§4. 4 大数定理与中心极限定理

大数定律，对第一章中提出的“频率稳定性”，给出理论上的论证

一般的大数定理讨论n个随机变量的平均值的稳定性.

一、切比雪夫(Chebyshev)不等式

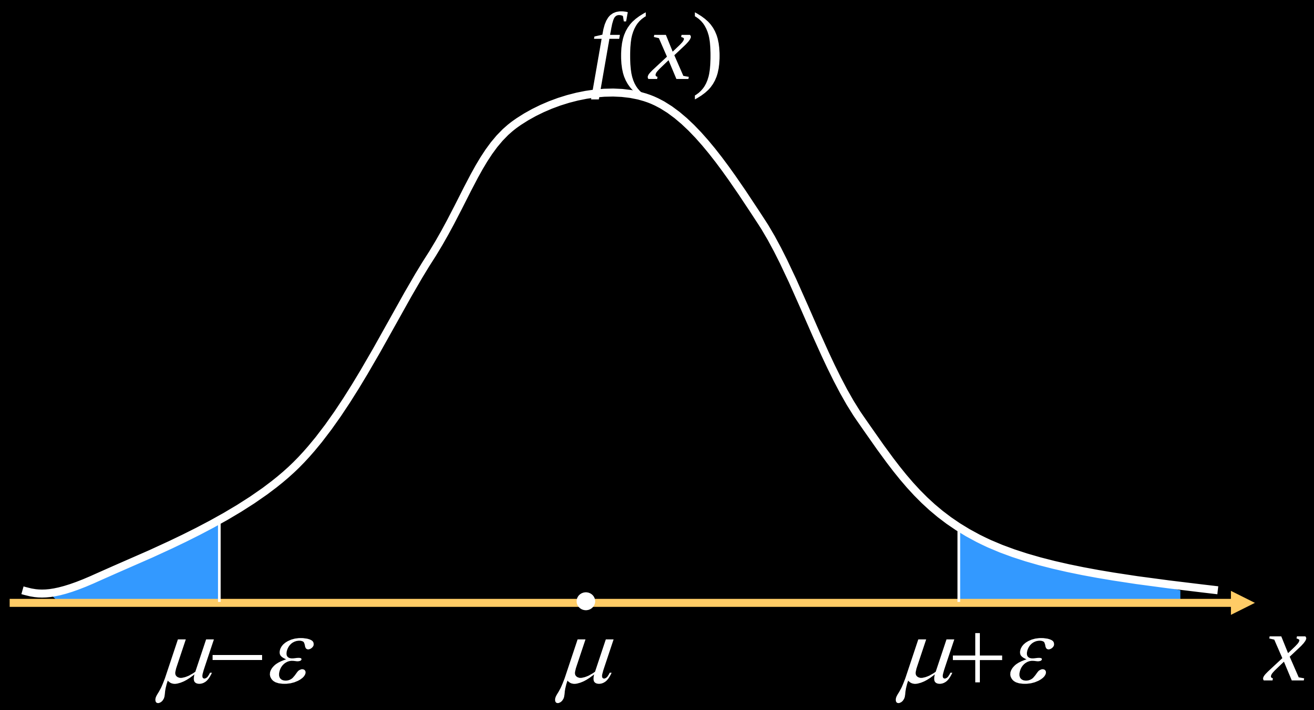
定理1 设随机变量X期望$E(X)=\mu$, 方差$D(X)=\sigma^2$, 则对于任给$\varepsilon>0$, 有

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

或

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

上述不等式称**切比雪夫不等式**.



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

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

证明　这里只证明X为连续型随机变量的情形，设X的概率密度为$f(x)$, 则有

$$P\{|X-\mu|\geq\varepsilon\}=\int_{|x-\mu|\geq\varepsilon} f(x)\,\mathrm{d}x$$

$$\leq\int_{|x-\mu|\geq\varepsilon}\frac{|x-\mu|^2}{\varepsilon^2}f(x)\,\mathrm{d}x$$

$$\leq\frac{1}{\varepsilon^2}\int_{-\infty}^{+\infty}(x-\mu)^2 f(x)\,\mathrm{d}x=\frac{\sigma^2}{\varepsilon^2}$$

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

注：①由切比雪夫不等式可以看出，若$D(X)$越小，则事件

$$\{|X-E(X)|<\varepsilon\}$$

的概率越大，即随机变量X集中在期望附近的可能性越大. 由此可见方差刻划了随机变量取值的离散程度.

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

②当方差已知时，即$D(X)=\sigma^2$, 切比雪夫不等式给出了X与它的期望的偏差不小于ε的概率的估计式．如取$\varepsilon=3\sigma$, 则有

$$P\{|X-E(X)|\geq 3\sigma\}\leq \frac{\sigma^2}{9\sigma^2}\approx 0.111.$$

故对任给的分布，只要期望和方差σ^2存在，则随机变量X取值偏离$E(X)$超过3倍均方差的概率小于0. 111.

例1 已知正常男性成人血液中，每一毫升白细胞数平均是7300，方差是700^2. 利用切比雪夫不等式估计每毫升白细胞数在5200~9400之间的概率.

解 设每毫升白细胞数为X, 依题意,

$$E(X)=7300,\ D(X)=700^2,$$

$$P\{5200\le X\le 9400\}$$

$$=P\{5200-7300\le X-7300\le 9400-7300\}$$

$$=P\{-2100\le X-E(X)\le 2100\}$$

$$=P\{|X-E(X)|\le 2100\}.$$

$E(X)=7300$, $D(X)=700^2$,

$P\{5200\le X\le 9400\}=P\{|X-E(X)|\le 2100\}$.

由切比雪夫不等式

$$P\{|X-E(X)|\le 2100\}\ge 1-\frac{D(X)}{(2100)^2}$$

$$=1-\left(\frac{700}{2100}\right)^2=1-\frac{1}{9}=\frac{8}{9}$$

即估计每毫升白细胞数在5200~9400之间的概率不小于8/9.

依概率收敛

与微积分学中的收敛性的概念类似，在概率论中，我们要考虑随机变量序列的收敛性.

定义1 设$X_1,X_2,\ldots,X_n,\ldots$是一个随机变量序列，$a$为一个常数，若对于任意给定的正数$\varepsilon$，有

$$\lim_{n\to\infty} P\{|X_n - a| < \varepsilon\} = 1,$$

则称序列$X_1,X_2,\ldots,X_n,\ldots$依概率收敛于$a$，记为

$$X_n \xrightarrow{P} a \quad (n \to \infty).$$

二 大数定理

若对于任意$n>1$, $X_1,X_2,\ldots,X_n$相互独立, 则称$X_1,X_2,\ldots,X_n,\ldots$相互独立.

定理2 设随机变量$X_1,X_2,\ldots,X_n,\ldots$相互独立, 且具有相同的期望和方差$E(X_i)=\mu$, $D(X_i)=\sigma^2$, $i=1,2,\ldots$ 记

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

则对任意$\varepsilon>0$, 有

$$\lim_{n\to\infty}P\left\{\left|Y_n-\mu\right|<\varepsilon\right\}=1$$ 即$Y_n \xrightarrow{P} \mu \quad (n\to\infty)$.

证明 $$E(Y_n)=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\mu,$$

$$D(Y_n)=\frac{1}{n^2}\sum_{i=1}^{n}D(X_i)=\frac{\sigma^2}{n},$$

由切比雪夫不等式，得

$$P\{|Y_n-\mu|<\varepsilon\}\ge 1-\frac{\sigma^2}{n\varepsilon^2}\xrightarrow[n\to\infty]{}1$$

即

$$Y_n\xrightarrow{P}\mu\quad(n\to\infty).$$

注: 定理表明：当 n 很大时，随机变量序列 $\{X_n\}$ 的算术平均值 $\frac{1}{n}\sum_{i=1}^{n} X_i$ 依概率收敛于其数学期望 μ.

推论（**伯努利大数定律**）设n_A是n重伯努利试验中事件A发生的次数，p是事件A在每次试验中发生的概率, 则对任意的$\varepsilon>0$, 有

$$\lim_{n\to\infty} P\left\{\left|\frac{n_A}{n}-p\right|<\varepsilon\right\}=1 \qquad (4.5)$$

或

$$\lim_{n\to\infty} P\left\{\left|\frac{n_A}{n}-p\right|\geq\varepsilon\right\}=0 \qquad (4.6)$$

证明 因为$n_A \sim b(n,p)$, 所以

$$n_A = X_1 + X_2 + \ldots + X_n,$$

其中$X_1, X_2, \ldots, X_n$相互独立, 且都服从以p为参数的0-1分布. 因而

$$E(X_k) = p,\ D(X_k) = p(1-p) \quad (k = 1, 2, \ldots, n),$$

由定理2即得

$$\lim_{n\to\infty} P\left\{ \left| \frac{1}{n}(X_1 + X_2 + \cdots + X_n) - p \right| < \varepsilon \right\} = 1,$$

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

注: ①伯努利大数定律是定理 2 的推论，它表明：当重复试验次数 n 充分大时，事件 A 发生的频率 $\frac{n_A}{n}$ 依概率收敛于事件 A 发生的概率 p. 定理以严格的数学形式表达了频率的稳定性. 在实际应用中，当试验次数很大时，便可以用事件发生的频率来近似代替事件的概率.

②如果事件A的概率很小, 则由伯努利大数定律知事件A发生的频率也是很小的, 或者说事件A很少发生. 即"概率很小的随机事件在个别试验中几乎不会发生", 这一原理称为小概率原理, 它的实际应用很广泛. 但应注意到, 小概率事件与不可能事件是有区别的. 在多次试验中, 小概率事件也可能发生.

四 中心极限定理

当一个随机变量X, 是由n个相互独立(可以不同分布不同数学期望不同方差, 但是方差要存在)的随机变量$X_1,X_2,\ldots,X_n$相加构成, 即

$$X=X_1+X_2+\ldots+X_n,$$

只要这n个随机变量的方差是差不多大的, n是足够大的, 则X近似服从正态分布, X的数学期望和方差, 就是各个$X_1,X_2,\ldots,X_n$的数学期望之和与方差之和. (一般n要大于20以上)

一个常用到中心极限定理的情况是二项分布, 即$X\sim b(n,p)$, 其中n较大, 通常要大于20, 则X是n个0-1分布的相互独立的随机变量

$$X_1, X_2, \ldots, X_n$$

之和. 而因为$E(X)=np$, $D(X)=np(1-p)$, 因此近似有

$$X\sim N(np,\ np(1-p)).$$

这被称为棣莫佛–拉普拉斯定理.

例2 一盒同型号螺丝钉共有100个, 已知该型号的螺丝钉的重量是一个随机变量, 期望值是100g, 标准差是10g, 求一盒螺丝钉的重量超过10.2kg的概率.

解 设X_i为第i个螺丝钉的重量, i=1,2,…,100, 且它们之间独立同分布, 于是一盒螺丝钉的重量$X=X_1+X_2+\dots+X_n$近似服从正态分布, 因$E(X_i)=100$, $D(X_i)=100$, i=1,2,…,100,则

$E(X)=10000$, $D(X)=10000$, 近似有

$X\sim N(10000, 100^2)$

求一盒螺丝钉的重量超过10.2kg的概率.

$X \sim N(10000, 100^2)$, 因此

$$P\{X > 10200\} = P\left\{\frac{X - 10000}{100} > \frac{10200 - 10000}{100}\right\}$$

$$= 1 - \Phi\left(\frac{200}{100}\right) = 1 - \Phi(2) = 1 - 0.97725$$

$$= 0.02275.$$

例 3 计算机在进行数学加法计算时，遵从四舍五入原则．为简单计，现在对小数点后面第一位进行舍入运算，则可以认为误差 X 服从[–0.5, 0.5]上的均匀分布．若在一项计算中进行了 100 次数字加法计算，求平均误差落在区间$\left[-\frac{\sqrt{3}}{20},\frac{\sqrt{3}}{20}\right]$上的概率．

解 n=100，用 X_i 表示第 i 次加法运算中产生的误差，X_i~U(–0.5,0.5), (i=1,2,…,100)，总误差 $X_1+X_2+\ldots+X_{100}$ 近似服从正态分布．

$X_i \sim U(-0.5,0.5)$, $E(X_i)=0$, $D(X_i)=\dfrac{1}{12}$, 令

$Y=X_1+X_2+\ldots+X_{100}$, 则 $E(Y)=0$, $D(Y)=\dfrac{100}{12}$,

近似有$Y \sim N(0,\dfrac{100}{12})$, $Y/100$ 为平均误差, 则

$E(Y)=0$, $D(Y)=\dfrac{100}{12}$, 近似有$Y\sim N(0,\dfrac{100}{12})$,

$Y/100$ 为平均误差, 则

$$P\left\{-\frac{\sqrt{3}}{20}\le\frac{Y}{100}\le\frac{\sqrt{3}}{20}\right\}=P\left\{|Y|\le 5\sqrt{3}\right\}$$

$$=P\left\{\left|\frac{Y}{\sqrt{100/12}}\right|\le\frac{5\sqrt{3}}{\sqrt{100/12}}\right\}=P\left\{\left|\frac{Y}{\sqrt{100/12}}\right|\le 3\right\}$$

$$=2\Phi(3)-1=0.9973$$

例4 某公司有200名员工参加一种资格证书考试. 按往年经验, 该考试通过率为0.8. 试计算这200名员工至少有150人考试通过的概率.

解 令

$$X_i=\begin{cases}1, & \text{第}i\text{人通过考试}\\ 0, & \text{第}i\text{人未通过考试}\end{cases},\ i=1,2,\ldots,200$$

则X_i~$b(1,0.8)$, 考试通过的人数

$Y=X_1+\ldots+X_{200}$~$b(200, 0.8)$, 但近似有

Y~$N(160, 32)$

考试通过的人数$Y \sim N(160, 32)$, 因此

$$P\{Y \geq 150\} = P\left\{\frac{Y-160}{\sqrt{32}} \geq \frac{150-160}{\sqrt{32}}\right\}$$

$$= P\left\{\frac{Y-160}{\sqrt{32}} \geq -1.77\right\} = \Phi(1.77) = 0.96$$

例5 某市保险公司开办一年人身保险业务, 被保险人每年需交付保险费160元, 若一年内发生重大人身事故, 其本人或家属可获2万元赔金. 已知该市人员一年内发生重大人身事故的概率为0.005, 现有5000人参加此项保险, 问保险公司一年内从此项业务中所得到的总收益在20万元到40万元之间的概率是多少?

解 记

$$X_i=\begin{cases}1, & \text{若第}i\text{个被保险人发生重大事故}\\ 0, & \text{若第}i\text{个被保险人未发生重大事故}\end{cases},\ i=1,2,\ldots,5000$$

则$X_i \sim b(1,0.005)$, $Y=X_1+\ldots+X_{5000}$是这5000人一年中发生重大人身事故的人数, 因此 $Y\sim b(5000,0.005)$, 近似有$Y\sim N(25, 24.875)$, 保险公司一年的总收益是$0.016\times5000-2Y$, 因此要求的概率是

$$P\{20\le 80-2Y\le 40\}$$

$Y \sim N(25, 24.875)$,

$$P\{20 \le 80 - 2Y \le 40) = P\{-60 \le -2Y \le -40\}$$

$$= P\{20 \le Y \le 30\}$$

$$= P\left\{\frac{20-25}{\sqrt{24.875}} \le \frac{Y-25}{\sqrt{24.875}} \le \frac{30-25}{\sqrt{24.875}}\right\}$$

$$\approx 2\Phi(1) - 1 = 0.6826$$

考研题(03304) 设总体 X 服从参数为 2 的指数分布, $X_1,X_2,\ldots,X_n$ 为来自总体 X 的简单随机样本, 则当 $n\to\infty$时, $Y_n=\dfrac{1}{n}\sum_{i=1}^{n}X_i^2$ 依概率收敛于________.

解 因

$$E(X_i^2)=D(X_i)+[E(X_i)]^2=\frac{1}{2^2}+\left(\frac{1}{2}\right)^2=\frac{1}{2}$$

根据大数定律应填 1/2.

考研题(01408) 一生产线生产的产品成箱包装, 每箱的重量是随机的, 假设每箱平均重50kg, 标准差为5kg. 若用最大载重量为5t的汽车承运, 试利用中心极限定理说明每辆车最多可以装多少箱, 才能保障不超载的概率大于0.977 ($\Phi(2)=0.977$, $\Phi(x)$是标准正态分布函数).

解 设$X_i(i=1,2,\ldots,n)$是装运的第i箱的重量, n为装运的箱数, 则$E(X_i)=50$, $D(X_i)=5^2$, 装运的总重量为$Y=X_1+X_2+\ldots+X_n$, $X_1,X_2,\ldots,X_n$独立同分布. $E(Y)=50n$, $D(Y)=25n$, 因此近似有$Y\sim N(50n, 25n)$, 不超载的概率为

$$P\{Y \le 5000\} = P\left\{\frac{Y-50n}{5\sqrt{n}} \le \frac{5000-50n}{5\sqrt{n}}\right\}$$

$$= \Phi\left(\frac{1000-10n}{\sqrt{n}}\right) > 0.977 = \Phi(2)$$

$$P\{Y \le 5000\} = P\left\{\frac{Y-50n}{5\sqrt{n}} \le \frac{5000-50n}{5\sqrt{n}}\right\}$$

$$= \Phi\left(\frac{1000-10n}{\sqrt{n}}\right) > 0.977 = \Phi(2)$$

故 $$\frac{1000-10n}{\sqrt{n}} > 2$$

解得n<98.0199, 也就是最多可以装98箱.

作业　习题4-4 第117页开始

第 5, 7, 9, 12 题