Chpt.5 Law of Large Number & Central Limit

第五章 大数定理及中心极限定理

上节回顾

- 依概率收敛
- 大数定理

一个随机变量的序列 $Y_1, Y_2, \cdots, Y_n, \cdots$，如果对任意 $\varepsilon > 0$，有 $\lim\limits_{n\to\infty} P\{|Y_n - p| \le \varepsilon\} = 1$，则称序列 Y_n **依概率收敛**到 p，记为 $Y_n \xrightarrow{P} p(n \to \infty)$

1 频率稳定性： 事件A发生的频率依概率收敛到事件A的概率

事件A发生的概率为 p 进行n次独立试验，A出现 n_A	$\Rightarrow \dfrac{n_A}{n} \xrightarrow{p} p(n \to \infty)$

2 算术均值稳定性：

随机变量 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立 具有相同的期望 μ 和方差 σ^2 或者同分布且期望 μ 存在	$\Rightarrow \dfrac{1}{n}(X_1 + X_2 + \cdots + X_n) \xrightarrow{p} \mu(n \to \infty)$

思考题

设随机变量$X_1, X_2, \cdots, X_n, \cdots$独立同分布，$X_1 \sim U(0,1)$，则$\sqrt[n]{X_1 X_2 \ldots X_n}$依概率收敛吗？如果依概率收敛，收敛于什么？

解： 记 $Y_n = \sqrt[n]{X_1 \ldots X_n}$，令 $Z_n = \ln Y_n = \frac{1}{n}(\ln X_1 + \ldots + \ln X_n)$.

则 $\ln X_1$，$\cdots$，$\ln X_n$，$\cdots$ 相互独立同分布，又

$$E(\ln X_1) = \int_0^1 \ln x dx = -1,$$

那么由辛钦大数定律知，故 $Z_n \xrightarrow{P} -1$，当 $n \to +\infty$.

利用依概率收敛的性质，得

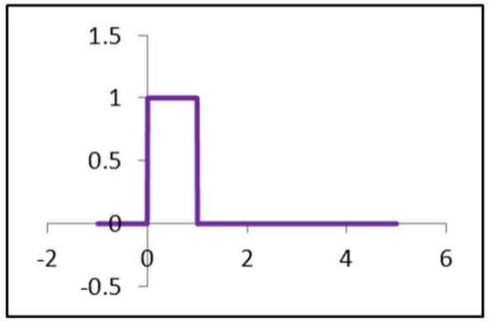
$Y_n = e^{Z_n} \xrightarrow{P} e^{-1}$，当 $n \to +\infty$.

中心极限定理

有许多随机变量，它们是由大量的相互独立的随机变量的综合影响所形成的，而其中每个个别的因素作用都很小，这种随机变量往往服从或近似服从正态分布，或者说它的极限分布是正态分布，中心极限定理正是从数学上论证了这一现象，它在长达两个世纪的时期内曾是概率论研究的中心课题.

现象观察

X_1



X1+X2=?

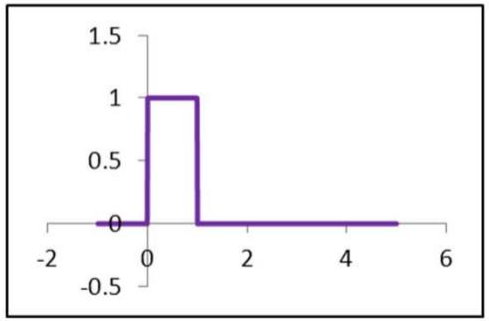
X1服从U(0,1)

X1,X2都服从U(0,1)

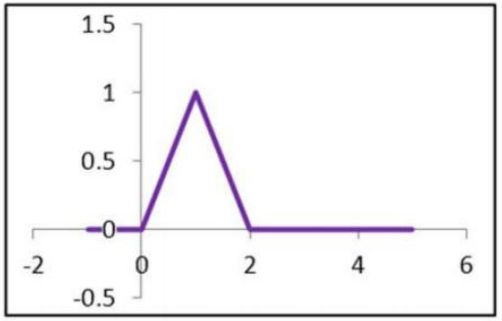
现象观察

Xi服从U(0,1)

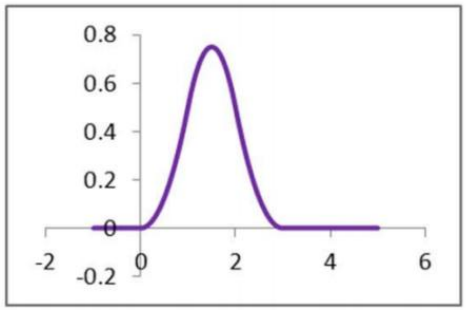
X_1



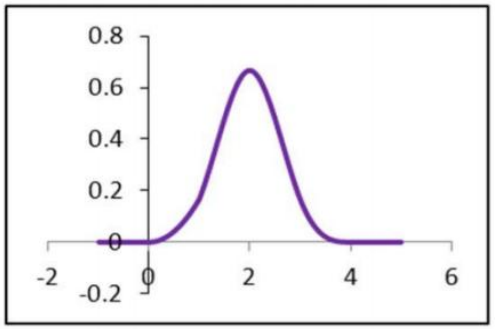
$X_1 + X_2$



$X_1 + X_2 + X_3$



$X_1 + X_2 + X_3 + X_4$



[独立同分布的中心极限定理CLT]

如果随机变量 $X_1,\cdots,X_n,\cdots$ 独立同分布，数学期望与方差存

在，$E(X_i)=\mu$ ，$D(X_i)=\sigma^2(i=1,2,\cdots,n,\cdots)$，记 $Y_n=\sum_{i=1}^{n}X_i$

其标准化变量 $Z_n=\dfrac{Y_n-E(Y_n)}{\sqrt{D(Y_n)}}=\dfrac{Y_n-n\mu}{\sqrt{n\sigma^2}}$，其分布函数记为 $F_n(x)=P\{Z_n\le x\}$

那么分布函数列的极限为

$$F(x)=\lim_{n\to\infty}F_n(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$$

也就是说，分布的极限为标准正态分布。

因此，当n充分大时近似有

$$\frac{Y_n-n\mu}{\sqrt{n\sigma^2}}\sim N(0,1)\qquad Y_n\sim N(n\mu,n\sigma^2)$$

或者 $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$ $\qquad \bar{X}\sim N(\mu,\dfrac{\sigma^2}{n})$

- 也就是说均值为 μ，方差为 σ^2 的独立同分布的随机变量 $X_1, X_2, \cdots, X_n$ 的和 $\sum_{i=1}^{n} X_i$，当n充分大时，近似地服从均值为 $n\mu$ 方差为 $n\sigma^2$ 的正态分布。

[定理2（Lyapunov定理）] 如随机变量 $X_1,\cdots,X_n,\cdots$ 相互独立，数学期望 $E(X_i)=\mu_i$ 与方差 $D(X_i)=\sigma_i^2\neq 0 \quad (i=1,2,\cdots,n,\cdots)$，记 $Y_n=\sum_{k=1}^{n}X_K$

$$B_n^2=D(Y_n)=\sum_{k=1}^{n}\sigma_K^2 \qquad Z_n=\frac{Y_n-E(Y_n)}{B_n}=\frac{\sum_{k=1}^{n}X_K-E(\sum_{k=1}^{n}X_K)}{\sqrt{D(\sum_{k=1}^{n}X_K)}}$$

分布函数 $F_n(x)=P\{Z_n\le x\}$，如果存在正数 $\delta>0$ 使 $n\to\infty$ 时

$$\frac{1}{B_n^{2+\delta}}\sum_{k=1}^{n}E\{(X_K-\mu_K)^{2+\delta}\}\to 0$$ 那么分布函数列的极限为

$$F(x)=\lim_{n\to\infty}F_n(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$$

上述定理说明，随机变量 $Z_n = \dfrac{Y_n - E(Y_n)}{B_n} = \dfrac{\sum_{k=1}^{n} X_k - E(\sum_{k=1}^{n} X_k)}{\sqrt{D(\sum_{k=1}^{n} X_k)}}$ ，

当n充分大时近似地服从标准正态分布$N(0,1)$。

由此，当n很大时，$Y_n = \sum_{k=1}^{n} X_K$ 近似地服从 $N(\sum_{k=1}^{n} \mu_k, B_n^2)$

定理说明：独立的随机变量 $X_1, \cdots, X_n, \cdots$，不管 X_k 本身是什么样的分布，只要满足一定的条件，当n充分大时，n 个随机变量的和 $\sum_{k=1}^{n} X_k$ 近似地服从正态分布。

[1] 在客观的实际应用中，所考虑的对象往往是由大量的相互独立的随机因素的综合影响形成（往往是和的形式），尽管这诸多的因素之分布是未知的，但是它们的和往往近似服从正态分布;

[2] 在实际应用中，我们进行分析往往是观察值的和或平均，而由上述定理保证当样本个数足够大时，这个和趋于正态分布，这在后面的统计推断中是极其重要的。

正是上述定理所陈述的是分布的极限，以及他们在应用统计中的重要性（或中心地位），Polya在1920年给他取名为“中心极限定理”

Example 设一次贝努里试验中成功的概率为p $(0 < p < 1)$, 令Sn表示n重贝努里试验中成功的次数，那么$Sn \sim B(n,p)$。

在实际问题中, 人们常常对成功次数介于整数α、β之间$(\alpha<\beta)$的概率感兴趣，即要计算

$$P\{\alpha < S_n < \beta\} = \sum_{\alpha<k<\beta} B(k,n,p)$$

这一和式往往涉及很多项，直接计算相当困难. 然而我们注意到

$S_n = X_1 + X_2 + \cdots + X_n$，$X_i$ 表示第i次试验的结果，X_i 独立同分布，

$E(X_i)=p$，$D(X_i)=p(1-p)$，则 $E(S_n)=np$，$D(S_n)=np(1-p)$，我们知道

$$\frac{S_n - E(S_n)}{\sqrt{D(S_n)}} = \frac{S_n - np}{\sqrt{np(1-p)}}$$ 近似服从$N(0,1)$

【德莫佛—拉普拉斯定理】

因此这个定理表示二项分布的标准化变量依分布收敛于标准正态分布. 简单地说成二项分布渐近正态分布.

定理的直接应用是:

n重贝努里实验中，成功次数介于整数α、β之间的概率

当n很大，p的大小适中时，可用正态分布近似计算：

$$P\{\alpha < S_n < \beta\} = P\left\{\frac{\alpha - np}{\sqrt{np(1-p)}} \le \frac{S_n - np}{\sqrt{np(1-p)}} \le \frac{\beta - np}{\sqrt{np(1-p)}}\right\}$$

$$\approx \Phi\left(\frac{\beta - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{\alpha - np}{\sqrt{np(1-p)}}\right)$$

$$P\{\alpha < S_n < \beta\} \approx \Phi\left(\frac{\beta - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{\alpha - np}{\sqrt{np(1-p)}}\right)$$

它的含义可用下图显示（为了直观，图中显示的是未标准化的随机变量)：作相邻小矩形，各小矩形的底边中心为k(α≤k≤β)，底边长为1，高度为b(k; n, p)，这些小矩形面积之和即为P(α≤Sn≤β). 再作标准正态分布N(np, npq)的密度曲线，其在[α,β]之间曲线覆盖的面积为上式右边之值.

泊松定理 中心极限定理

[泊松(Poisson)定理] 如果存在正常数 λ，当 $n\to\infty$ 时有 $np_n\to\lambda$，则

$$\lim_{n\to\infty} b(k,n,p)=\frac{\lambda^k}{k!}e^{-\lambda} \qquad k=0,1,2,\cdots$$

Remark：泊松定理，它与定理5是没有矛盾的. 因为泊松定理要求 $\lim_{n\to\infty} np_n=\lambda$ 是常数，而定理5中p是固定的.

实际应用中，当n很大时

[1] 若p大小适中，可以用正态分布去逼近;

[2] 如果p接近0, 且np较小，则二项分布的图形偏斜度太大，用正态分布去逼近效果就不好，此时用泊松分布去估计精度会更高.

Example (pp.151-152) 设一货轮在某海区航行，已知每遭受一次波浪的冲击，纵摇角度大于3° 的概率为p=1/3 。若货轮在航行中遭受了90000 次波浪冲击，问其中有 29500 ~ 30500次纵摇角度大于3°的概率是多少？

[解]：可将货轮每遭受一次波浪冲击看作是一次试验，并认为实验是独立的。在 90000次波浪冲击中，纵摇角度大于3°的次数记为X，则 X 为一随机变量，它服从二项分布B(90000,1/3)。其分布列为

$$P(X=k)=C_{90000}^{k}\left(\frac{1}{3}\right)^{k}\left(\frac{2}{3}\right)^{90000-k},k=0,1,2,\cdots,90000$$

所求概率精确的算式为

$$P(29500 < X \le 30500) = \sum_{k=29501}^{30500} C_{90000}^{k} \left(\frac{1}{3}\right)^{k} \left(\frac{2}{3}\right)^{90000-k}.$$

显然，要直接计算是困难的。可以利用德莫佛—拉普拉斯定理来求它的近似值。即有

$$P(29500 < X \le 30500) = P\left(\frac{29500-np}{\sqrt{np(1-p)}} < \frac{X-np}{\sqrt{np(1-p)}} \le \frac{30500-np}{\sqrt{np(1-p)}}\right)$$

$$\approx \Phi\left(\frac{30500-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{29500-np}{\sqrt{np(1-p)}}\right)$$

$$\approx \Phi\left(\frac{5}{\sqrt{2}}\right) - \Phi\left(-\frac{5}{\sqrt{2}}\right) = 0.9995.$$

拓展：如果摇摆超过3°，则每次翻船的概率为0.001%，则最终翻船的概率是多少？

Example 近似计算时，原始数据y_k四舍五入到小数第m位，这时舍入误差X_k可以看作在$[-0.5\times10^{-m}, 0.5\times10^{-m}]$上均匀分布，而据此n个$X_k$的和$\Sigma X_k$，按四舍五入所得的误差是多少呢？

[解]：习惯上人们总是以各误差上限的和来估计ΣX_k，即$0.5\times n\times10^{-m}$. 当n很大时，这个数自然很大.

事实上，误差不太可能这么大. 因为$\{X_k\}$独立同分布，$E(X_k)=0, D(X_k)=\sigma^2=10^{-2m}/12$

$$P\left\{\frac{|\sum_{i=1}^{n} X_i|}{\sqrt{n}\sigma}\le x\right\}\approx 2\Phi(x)-1$$

若取x=3,上述概率为0.997. 此时和的误差超过 $3\sigma\sqrt{n}=3\times\sqrt{n}\times1/\sqrt{12}\times10^{-m}$ 的可能性仅为0.003. 显然，对较大的n，这一误差界限远小于习惯上的保守估计$0.5\times n\times10^{-m}$.

两数相比，大约有$\left(\sqrt{3}\times\sqrt{n}\times1/2\times10^{-m}\right)/0.5\times n\times10^{-m}=\sqrt{3}/\sqrt{n}$,

若100个数相加，则大约为1.7/10=0.17倍，10000个数相加则相差为0.017倍

Example 设 $X_1, X_2, \cdots, X_n, \cdots$ 独立同分布，$E(X_i)=\mu$，$D(X_i)=\sigma^2$令，

$$\overline{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i \quad , \quad S_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2$$

求证: $S_n^2 \xrightarrow{P} \sigma^2$　　统计中的重要观点！

$$S_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2=\frac{1}{n}\sum_{i=1}^{n}((X_i-\mu)-(\overline{X}_n-\mu))^2$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left[(X_i-\mu)^2-2(\overline{X}_n-\mu)(X_i-\mu)+(\overline{X}_n-\mu)^2\right]$$

$$=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-\frac{2}{n}\sum_{i=1}^{n}(X_i-\mu)(\overline{X}_n-\mu)+(\overline{X}_n-\mu)^2$$

$$=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-(\overline{X}_n-\mu)^2$$

由辛钦大数定律知 $\overline{X}_n \xrightarrow{p} \mu$，从而 $(\overline{X}_n - \mu)^2 \xrightarrow{p} 0$. 再因 $\{(X_i - \mu)^2\}$ 独立同分布，$E(X_i - \mu)^2 = D(X_i) = \sigma^2$，故 $\{Y_i = (X_i - \mu)^2\}$也服从辛钦大数定律，即 $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \xrightarrow{p} \sigma^2$，故此 $S_n^2 \xrightarrow{P} \sigma^2$。

中心极限定理重要的作用：

[1] 对数理统计学的许多分支，如参数（区间）估计、假设检验、抽样调查等

[2] 是保险精算等学科的理论基础之一.

假定某保险公司为某险种推出保险业务，现有n个顾客投保，第i份保单遭受风险后损失索赔量记为Xi. 对该保险公司而言，随机理赔量应该是所有保单索赔量之和，记为S，即 $S=\sum_{i=1}^{n} X_i$，弄清S的概率分布对保险公司进行保费定价至关重要.

在实际问题中，通常假定所有保单索赔相互独立. 这样，当保单总数n充分大时，我们并不需要计算S 的精确分布（一般情况下这是困难甚至不可能的）. 此时，可应用中心极限定理，对S进行正态逼近：$\frac{S-E(S)}{\sqrt{D(S)}}$ 渐近具有正态分布N(0,1)，并以此来估计一些保险参数.

Example 某保险公司发行一年期的保险索赔金分别为1万元与2万元的两种人身意外险. 索赔概率q_k及投保人数n_k如下表所示（金额单位：万元）

类别k	索赔概率q_k	索赔额b_k	投保数n_k
1	0.02	1	500
2	0.02	2	500
3	0.10	1	300
4	0.10	2	500

保险公司希望只有0.05的可能使索赔金额超过所收取的保费总额. 设该保险公司按期望值原理进行保费定价，即收取的总保费，应为索赔金额期望的函数。此时，有即保单i的保费 $\pi(X_i)=(1+\theta)E(X_i)$. 要求估计θ.

解：设每笔保单的索赔额为Xi,则索赔总额为 $S=\sum_{i=1}^{1800}X_i$ ，计算其均值与方差

$$ES = \sum_{i=1}^{1800} EX_i = \sum_{k=1}^{4} n_k b_k q_k$$

总索赔的期望

$$= 500 \cdot 1 \cdot 0.02 + 500 \cdot 2 \cdot 0.02 + 300 \cdot 1 \cdot 0.10 + 500 \cdot 2 \cdot 0.10 = 160,$$

$$VarS = \sum_{i=1}^{1800} Var\, X_i = \sum_{k=1}^{4} n_k b_k^{\,2} q_k (1 - q_k)$$

$$= 500 \cdot 1^2 \cdot 0.02 \cdot 0.98 + 500 \cdot 2^2 \cdot 0.02 \cdot 0.98$$

$$+ 300 \cdot 1^2 \cdot 0.10 \cdot 0.90 + 500 \cdot 2^2 \cdot 0.10 \cdot 0.90$$

$$= 256$$

总索赔的方差

由此得保费总额

$$\pi(S)=(1+\theta)ES=160(1+\theta).$$

依题意， 我们有

$$P(S\le(1+\theta)ES)=0.95$$

也即

$$P\left(\frac{S-ES}{\sqrt{VarS}}\le\frac{\theta ES}{\sqrt{VarS}}\right)=P\left(\frac{S-ES}{\sqrt{VarS}}\le 10\theta\right)=0.95.$$

将 $\frac{S-E(S)}{\sqrt{D(S)}}$近似看作标准正态随机变量，查表可得 $10\theta=1.645$

故 $\theta=0.1645$

应收总保费为

$$\pi(S) = (1+\theta)ES = 160(1+\theta).$$
$$= 160 * (1 + 0.1654) = 160 * 0.1645 = 186.32$$

拓展问题1：请问总共186.32万元的预期应收总保费，如何在四类保险中分别收取？

类别k	索赔概率q_k	索赔额b_k	投保数n_k
1	0.02	1	500
2	0.02	2	500
3	0.10	1	300
4	0.10	2	500

拓展问题2：如果保险公司希望盈利概率至少为99%，该如何收取？
若为99.9%又该如何收取？
其中盈利概率变化，盈利的期望如何变化？理解期不同

拓展问题3：如果购买人数又变成价格影响的随机变量，则如何定价？