

第5章 大数定律与中心极限定理

Chpt. 5 Law of Large Number & Central Limit Theory



本章学习目标包括两个：

- [1] 对概率论中的一些结论作出严格的证明；
- [2] 为后面的统计作出准备。

概率论早期发展的目的：揭示由于大量随机因素产生影响而呈现的规律性.

概率与频率之间的关系 → 大数定律：研究无穷随机试验序列，刻画事件的概率与它发生的频率之间的关系。

大量的相互独立的随机因素的综合影响 → 中心极限定理：将观察的误差看作大量独立微小误差的累加，其分布渐近正态。



5.1 大数定理 (Law of Large Number)

最基本的假设：频率 \rightarrow 概率。

在相同的条件下，进行 n 次独立试验，其中事件 A 发生的次数记为 n_A ，定义 $f_n(A) = \frac{n_A}{n}$ ，我们说 $f_n(A) \rightarrow p$ ， p 就是事件 A 发生的概率

p 是一个抽象得到的数，按一般的极限描述就是：

$$\forall \varepsilon > 0, \exists N, \text{当 } n > N \text{ 时, } \left| \frac{n_A}{n} - p \right| \leq \varepsilon$$



上式存在问题

n_A 不是一般的数而是随机变量，因此 $\left|\frac{n_A}{n} - p\right| \leq \varepsilon$ 表示的是一个事件，说
 $n > N$ 时总有 $\left|\frac{n_A}{n} - p\right| \leq \varepsilon$ 是不符合逻辑的，只能说以多大的概率成立上式。

$$P\left\{\left|\frac{n_A}{n} - p\right| \leq \varepsilon\right\} \rightarrow 1$$

$$\lim_{n \rightarrow \infty} P\left\{\left|\frac{n_A}{n} - p\right| \leq \varepsilon\right\} = 1$$

这个等式是否成立？



切比雪夫(Chebyshev)不等式 若随机变量的方差存在, 则

对任意给定的正数 ε , 恒有 $P\{|X - \mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2}$

证明:

$$\begin{aligned} P\{|X - \mu| \geq \varepsilon\} &= \int_{|x-\mu| \geq \varepsilon} f(x) dx \\ &\leq \int_{|x-\mu| \geq \varepsilon} \frac{(x-\mu)^2}{\varepsilon^2} f(x) dx \\ &\leq \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\varepsilon^2} f(x) dx = \frac{\sigma^2}{\varepsilon^2} \end{aligned}$$

上面的式子等价于

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



马尔可夫(Markov)不等式 设 X 为取非负值的随机变量, 则对任意给定的正数 ε , 恒有 $P(X \geq \varepsilon) \leq \frac{E(X)}{\varepsilon}$

证明: $\forall \varepsilon > 0$, 令 $I = \begin{cases} 1, & \text{若 } X \geq \varepsilon \\ 0, & \text{其他} \end{cases}$

由于 $X \geq 0, \varepsilon > 0$, 可得 $I \leq \frac{X}{\varepsilon}$

所以 $E(I) \leq E\left(\frac{X}{\varepsilon}\right) = \frac{1}{\varepsilon} E(X) \Rightarrow P(X \geq \varepsilon) \leq \frac{1}{\varepsilon} E(X)$

令 $Y = (X - E(X))^2, \varepsilon = \varepsilon^2$, 可得

$$P\left\{(X - E(X))^2 \geq \varepsilon^2\right\} \leq \frac{E\left((X - E(X))^2\right)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}$$

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2} \quad (\text{Chebyshev)不等式}$$



$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{n_A}{n} - p \right| \leq \varepsilon \right\} = 1 \quad \text{是否成立呢?}$$

注意到对任意的 n 重独立重复试验可定义,

$$X_i = \begin{cases} 1, & \text{第} i \text{次试验中} A \text{出现} \\ 0, & \text{第} i \text{次试验中} A \text{没出现} \end{cases}$$

$$n_A = X_1 + X_2 + \cdots + X_n$$

$$Y_n = \frac{n_A}{n} = \frac{1}{n} (X_1 + X_2 + \cdots + X_n)$$

我们知道: $E(X_i) = p, D(X_i) = p(1-p) = \sigma^2$

故: $E(Y_n) = p, D(Y_n) = \frac{1}{n} \sigma^2$



$$P\left\{\left|\frac{n_A}{n} - p\right| \leq \varepsilon\right\} = P\{|Y_n - p| \leq \varepsilon\}$$

$$\geq 1 - \frac{D(Y_n)}{\varepsilon^2} = 1 - \frac{1}{n} \left(\frac{\sigma^2}{\varepsilon^2} \right)$$

$$1 \geq P\left\{\left|\frac{n_A}{n} - p\right| \leq \varepsilon\right\} = P\{|Y_n - p| \leq \varepsilon\} \geq 1 - \frac{1}{n} \left(\frac{\sigma^2}{\varepsilon^2} \right)$$

$$\lim_{n \rightarrow \infty} P\left\{\left|\frac{n_A}{n} - p\right| \leq \varepsilon\right\} = 1$$



[Bernoulli大数定理] 设 n_A 是 n 次独立试验中事件A发生的次数, p 是事件A在一次试验中发生的概率, 则对任意的 $\varepsilon > 0$, 有

$$\lim_{n \rightarrow \infty} P\left\{\left|\frac{n_A}{n} - p\right| \leq \varepsilon\right\} = 1$$

Remark1: 一个随机变量的序列 $Y_1, Y_2, \dots, Y_n, \dots$ 如果对任意 $\varepsilon > 0$, 有 $\lim_{n \rightarrow \infty} P\{|Y_n - p| \leq \varepsilon\} = 1$, 则称序列依概率收敛到 p (稳定性一), 记为

$$Y_n \xrightarrow{P} p (n \rightarrow \infty)$$

Remark2 (大数定律含义之一): Bernoulli定理说明事件A发生的频率 n_A / n 依概率收敛到事件的概率 p 。以严格的数学形式表达了我们的直观看法。在实际应用中, 当试验次数足够大时, 便可以用事件的频率来代替事件的概率 p 。

上面的证明中，我们用到了

$$\lim_{n \rightarrow \infty} P\left\{\left|\frac{n_A}{n} - p\right| \leq \varepsilon\right\} = \lim_{n \rightarrow \infty} P\left\{\left|\frac{1}{n}(X_1 + \cdots + X_n) - p\right| \leq \varepsilon\right\} = 1$$

这对于一般的 X_1, \cdots, X_n, \cdots 是否成立？

[**辛钦 (Khintchine) 大数定理**] 设随机变量 X_1, \cdots, X_n, \cdots 相互独立且同分布，具有相同的数学期望 $E(X_i) = \mu$ ，则对任意的 $\varepsilon > 0$ 有

$$\lim_{n \rightarrow \infty} P\left\{\left|\frac{1}{n}(X_1 + \cdots + X_n) - \mu\right| \leq \varepsilon\right\} = 1$$



我们只在随机变量 $X_1, X_2, \dots, X_n, \dots$ 的方差存在且 $D(X_1) = \sigma^2$ 这一条件下证明上述结果

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

则 $E(\bar{X}_n) = \mu, D(\bar{X}_n) = \frac{1}{n} \sigma^2$

由切比雪夫不等式可得：

$$1 \geq P(|\bar{X}_n - \mu| \leq \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}$$

故当 $n \rightarrow \infty$ 时, $P(|\bar{X}_n - \mu| \leq \varepsilon) = 1$

即 $\lim_{n \rightarrow \infty} P\left\{\left|\frac{1}{n}(X_1 + \dots + X_n) - \mu\right| \leq \varepsilon\right\} = 1$



Remark（大数定律含义之二）：

当 n 很大时，随机变量 X_1, \dots, X_n, \dots 的算术平均 $\frac{1}{n}(X_1 + \dots + X_n)$ 接近于数学期望 $E(X_i) = \mu$ 。

通俗地说 n 个独立随机变量的算术平均当 n 很大时接近于一个常数（稳定性二）。进一步地，可以放宽对 X_i 方差的要求，但**要求 X_i 同分布**。



概括前面的几个定理，可以归结为两点：

1 频率稳定性：事件A发生的频率以概率收敛到概率p

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

2 算术均值稳定性：

随机变量 $X_1, X_2, \dots, X_n, \dots$ 相互独立	$\Rightarrow \frac{1}{n} (X_1 + X_2 + \dots + X_n) \xrightarrow{p} \mu (n \rightarrow \infty)$
具有相同的均值 μ 和方差 σ^2	



5.2中心极限定理

一般地, 我们考虑随机变量和的标准化形式

$$\frac{\sum_{k=1}^n X_K - \sum_{k=1}^n E(X_K)}{\sqrt{D(\sum_{k=1}^n X_K)}} \quad Y_n = \sum_{k=1}^n X_K$$

则上式可以写为

$$Z_n = \frac{Y_n - E(Y_n)}{\sqrt{D(Y_n)}} = \frac{\sum_{k=1}^n X_k - \sum_{k=1}^n E(X_k)}{\sqrt{D(\sum_{k=1}^n X_k)}}$$

可知 $E(Z_n) = 0$, $D(Z_n) = 1$

那么它的极限分布是什么呢?

[独立同分布的中心极限定理（林德伯格-莱维定理）]

如随机变量 X_1, \dots, X_n, \dots 相互独立服从同一分布，数学期望与方差 $E(X_i) = \mu, D(X_i) = \sigma^2 (i = 1, 2, \dots, n, \dots)$ ，记 $Y_n = \sum_{k=1}^n X_k$ ，其标准化变量 $Z_n = \frac{Y_n - E(Y_n)}{\sqrt{D(Y_n)}} = \frac{Y_n - n\mu}{\sqrt{n\sigma^2}}$ 的分布函数记为 $F_n(x) = P\{Z_n \leq x\}$ ，那么该分布函数列的极限为

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

定理说明，分布的极限为标准正态分布，则当n充分大时近似有

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

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



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, \dots, X_n 的算术平均当 n 充分大时近似地服从均值为 μ 方差为 σ^2/n 的正态分布。

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

$$B_n^2 = D(Y_n) = \sum_{k=1}^n \sigma_k^2 \quad Z_n = \frac{Y_n - E(Y_n)}{B_n} = \frac{\sum_{k=1}^n X_k - E(\sum_{k=1}^n X_k)}{\sqrt{D(\sum_{k=1}^n X_k)}}$$

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

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

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



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

当 n 充分大时近似地服从正态分布 $N(0,1)$ 。

由此当 n 很大时, $\sum_{k=1}^n X_k = B_n Z_n + \sum_{k=1}^n \mu_k$ 近似地服从 $N(\sum_{k=1}^n \mu_k, B_n^2)$

定理说明当 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, 而不管 X_k 本身是什么样的分布, 只要满足一定的条件即可。



[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，但是他们的和服从正态分布；**

[2] 在实际应用中，我们进行分析往往是对观察值的和或平均进行的，而这个和已经由上述定理保证是趋向于正态分布的，这就是说当样本个数足够大时，样本和就趋于正态分布，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 5.1 设一次伯努利试验中成功的概率为 p ($0 < p < 1$), 令 S_n 表示 n 重伯努利试验中成功的次数, 那么 $S_n \sim B(n, p)$ 。在实际问题中, 人们常常对成功次数介于两整数 α 、 β 和之间($\alpha < \beta$)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 n 重实验中第 i 次试验的结果), $E(X_i) = p$, $D(X_i) = p(1-p)$, 则 $E(S_n) = np$, $D(S_n) = np(1-p)$, 我们知道

$$\frac{S_n - E(S_n)}{\sqrt{D(S_n)}} = \frac{S_n - np}{\sqrt{np(1-p)}}$$

近似为 $N(0,1)$ 【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. 简单地说是二项分布渐近正态分布.

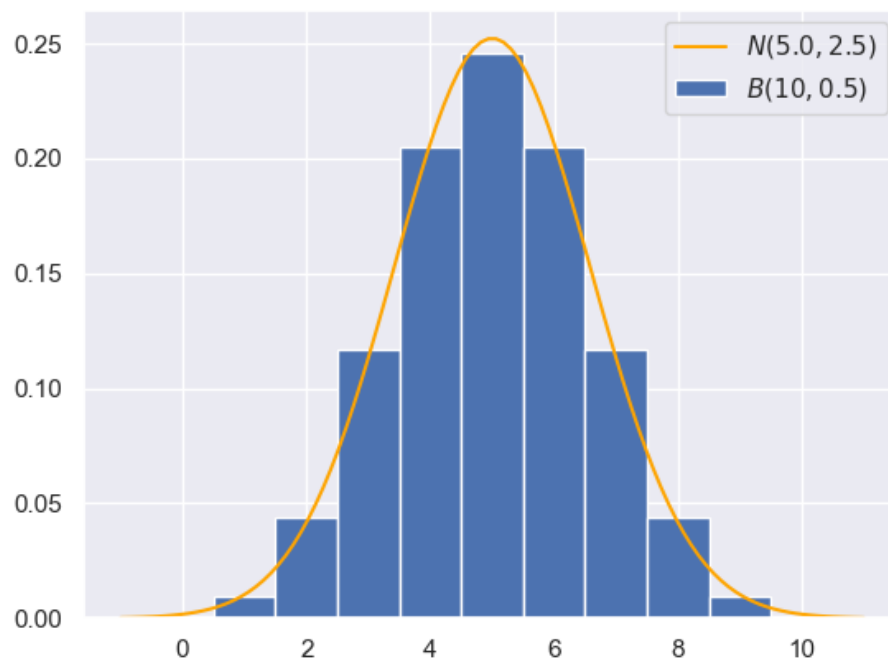
定理的直接应用是:

当n很大, p的大小适中时, 可用正态分布近似计算:

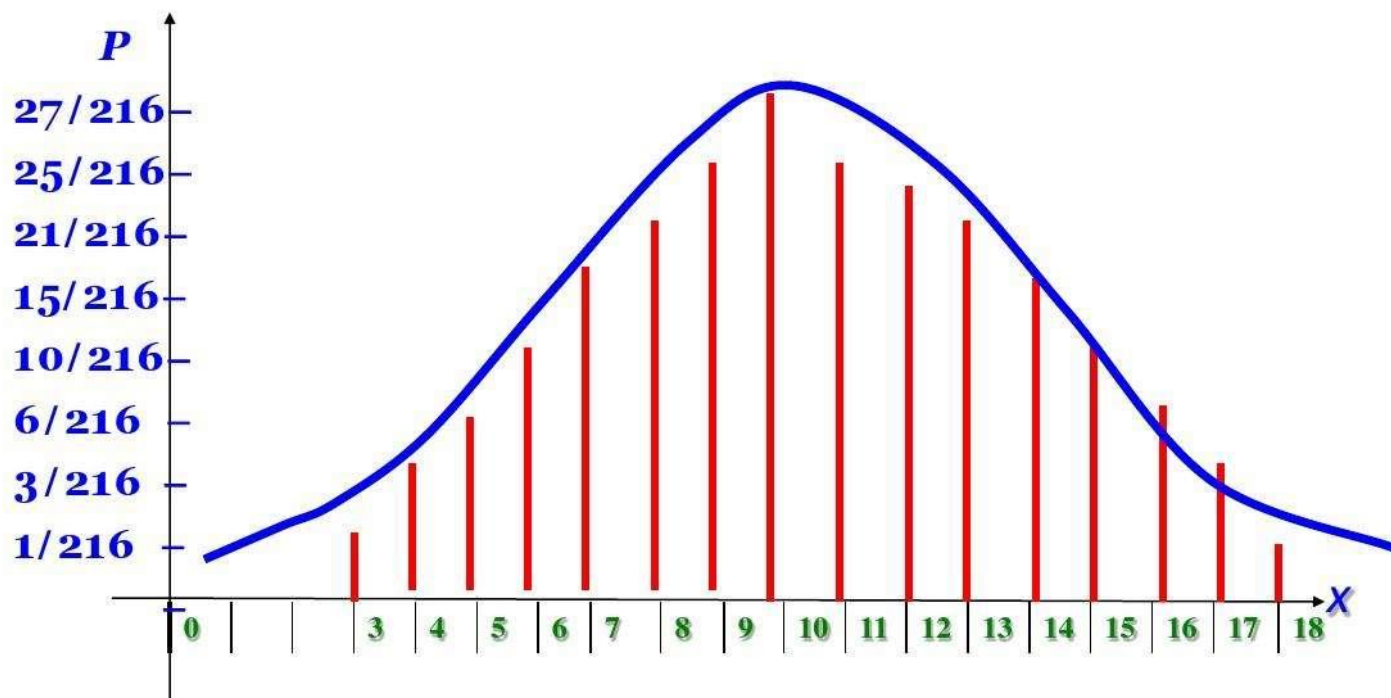
$$\begin{aligned} P\{\alpha < S_n < \beta\} &= \sum_{\alpha < k < \beta} B(k, n, p) \\ &= P\left\{ \frac{\alpha - np}{\sqrt{np(1-p)}} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{\beta - np}{\sqrt{np(1-p)}} \right\} \\ &\approx \Phi\left(\frac{\beta - np}{\sqrt{np(1-p)}} \right) - \Phi\left(\frac{\alpha - np}{\sqrt{np(1-p)}} \right) \\ &\approx \Phi\left(\frac{\beta + \frac{1}{2} - np}{\sqrt{np(1-p)}} \right) - \Phi\left(\frac{\alpha - \frac{1}{2} - np}{\sqrt{np(1-p)}} \right) \end{aligned}$$



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 $k(\alpha \leq k \leq \beta)$ ，底边长为1，高度为 $b(k; n, p)$ ，这些小矩形面积之和即为 $P(\alpha \leq S_n \leq \beta)$. 再作 $N(np, npq)$ 的密度曲线，在 $[\alpha, \beta]$ 之间曲线覆盖的面积为上式右边之值.

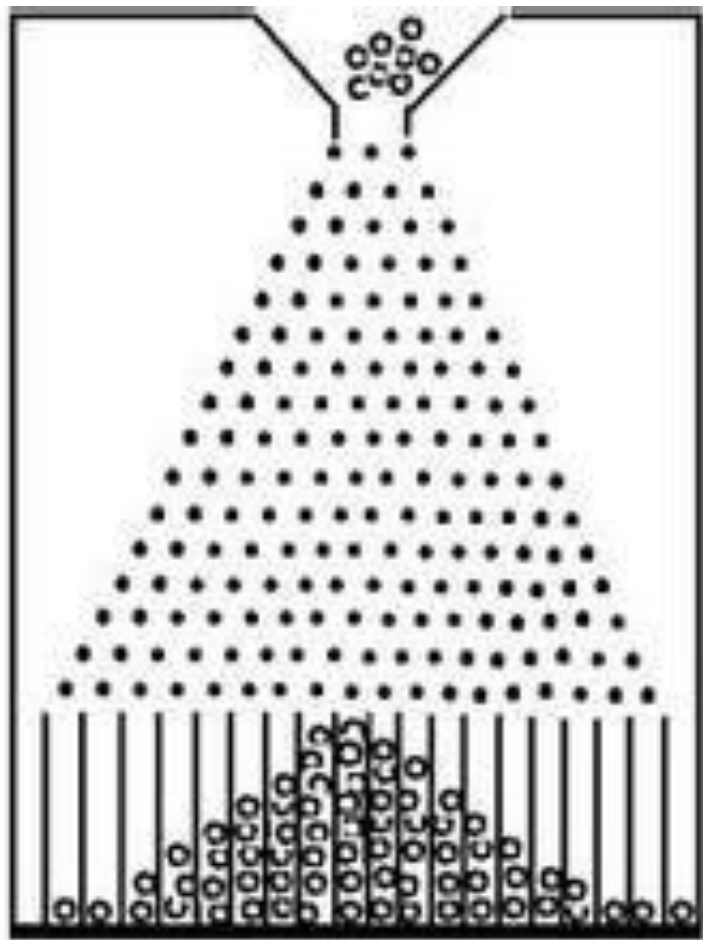


掷三颗骰子，出现点数和 $X=X_1+X_2+X_3$ 的分布律为：
 $\Rightarrow X$ 近似服从正态分布



备注：来自网上图片





Remark 1 第二章讲过二项分布渐近于泊松分布的泊松定理，它与上述定理是没有矛盾的. 因为泊松定理要求 $\lim_{n \rightarrow \infty} np_n = \lambda$ 是常数，而定理中 p 是固定的. 实际应用中，当 n 很大 ($n \geq 30, np \leq 5$) 时

- [1] 若 p 大小适中，可以用正态分布 $\Phi(x)$ 去逼近二项分布的概率;
- [2] 如果 p 接近 0 (或 1), 且 np 较小 (或较大), 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.

Remark 2 中心极限定理有着广泛的应用，在实际工作中，只要 n 足够大，便可以把独立同分布的随机变量和的标准化当作正态变量.



Example 5.2 (正态随机数的产生) 在采样和机器学习中经常需要产生正态分布 $N(\mu, \sigma^2)$ 的随机数，一般统计软件都有产生正态随机数的功能，它是如何产生的呢？

下面介绍用中心极限定理通过 $U(0,1)$ 的随机数来产生正态分布 $N(\mu, \sigma^2)$ 的随机数

方法： 设随机变量 $X \sim U(0,1)$ ，则 $E(X) = \frac{1}{2}$, $D(X) = \frac{1}{12}$

因此，12个相互独立的 $(0,1)$ 上的均匀分布随机变量和的数学期望和方差是6和1

- (1) 从计算机中产生12个服从 $U(0,1)$ 的随机数 x_1, x_2, \dots, x_{12}
- (2) 计算 $y = x_1 + x_2 + \dots + x_{12} - 6$ ，根据中心极限定理， y 可以近似看成来自标准正态分布的一个随机数
- (3) 计算 $z = \sigma y + \mu$ ，则可以将 z 看成来自一个自 $N(\mu, \sigma^2)$ 的随机数
- (4) 重复n次(1)-(3)，就可以得到n个来自 $N(\mu, \sigma^2)$ 的随机数



Example 5.3 (误差分析) 近似计算时, 原始数据 x_k 四舍五入到小数第 m 位, 记为 x'_k , 这时舍入误差 $\varepsilon_k = x_k - x'_k$ 可以看作在 $[-0.5 \times 10^{-m}, 0.5 \times 10^{-m}]$ 上均匀分布, 按四舍五入计算误差和 $\sum_k \varepsilon_k$ 是多少呢?

解: 习惯上人们总是以各误差 ε_k 的和来估计 $\sum_k \varepsilon_k$ 的误差限, 即 $0.5 \times n \times 10^{-m}$. 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 $\{\varepsilon_k\}$ 独立同分布, $E(\varepsilon_k)=0, D(\varepsilon_k)=\sigma^2=10^{-2m}/12$

$$\begin{aligned} P\left(\left|\sum_{k=1}^n \varepsilon_k\right| \leq z\right) &= P\left(\frac{-z}{\sqrt{\frac{10^{-2m}}{12} \cdot n}} \leq \frac{\sum_{k=1}^n \varepsilon_k}{\sqrt{\frac{10^{-2m}}{12} \cdot n}} \leq \frac{z}{\sqrt{\frac{10^{-2m}}{12} \cdot n}}\right) \\ &\approx 2\Phi\left(\frac{z}{\sqrt{\frac{10^{-2m}}{12} \cdot n}}\right) - 1 \\ &= 2\Phi(z') - 1 \end{aligned}$$



若取 $z' = 3$, 上述概率为0.997, 即所有近似数的和的误差超过 $3\sigma\sqrt{n} = 0.5 \times \sqrt{3} \times \sqrt{n} \times 10^{-m}$ 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 $0.5 \times n \times 10^{-m}$ 。

比如在数值计算中, 保留5位小数, 求10000个近似数之和的总误差, 用上式可以估计出有99.7%的概率, 总误差为0.000866, 即万分之8.7

中心极限定理的应用


[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。

给定 n 个相互独立的随机变量 X_1, X_2, \dots, X_n ，令 $Y_n = \sum_{i=1}^n X_i$

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

$$P(Z_n \leq z) \approx \Phi(z') = \beta$$

- 
- 1. 已知 n, z ，求 β
 - 2. 已知 n, β ，求 z
 - 3. 已知 z, β ，求 n



1. 已知 n, z , 求 β

Example 5.4 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

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



所求概率精确的算式为

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

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

$$\begin{aligned} P(29500 < X \leq 30500) &= P\left(\frac{29500 - np}{\sqrt{np(1-p)}} < \frac{X - np}{\sqrt{np(1-p)}} \leq \frac{30500 - np}{\sqrt{np(1-p)}}\right) \\ &\approx \Phi\left(\frac{30500 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{29500 - np}{\sqrt{np(1-p)}}\right) \\ &\approx \Phi\left(\frac{5}{\sqrt{2}}\right) - \Phi\left(-\frac{5}{\sqrt{2}}\right) = 0.9995. \end{aligned}$$

2. 已知 n, β , 求 z

Example 5.5 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

为简化模型，认为每笔保单相互独立。若保险公司希望只有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，即保单 i 的保费 $\pi(X_i) = (1+\theta)E(X_i)$ ，要求估计 θ 的取值。

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

$$\begin{aligned}
ES &= \sum_{i=1}^{1800} EX_i = \sum_{k=1}^4 n_k b_k q_k \\
&= 500 \cdot 1 \cdot 0.02 + 500 \cdot 2 \cdot 0.02 + 300 \cdot 1 \cdot 0.10 + 500 \cdot 2 \cdot 0.10 = 160, \\
VarS &= \sum_{i=1}^{1800} VarX_i = \sum_{k=1}^4 n_k b_k^2 q_k (1 - q_k) \\
&= 500 \cdot 1^2 \cdot 0.02 \cdot 0.98 + 500 \cdot 2^2 \cdot 0.02 \cdot 0.98 \\
&\quad + 300 \cdot 1^2 \cdot 0.10 \cdot 0.90 + 500 \cdot 2^2 \cdot 0.10 \cdot 0.90 \\
&= 256
\end{aligned}$$



由此得保费总额

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

依题意，我们有

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

也即

$$P\left(\frac{S - ES}{\sqrt{\text{Var}S}} \leq \frac{\theta ES}{\sqrt{\text{Var}S}}\right) = P\left(\frac{S - ES}{\sqrt{\text{Var}S}} \leq 10\theta\right) = 0.95.$$

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

$$\theta = 0.1645$$

Remark

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 X_i . 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.

3. 已知 z, β , 求 n

Example 5.6 某调查公司受委托, 调查某电视节目在S市的收视率 p , 调查公司将所有调查对象中收看此节目的频率 \hat{p} 作为 p 的估计, 假定各调查对象是否收看此节目是独立同分布的, 现在要保证有90%的把握, 使得调查所得收视率与真实收视率之间的差异不大于5%, 问至少要调查多少对象?

解: 设共调查 n 个对象, 记 $X_i = 1$ 表示第 i 个调查对象收看此节目, 否则 $X_i = 0$ 。 X_1, \dots, X_n 独立同分布, $E(X_i) = p, D(X_i) = p(1 - p)$

又记 n 个被调查对象中, 收看此节目的人数为 $Y_n = \sum_{i=1}^n X_i \sim B(n, p)$

由大数定理可知, $\hat{p} = \frac{Y_n}{n} \xrightarrow{P} p (n \rightarrow \infty)$

根据中心极限定理

$$P\left(\left|\frac{Y_n}{n} - p\right| \leq 0.05\right) = 2\Phi\left(0.05 \sqrt{\frac{n}{p(1-p)}}\right) - 1 \geq 0.9$$

$$\Phi\left(0.05\sqrt{\frac{n}{p(1-p)}}\right) \geq 0.95$$

查正态分布表可得0.95的分位数为1.645，所以

$$0.05\sqrt{\frac{n}{p(1-p)}} \geq 1.645$$

$$n \geq p(1-p) \frac{1.645^2}{0.05^2} = p(1-p) \times 1082.41$$

又因为 $p(1-p) \leq 0.25$ ，所以 $n \geq 270.6$ ，即至少调查271个对象



Example 5.7 设 $X_1, X_2, \dots, X_n, \dots$ 独立同分布, $E(X_i) = \mu$, $D(X_i) = \sigma^2$ 令,

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

求证: $S_n^2 \xrightarrow{P} \sigma^2$

$$\begin{aligned} S_n^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n} \sum_{i=1}^n ((X_i - \mu) - (\bar{X}_n - \mu))^2 \\ &= \frac{1}{n} \sum_{i=1}^n [(X_i - \mu)^2 - 2(\bar{X}_n - \mu)(X_i - \mu) + (\bar{X}_n - \mu)^2] \\ &= \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - \frac{2}{n} \sum_{i=1}^n (X_i - \mu)(\bar{X}_n - \mu) + (\bar{X}_n - \mu)^2 \\ &= \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - (\bar{X}_n - \mu)^2 \end{aligned}$$



由辛钦大数定律知 $\bar{X}_n \xrightarrow{p} \mu$ ，从而 $(\bar{X}_n - \mu)^2 \xrightarrow{p} 0$ 。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 $S_n^2 \xrightarrow{P} \sigma^2$ 。

[注] 在数理统计中，称 \bar{X}_n 为样本均值，称 $\frac{n}{n-1} S_n^2$ 为样本方差。辛钦大数定律表明样本均值依概率收敛于总体均值。上述例子则表明样本方差依概率收敛于总体方差。



练习

Example 5.8 设某地区内拟筹建一家大型电影院，已知该地区每日平均观影人数为 $n=1600$ 人，预计电影院建成后，平均约有 $3/4$ 的观众将去这家电影院，各个观众是否选择该电影院是独立的。现该电影院在计划其座位数，要求座位尽可能多，但希望空座数达到200甚至更多的可能性不超过0.1，问设置多少座位为好？



谢谢！

