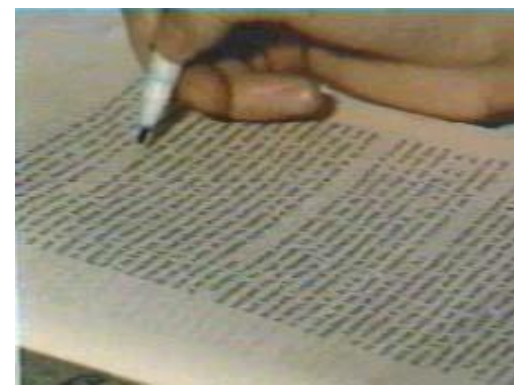
1. 随着试验次数增大, 事件发生的频率稳定于某个常数;
2. 实践中大量测量值的算术平均值也具有稳定性。

人们正是研究了这些稳定性得到了大数定律。

大量抛掷硬币
正面出现频率



生产过程中的废率



字母使用频率

Story

二、弱大数定理（辛钦大数定理）

辛钦

设$X_1, X_2, \ldots X_n$是相互独立，服从同一分布的随机变量序列，且具有数学期望$E(X_k)=\mu\ (k=1,2,\ldots)$

作**前n个变量的算术平均** $\frac{1}{n}\sum_{k=1}^{n} X_k$，则对于任意ε>0，有：

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

设$\{X_n\}$为随机变量序列，a为常数，若任给$\varepsilon>0$, 使得 $\lim_{n\to\infty} P\{|X_n-a|<\varepsilon\}=1$，则称$\{X_n\}$依概率收敛于$a$。可记为$X_n \xrightarrow{P} a$

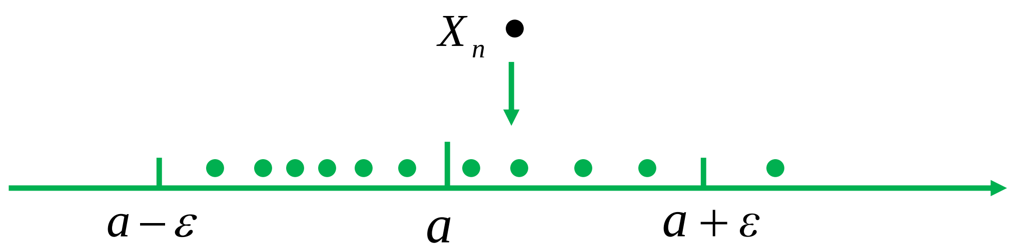
二、弱大数定理（辛钦大数定理）

性质 设 $X_n \xrightarrow{P} a$，$Y_n \xrightarrow{P} b$，又设函数$g(x, y)$ 在 (a, b) 连续, 则

$$g(X_n, Y_n) \xrightarrow{P} g(a, b)$$

$X_n \xrightarrow{P} a$ 意思是：当$n\to\infty$时，X_n落在$(a-\varepsilon, a+\varepsilon)$内的概率越来越大。

$$\forall n_0, n > n_0$$



$X_n \to a$ 意思是：$\forall \varepsilon > 0, \exists n_0$，当$n > n_0$，$|X_n - a| < \varepsilon$即$X_n$落在$(a-\varepsilon, a+\varepsilon)$内。

三、伯努利大数定理

伯努利

设 f_A 是n次独立重复试验中事件A发生的次数，p是事件A在每次试验中发生的概率，则对于任意正数$\varepsilon>0$，有

$$\lim_{n\to\infty} P\left\{\left|\frac{f_A}{n}-p\right|<\varepsilon\right\}=1 \quad \text{或} \lim_{n\to\infty} P\left\{\left|\frac{f_A}{n}-p\right|\geq\varepsilon\right\}=0$$

证明： 设X_i (i=1,2,...)是第i次试验的随机变量

$$X_i=\begin{cases}1, & \text{当第}i\text{次试验事件}A\text{发生}\\ 0, & \text{当第}i\text{次试验事件}A\text{不发生}\end{cases}$$

设f_A 为n次独立重复试验中事件A发生的次数，则有：

$$f_A=\sum_{i=1}^{\infty}X_i,\quad f_n(A)=\frac{f_A}{n}=\frac{1}{n}\sum_{i=1}^{n}X_i=\overline{X}_n$$

§1 大数定律

三、伯努利大数定理

随机变量X_i $(i=1,2,\ldots)$相互独立同分布且服从参数p的二项分布

$$E(X_1)=E(X_2)=\cdots E(X_n)=p,$$
$$D(X_1)=D(X_2)=\cdots D(X_n)=p(1-p);$$

由辛钦大数定理得：

$$\lim_{n\to\infty}P\left\{\left|\frac{f_A}{n}-p\right|<\varepsilon\right\}=1 \quad 或 \lim_{n\to\infty}P\left\{\left|\frac{f_A}{n}-p\right|\geq\varepsilon\right\}=0$$

伯努利大数定律表明，当重复试验次数n充分大时，事件A发生的频率f_A/n与事件A的概率p**有较大偏差的概率很小。**

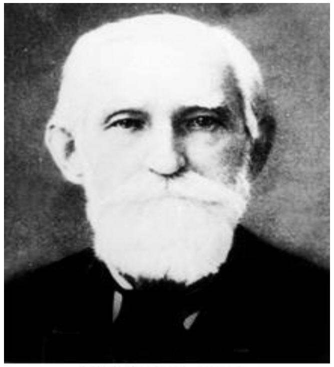
三、伯努利大数定理

- 伯努利大数定理是最早的一个大数定律，它**刻画了频率的稳定性**

- 当独立试验次数n很大时，可以用事件A发生的**频率$f_n(A)$**来**估算**一次试验时事件A发生的**概率**。

例　估计某种商品的不合格率时，可以从这种商品中抽取n件进行测试，当n足够大时，不合格产品的频率f_n可以作为该商品的不合格率(概率)P的估计值。

四、切比雪夫大数定理



切比雪夫，П. Л.

切比雪夫

设 $X_1, X_2, \ldots$是相互不相关的随机变量序列，它们都有有限的方差，并且方差有共同的上界，即$D(X_i) \le c$，$i=1, 2, \ldots$，则对任给$\varepsilon>0$, 有

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} E(X_i)\right| < \varepsilon\right\} = 1$$

切比雪夫大数定理**表明**

独立随机变量序列$\{X_n\}$，如果方差有共同的上界，则$\frac{1}{n}\sum_{i=1}^{n} X_i$与其数学期望$\frac{1}{n}\sum_{i=1}^{n} E(X_i)$偏差很小。**概率接近于1。**

§1 大数定律

四、切比雪夫大数定理

在定理条件下，当n很大时，n个随机变量的**算术平均值**$\overline{X}_n$与各随机变量的相同的**数学期望**(或**统计均值**)**相差很小**。

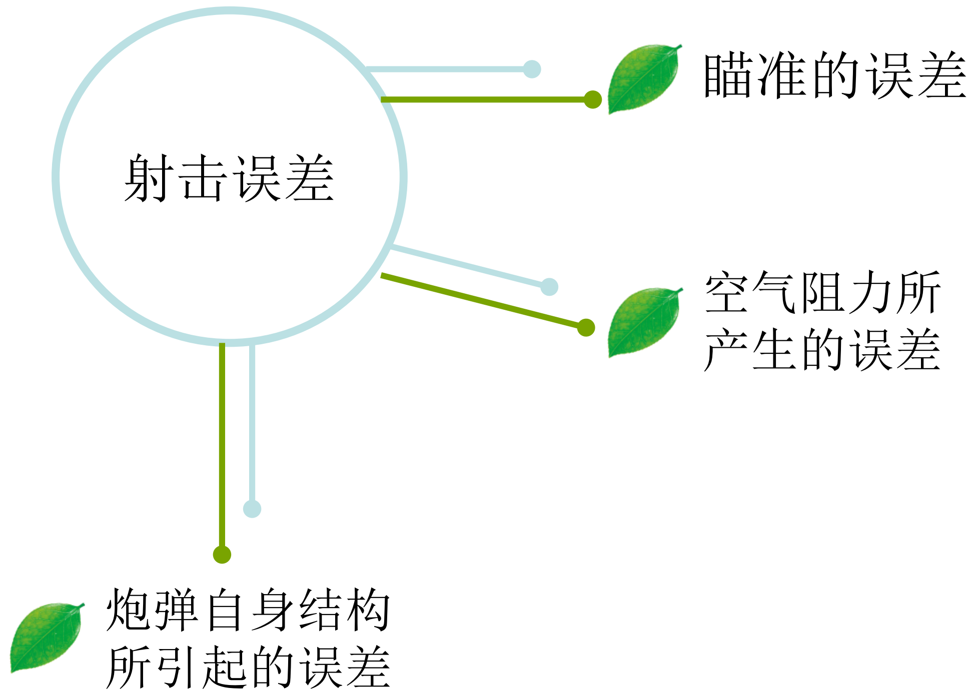
在测量中，对某物理量独立测试n次，设测量值为$X_1, X_2, \ldots, X_n$，视这n个值为互不相关的随机变量$X_1, X_2, \ldots, X_n$的观察值，***Chebyshev*大数定理**可以说明$X_1, X_2, \ldots, X_n$的**算术平均值作为该物理量的近似值是合理的。**

Story

§2 中心极限定理

一、中心极限定理的客观背景

在实际问题中，常常需要考虑许多随机因素所产生总影响。



Story

§2 中心极限定理

二、独立同分布下的中心极限定理

设 $X_1, X_2, \ldots$是独立同分布的随机变量序列，且 $E(X_k)=\mu<\infty$，$D(X_k)=\sigma^2>0$，$k=1, 2, \ldots$，则随机变量之和 $\sum_{k=1}^{n} X_k$ 的标准化变量的分布函数$F_n(x)$ 对于任意x 满足

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

$$=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt=\Phi(x)$$

二、独立同分布下的中心极限定理

设$\{X_n\}$为独立同分布的随机变量序列，其共同分布为区间(0,1)上的均匀分布。

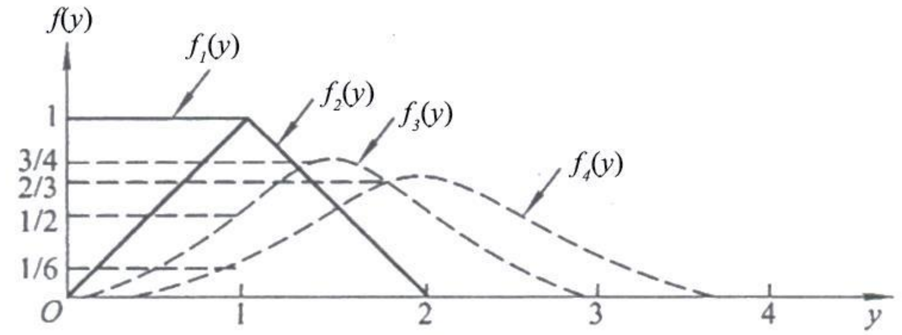
记$Y_n=\sum_{i=1}^{n}X_i$，$f_n(y)$为Y_n的密度函数，用卷积公式可以求出

$$f_1(y)=\begin{cases}1, & 0<y<1,\\ 0, & \text{其他}.\end{cases}$$

$$f_2(y)=\begin{cases}y, & 0<y<1,\\ 2-y, & 1\le y<2,\\ 0, & \text{其他}.\end{cases}$$

$$f_3(y)=\begin{cases}y^2/2, & 0<y<1,\\ -(y-3/2)^2+3/4, & 1\le y<2,\\ (3-y)^2/2, & 2\le y<3,\\ 0, & \text{其他}.\end{cases}$$

$$f_4(y)=\begin{cases}y^3/6, & 0<y<1,\\ \left[y^3-4(y-1)^3\right]/6, & 1\le y<2,\\ \left[(4-y)^3-4(3-y)^3\right]/6, & 2\le y<3,\\ (4-y)^3/6, & 3\le y<4,\\ 0, & \text{其他}.\end{cases}$$



§2 中心极限定理

三、棣莫佛—拉普拉斯定理

设随机变量η_n (n=1, 2, ...)服从参数为$n, p(0<p<1)$的二项分布，则

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

定理表明： 当n很大，$0<p<1$是一个定值时（或者说，$np(1-p)$也不太小时），二项变量η_n的分布近似正态分布$N(np, np(1-p))$

证明： 由$X_k \sim B(1,\ p)$　(k=1，2，...)，则$E(X_k)=p$，　$D(X_k)=p(1-p)$ (k=1, 2, ...)。由独立同分布下的中心极限定理得

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

四、李雅普诺夫定理

设随机变量$X_1, X_2, \dots, X_n, \dots$相互独立，它们具有数学期望和方差：$E(X_k)=\mu_k$, $D(X_k)=\sigma_k^2>0$, $k=1, 2, \dots$,

记 $B_n^2=\sum\limits_{k=1}^{n}\sigma_k^2$ 若存在正数δ，使得当$n\to\infty$时，

$$\frac{1}{B_n^{2+\delta}}\sum_{k=1}^{n}E\left\{\left|X_k-\mu_k\right|^{2+\delta}\right\}\to 0$$

则随机变量之和的标准化变量

$$Z_n=\frac{\sum\limits_{k=1}^{n}X_k-E\left(\sum\limits_{k=1}^{n}X_k\right)}{\sqrt{D\left(\sum\limits_{k=1}^{n}X_k\right)}}=\frac{\sum\limits_{k=1}^{n}X_k-\sum\limits_{k=1}^{n}\mu_k}{B_n}$$

的分布函数$F_n(x)$对于任意x满足

$$\lim_{n\to\infty}F_n(x)=\lim_{n\to\infty}P\left\{\frac{\sum\limits_{k=1}^{n}X_k-\sum\limits_{k=1}^{n}\mu_k}{B_n}\le x\right\}=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t=\Phi(x).$$

四、李雅普诺夫定理

定理表明：

无论各个随机变量$X_1, X_2, \ldots, X_n, \ldots$服从什么分布，只要满足定理的条件，那么它们的和 $\sum_{k=1}^{n} X_k$ 当n很大时，近似地服从正态分布

在实际中，考虑到一个随机变量可以表示成很多独立随机变量之和，它往往可以近似服从正态分布。

Lyapunov定理小结

$$Z_n = \frac{\sum_{k=1}^{n} X_k - \sum_{k=1}^{n} \mu_k}{B_n} \sim N(0,1)$$

$$\sum_{k=1}^{n} X_k = B_n Z_n + \sum_{k=1}^{n} \mu_k \sim N\left(\sum_{k=1}^{n} \mu_k, B_n^2\right)$$

§2 中心极限定理

例　某英语考试中共四选一选择题100题（每题1分），

1)若随机选

　　a)选对20题或以上的概率；

　　b)选对40题以上的概率;

2)某同学在多次模拟考试中平均成绩为85分，问这次考试中选对90题以上的概率。

例　　吸烟率为p，需要调查多少人才能保证吸烟频率与p的差不超过0.05的概率不低于95%？。

例 一个加法器同时收到20个噪声电压V_k (k=1,2,...,20)，设这20个噪声电压是独立产生的，均服从区间(0, 10)上的均匀分布。若 $V=\sum_{k=1}^{n}V_k$ 求$P(V>105)$。

解 $V_k \sim U(0,10)$，则$E(V_k)=\frac{0+10}{2}=5$，$D(V_k)=\frac{(10-0)^2}{12}=\frac{25}{3}$

由*Lyapunov*定理得：$$Y_{20}=\frac{\sum_{k=1}^{n}V_k-20\times 5}{\sqrt{20\times\frac{25}{3}}}=\frac{V-100}{\sqrt{\frac{500}{3}}}\sim N(0,1)$$

$$P(V>105)=1-P(V\le 105)=1-P\{\frac{V\text{-}100}{\sqrt{\frac{500}{3}}}\le\frac{105\text{-}100}{\sqrt{\frac{500}{3}}}\}=1-P\{\frac{V\text{-}100}{\sqrt{\frac{500}{3}}}\le 0.387\}$$

$$\approx 1-\Phi(0.387)=0.348.$$

四、李雅普诺夫定理

例 某车间有150台机床独立工作，每台机床工作时耗电量均为5kw，每台机床平均只有60%的时间运转。该车间应供应电多少千瓦，才能以99.9%的概率保证车间机床能够正常运转。

解 150台机床运转与否可以认为是150次独立重复试验。运转机床的台数m服从n=150，p=0.6的二项分布

设供电量5k (kw)，即认为运转的机床台数不超过k台时，机床正常运转，只需求出满足：$P(0 \le m \le k)=0.999$的k，就能保证车间的机床以99.9%的概率正常工作。

§2 中心极限定理

四、李雅普诺夫定理

由*Lyapunov*定理得：$P\left(\frac{0-np}{\sqrt{np(1-p)}} \le \frac{m-np}{\sqrt{np(1-p)}} \le \frac{k-np}{\sqrt{np(1-p)}}\right) = 0.999$

即：$P\left(-15 \le \frac{m-np}{\sqrt{np(1-p)}} \le \frac{k-90}{6}\right) = \Phi\left(\frac{k-90}{6}\right) - \Phi(-15) = 0.999$

$\Phi(-15) \approx 0$，则$\Phi\left(\frac{k-90}{6}\right) \approx 0.999 \Rightarrow \frac{k-90}{6} = 3.01 \Rightarrow k \approx 108.6$

取k=109，因此车间供电量为5×109=545 (kw)时，可以99.9%的概率正常工作

Story

谢　　谢！