大数定律、中心极限定理

数学科学学院

电子科技大学
UNIVERSITY OF ELECTRONIC SCIENCE
AND TECHNOLOGY OF CHINA

2. 设二维随机变量(X,Y)的联合概率密度为

$$f(x,y)=\begin{cases}\dfrac{6}{(x+y+1)^4}, & (x\ge 0, y\ge 0),\\ 0, & 其它,\end{cases}$$

试求 (1) 条件概率密度 $f_{X|Y}(x|y)$；（2）$P\{0\le X\le 1|Y=1\}$ 。

$$f_{X|Y}(x|y)=\begin{cases}\dfrac{3(y+1)^3}{(x+y+1)^4}, & x\ge 0, y\ge 0\\ 0, & else\end{cases}$$

例 将一枚均匀硬币连续抛 n 次，设 A ={ 出现正面 }， $P(A)=1/2$， $f_n(A)$是事件A发生的频率

问题：是否有 $f_n(A) \to P(A)=1/2$? 是什么收敛性？

实验者	抛掷次数	出现正面次数	正面频率
德. 摩根	2048	1061	0.5181
蒲丰	4040	2048	0.5069
皮尔逊	24000	12012	0.5005
维尼	30000	14994	0.4998

分析：$P(A)=1/2$ 为常数，但 $f_n(A)$ 是数列吗？

$f_n(A)$ 可能偏离$P(A)$很多吗 ?

定义 设$\{X_n\}$是一个随机变量序列，X 是一个随机变量或常数，若对于任意的$\varepsilon > 0$，有

$$\lim_{n\to\infty} P\{|X_n - X| \geq \varepsilon\} = 0 \quad \text{或} \lim_{n\to\infty} P\{|X_n - X| < \varepsilon\} = 1$$

则称随机变量序列$\{X_n\}$依概率收敛于X，记为

$$X_n \xrightarrow{p} X \quad \text{或者} \quad \lim_{n\to\infty} X_n = X, \ (p)$$

$\{X_n\}$ 依概率收敛于X 的含义：

n很大时，X_n 与 X 出现较大偏差的可能性很小

n很大时，有很大把握保证 X_n 与 X 非常接近

例 将一枚均匀硬币连续抛 n 次，设 A ={ 出现正面 }， $P(A)=1/2$， $f_n(A)$是事件A发生的频率。

考虑： $$P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}X_k-\frac{1}{2}\right|\geq\varepsilon\right\}$$

注意到 $$E\left(\frac{1}{n}\sum_{k=1}^{n}X_k\right)=\frac{1}{2}$$

问题：可以用什么工具估计上述概率？

定理 Chebyshev(切比雪夫)不等式

设随机变量X的方差存在，则对于$\varepsilon > 0$有

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

方差刻画了随机变量X关于其数学期望的偏离程度：

方差越小，随机变量越可能待在数学期望的附近

证明：

$$P\left\{\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{2}\right| \geq \varepsilon\right\} \leq \frac{1}{\varepsilon^2} D\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right)$$

代入二项分布的结果，整理得到

$$P\left\{\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{2}\right| \geq \varepsilon\right\} \leq \frac{1}{4n\varepsilon^2}$$

可以证得：频率序列$\frac{1}{n}\sum_{k=1}^{n} X_k$依概率收敛于1/2。

大数定律的定义

设X_k，n=1,2...是一个随机变量序列，其数学期望$E(X_k)$都存在，若对于任意的$\varepsilon > 0$，有

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

即随机变量序列$\{X_k\}$依概率收敛到其数学期望,则称该序列$\{X_k\}$服从大数定律.

$\{X_k\}$服从大数定律的含义：$\{X_k\}$ 前n 项的算术平均将紧密地聚集在其数学期望的附近.

伯努利大数定律：设m是n重伯努利试验中事件A出现的次数，p是试验发生的概率，则频率序列m/n满足大数定律。

回顾： $$P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}X_k - E\left(\frac{1}{n}\sum_{k=1}^{n}X_k\right)\right| \geq \varepsilon\right\} \leq \frac{1}{\varepsilon^2}D\left(\frac{1}{n}\sum_{k=1}^{n}X_k\right)$$

只要有 $\lim\limits_{n\to\infty} D\left(\frac{1}{n}\sum_{k=1}^{n}X_k\right) = 0$ 即满足大数定律

充分条件：

1） 独立时，若$D(X_k)$有统一的上界，则满足

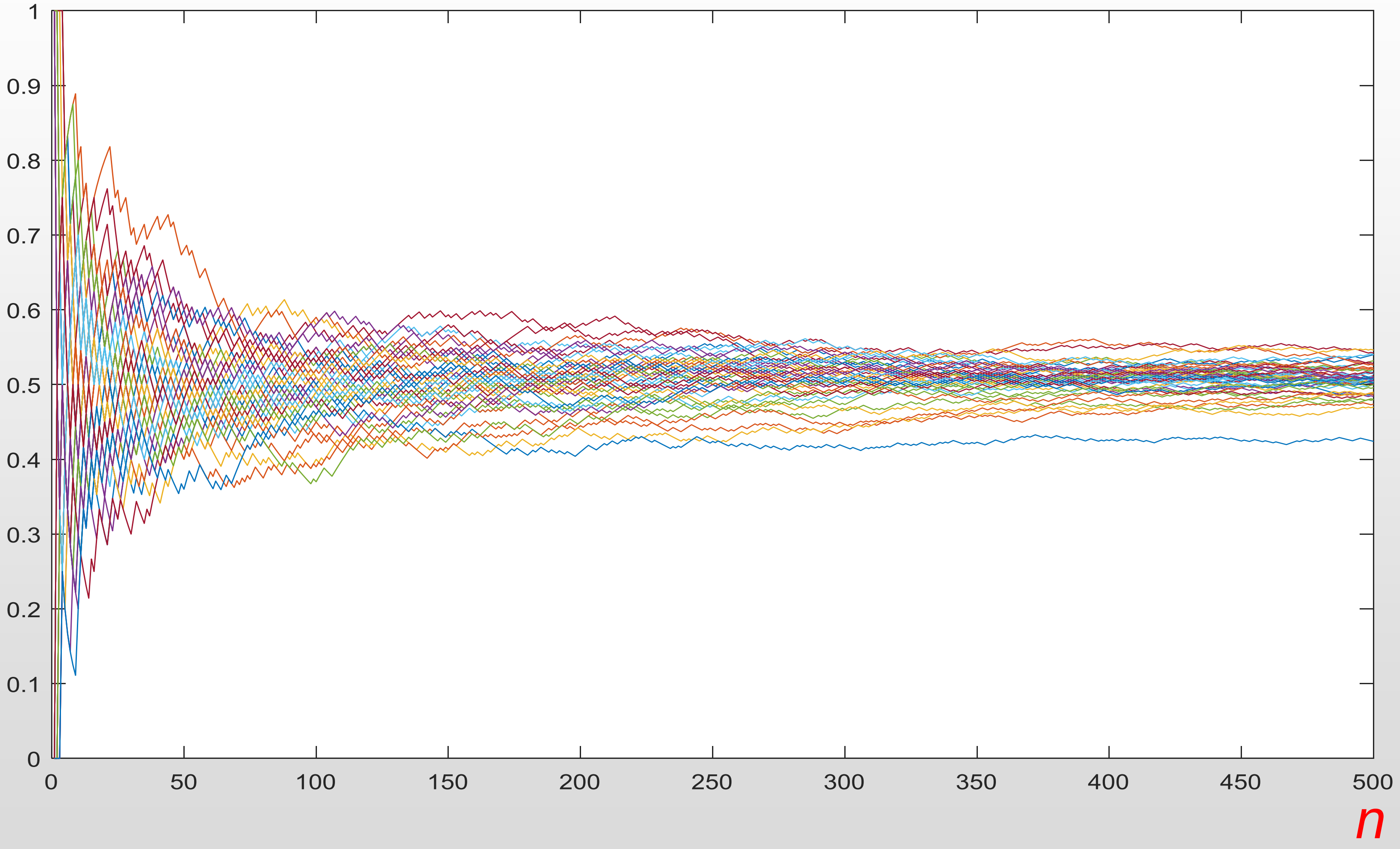
切比雪夫大数定律

2） 独立同分布时，若期望和方差均存在，则满足

独立同分布大数定律

注：不独立时也可能满足大数定律

抛硬币试验正面频率模拟



例： 设随机变量序列X_k相互独立，并且满足

$$P(X_k=\pm\sqrt{k})=\frac{1}{k},\qquad P(X_k=0)=1-\frac{2}{k}$$

证明随机变量序列都满足大数定律

解： $E(X_k^2)=2$，因此期望和方差均存在

进一步有$E(X_k)=0$，$D(X_k)=2$，方差有统一的上界

满足切比雪夫大数定律

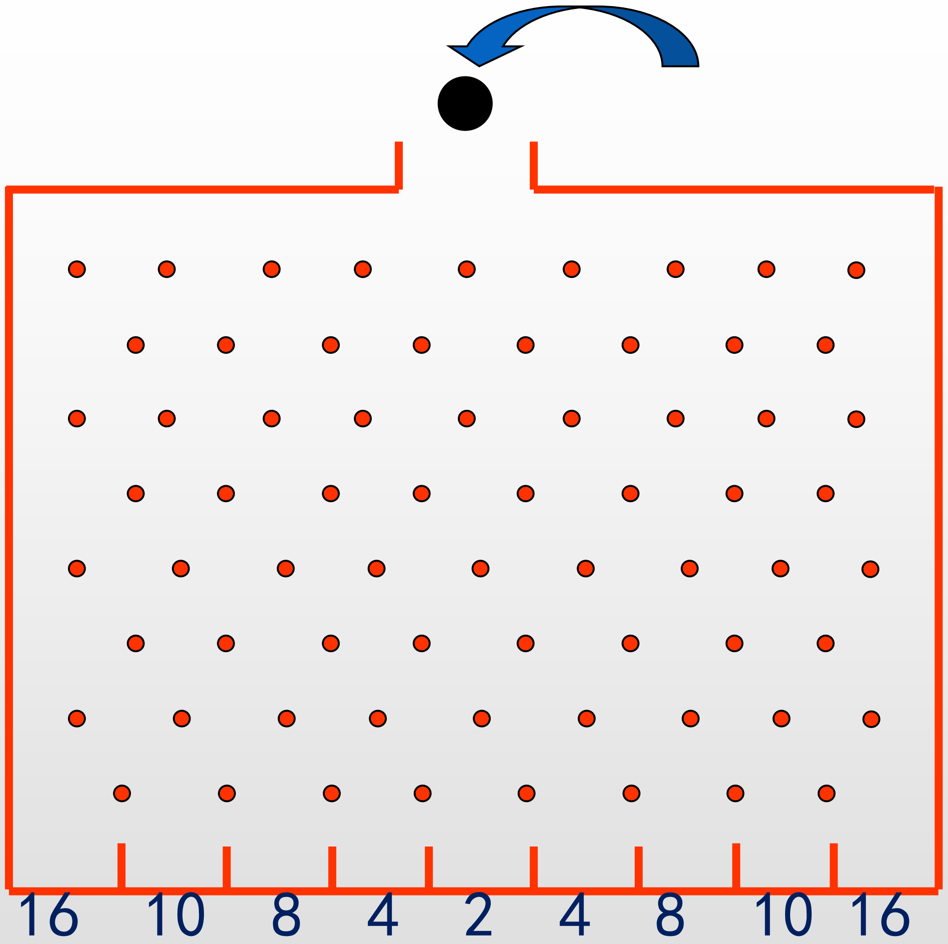
中心极限定理

引例　　奖券设置问题：

把小球从游戏板上方入口投入，逐层碰到板上钉子下落，落入底层格中对应数字则为游戏者获得奖券数

奖券设置有何规律

为何如此设置



关注：小球下落到底层的位置

需描述小球的下落过程

假设： 1.共有n层钉子；

2.小球入口处对应水平位置为坐标原点0；

3.小球在每层碰到钉子后，向左或向右等可能位移一格，不会出现跳格（位移２格以上）的情况；

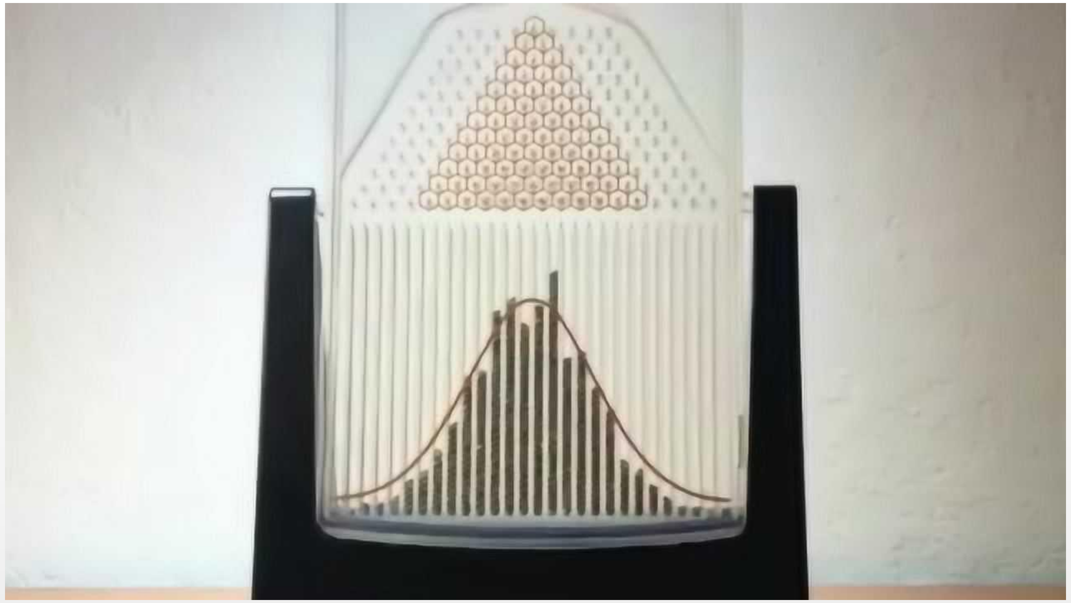
4.小球在不同层向左或向右是相互独立的.

则随机变量序列：
$$X_k=\begin{cases}1 & \text{第}k\text{层向右}\\ -1 & \text{第}k\text{层向左}\end{cases}\qquad k=1,\ldots,n$$

完整描述了小球的运动过程，且

X_k	-1	1
p	1/2	1/2

最终位置：
$$Y_n=\sum_{k=1}^{n}X_k$$



事实上，可严格证明：在高尔顿钉板试验中，

小球位置的标准化随机变量序列

$$\frac{Y_n - E(Y_n)}{\sqrt{D(Y_n)}}$$ 趋向于标准正态分布随机变量

问题：什么叫趋向于？

定义：设随机变量序列X_n的分布函数为$F_n(x)$，X的分布函数为$F(x)$，如果在$F(x)$的连续点x处均有

$$\lim_{n\to\infty} F_n(x) = F(x)$$

则称随机变量序列X_n依分布收敛到X

独立同分布中心极限定理

设$\{X_k\}$，$k=1,2\ldots$是一个相互独立，具有相同分布的随机变量序列，且$E(X_k)=\mu$，$D(X_k)=\sigma^2\neq 0$，（$k=1,2,\ldots$），则有

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

即序列和的标准化依分布收敛到标准正态分布。

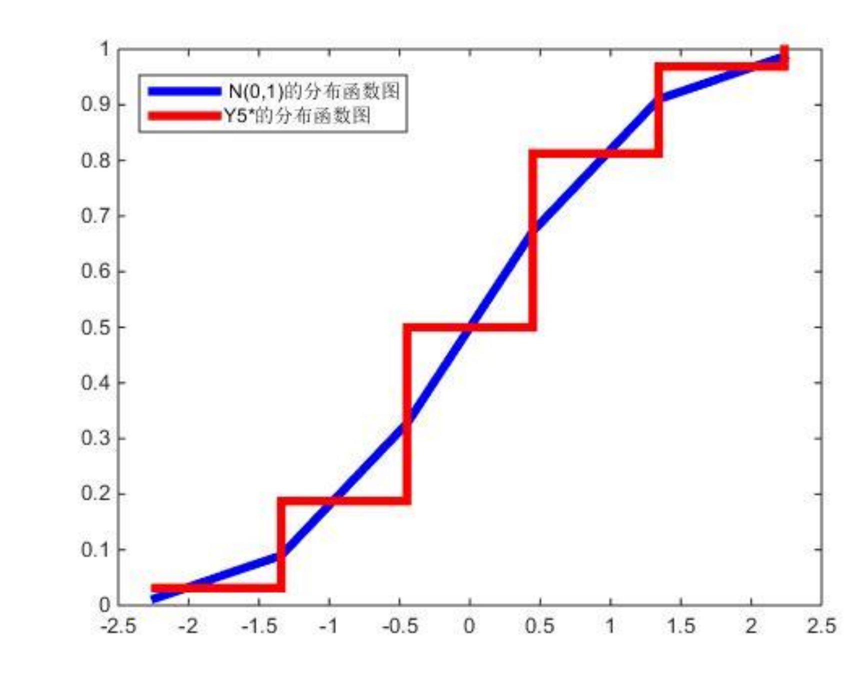
注：对X_k的分布没任何要求！

具体展示

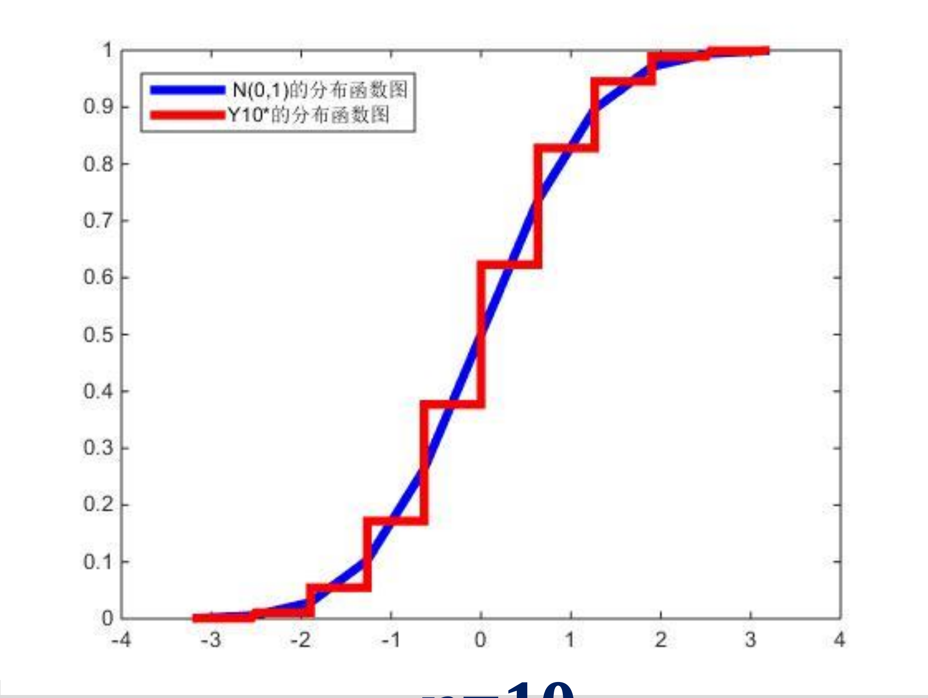
基于钉板试验中X_k所满足条件：独立，同两点分布

计算小球位置的标准化随机变量序列 $Y_n^* = \dfrac{\sum_{k=1}^{n} X_k}{\sqrt{n}}$

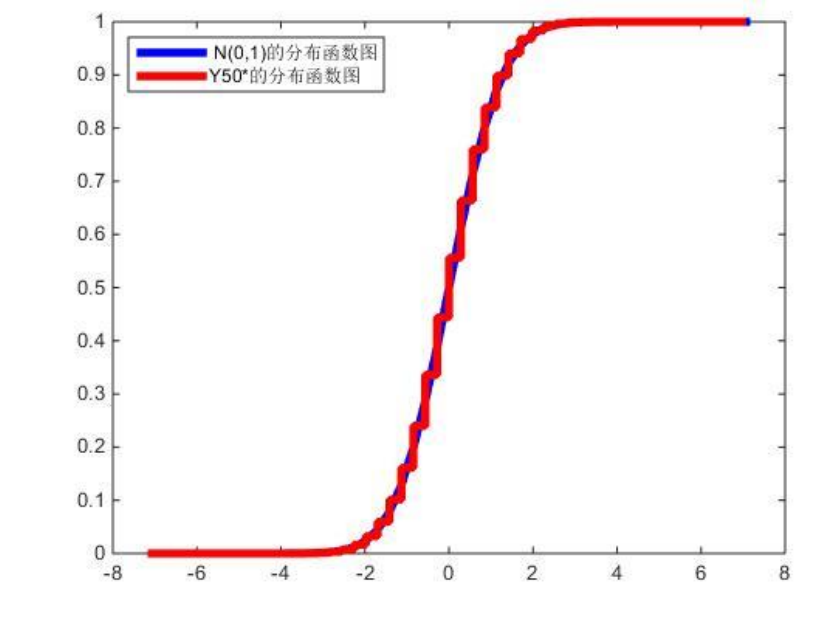
的分布函数(红色)，与**标准正态**分布函数(蓝色)对比：



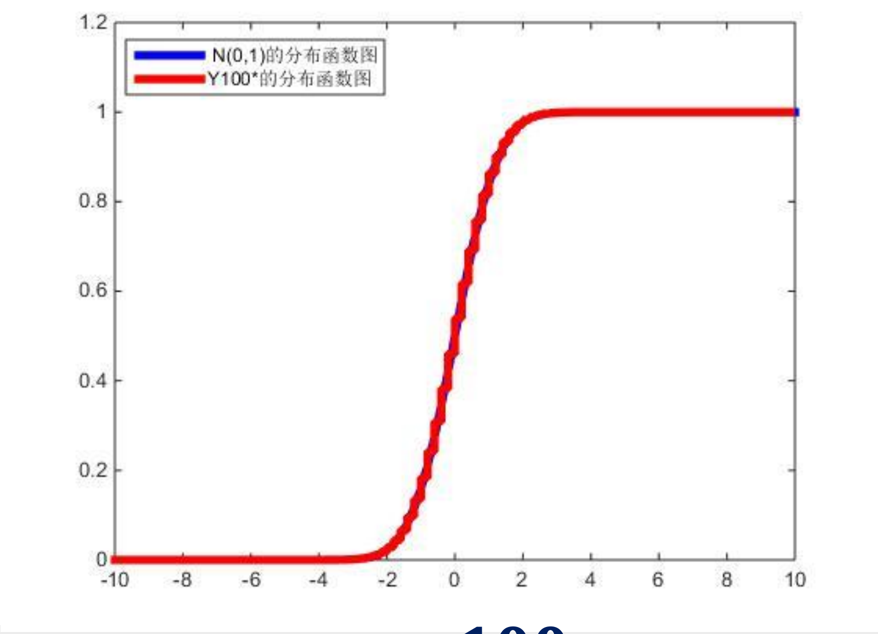
n=5



n=10



n=50



n=100

可从理论上严格证明：当钉子层数n→∞时，

Y_n^*的分布函数 收敛于 标准正态分布函数$\Phi(x)$

称 Y_n^*依分布收敛于 标准正态分布随机变量

注1 若随机变量序列$\{X_k\}$，$k=1,2,\ldots$服从中心极限定理，则可进行概率的近似计算

$$\lim_{n\to\infty} P\left\{x_1 \le \frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\sigma} \le x_2\right\} = \Phi(x_2) - \Phi(x_1)$$

即当n足够大时，可认为

$$\frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\sigma} \sim N(0,1)$$ 近似成立

即$\sum_{k=1}^{n} X_k \sim N(n\mu, n\sigma^2)$ 近似成立

若某随机变量Y可被拆分为多个（n个）独立同分布随机变量之和，即$Y=\sum_{i=1}^{n}X_i$，则根据独立同分布中心极限定理，当n很大时，Y/n的标准化随机变量近似服从正态分布。

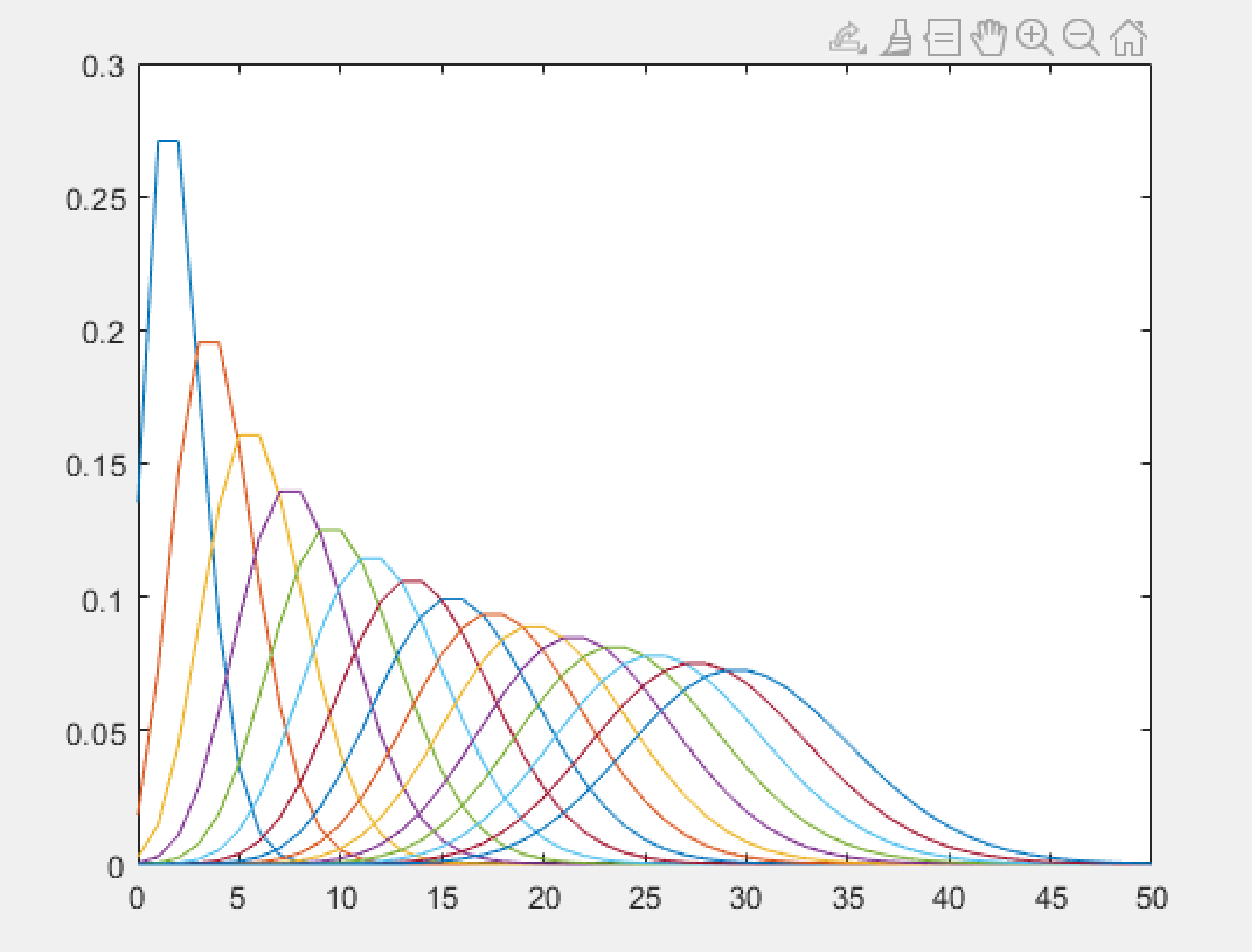
可加性

例1：$Y \sim P(n)$ $X_k \sim P(1)$

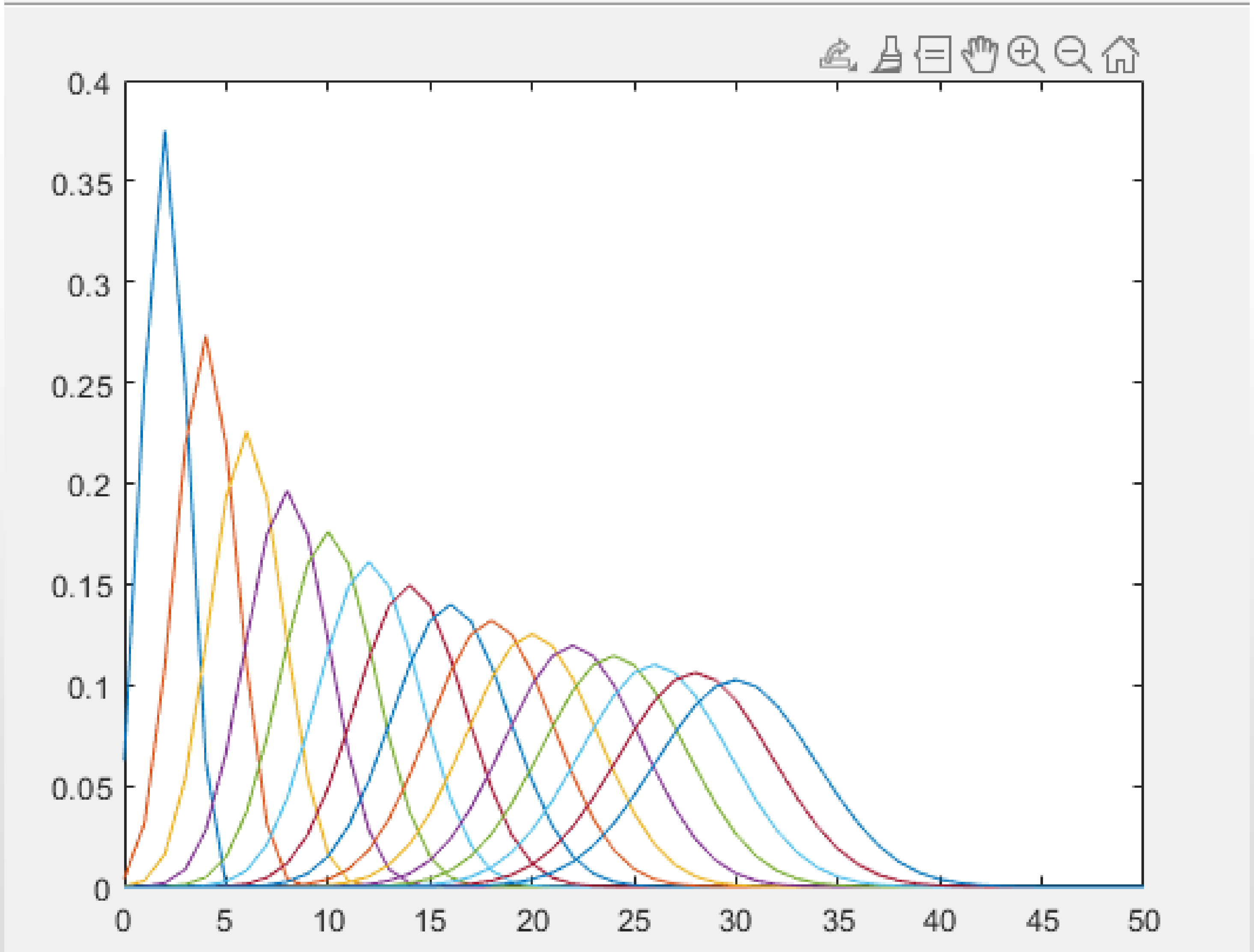
当n很大时，近似有 $$\frac{\sum_{k=1}^{n}X_k-n}{\sqrt{n}} \sim N(0,1)$$

此时Y近似服从正态分布

例2：$Y \sim B(n,p)$ $X_k \sim B(1,p)$ 棣莫弗–拉普拉斯中心极限定理



参数为$2k$的泊松分布的分布律，k=1，2，...,15



参数为(4*k*,0.5)的二项分布的分布律，k=1，2，...,15

中心极限定理解释了现实中哪些随机变量可看作正态分布，为什么统计学中大量出现正态分布.

“独立同分布”条件的理解

$\sum_{i=1}^{n} X_i$ 中各X_i相互独立，且“均匀的小”

叠加的结果：正态分布

实际中的例子？ 人的身高和体重、智力测验分数、考试成绩、制造过程中的测量误差、金融市场的回报率、生物科学中的测量、农业产量、交通流量、电子元件的性能参数

正态分布背景及定义

正态分布（**Normal distribution**），由法国裔英国籍的数学家**棣莫弗**（**De Moivre**）于1733年首次提出。他在考虑二项分布的极限分布时，用阶乘的近似公式导出了「正态分布」的密度曲线。

但因德国数学家**高斯**（**Gauss**）率先将正态分布应用于误差分布与最小二乘法，且此工作对现代数理统计学影响极大，故正态分布又叫**高斯分布**（**Gaussian distribution**）

例 路边有一个售报亭，每个过路人在报亭买报的概率是1/3，求：正好售出100份报纸时的过路人数在280到300之间的概率。

解 设 X 是正好售出100份报纸时的过路人数，X_i 是售出第 $i-1$ 份报纸后到售出第 i 份报纸时的过路人数，则

$$X=\sum_{i=1}^{100} X_i$$

并且随机变量 $X_1, X_2, \cdots, X_{100}$ 独立同分布，具有分布律:

$$P\{X_i = k\} = \frac{1}{3}\left(\frac{2}{3}\right)^{k-1}, \qquad k = 1,2,\cdots\cdots$$

因

$$E(X_i) = \frac{1}{\frac{1}{3}} = 3, \qquad D(X_i) = \frac{\frac{2}{3}}{\left(\frac{1}{3}\right)^2} = 6$$

$$i = 1, 2, \cdots, 100;$$

根据独立同分布中心极限定理，所求概率

$$P\{280 < \sum_{i=1}^{100} X_i < 300\}$$

$$= P\left(\frac{280 - 100 \times 3}{\sqrt{100 \times 6}} < \frac{\sum_{i=1}^{100} X_i - 100 \times 3}{\sqrt{100 \times 6}} < \frac{300 - 100 \times 3}{\sqrt{100 \times 6}}\right)$$

$$\approx \Phi\left(\frac{300 - 100 \times 3}{\sqrt{100 \times 6}}\right) - \Phi\left(\frac{280 - 100 \times 3}{\sqrt{100 \times 6}}\right)$$

$$= \Phi(0) - \Phi(-0.8165)$$

$$= 0.5 - 1 + \Phi(0.8165)$$

$$= 0.293$$

例 将一枚均匀硬币连续抛 n 次，试用中心极定理来估计 n，使下式成立.

若用切比雪夫不等式估计？

$$P\{|f_n(A)-P(A)|<0.01\}\geq 0.99$$

其中 $A=\{$ 出现正面 $\}$, 已知$\Phi(2.58)=0.995$

解 有$P(A)=1/2$，令

$$X_i=\begin{cases}1, & 第i次出现正面;\\ 0, & 否则,\end{cases}\qquad (i=1,2,\cdots n)$$

则随机变量序列$\{X_i\}$，$i=1,2,\ldots$是相互独立且同分布的. 而且有

$$E(X_i)=\frac{1}{2},\quad D(X_i)=\frac{1}{4},\quad i=1,2,\cdots$$

所以随机变量序列$\{X_i\}$，满足独立同分布中心极限定理.

有 $$f_n(A)=\frac{1}{n}\sum_{i=1}^{n}X_i$$,由题意可得

$$0.99\le P\{|f_n(A)-P(A)|<0.01\}$$

$$=P\left\{\frac{1}{2}-0.01<\frac{1}{n}\sum_{i=1}^{n}X_i<\frac{1}{2}+0.01\right\}$$

$$=P\left\{\frac{n}{2}-0.01n<\sum_{i=1}^{n}X_i<\frac{n}{2}+0.01n\right\}$$

$$= P\{-\frac{0.01n}{1/2\sqrt{n}} < \frac{\sum_{i=1}^{n} X_i - \frac{n}{2}}{1/2\sqrt{n}} < \frac{0.01n}{1/2\sqrt{n}}\}$$

因为 $\dfrac{\sum_{i=1}^{n} X_i - \frac{n}{2}}{1/2\sqrt{n}}$ 近似服从$N(0,1)$分布，

$$\text{所以} 0.99 \le P\{-\frac{0.01n}{1/2\sqrt{n}} < \frac{\sum_{i=1}^{n} X_i - \frac{n}{2}}{1/2\sqrt{n}} < \frac{0.01n}{1/2\sqrt{n}}\}$$

$$\approx 2\Phi(0.02\sqrt{n}) - 1$$

$$\Rightarrow \Phi(0.02\sqrt{n}) \ge 0.995 \Rightarrow 0.02\sqrt{n} \ge 2.58$$

解得 $n \ge 16641$ (次)

(对比切比雪夫不等式所得结果250000次？)

例 随机抽查验收产品，如果在一批产品中查出10个以上的次品，则拒绝接收.问至少检查多少个产品，能保证次品率为 10%的一批产品被拒收的概率不低于0.9. ($\varphi(1.28)=0.9$)

解 设检查的产品数为 n，查出的次品数为X，则$X\sim B(n,\ 0.1)$，按题意，有

$$P\{10\le X\le n\}\ge 0.9$$

由中心极限定理，当n很大时近似地有

$$X\sim N(0.1n,\ 0.09n)$$

因此 $P\{10 \le X \le n\} \approx \Phi\left(\dfrac{n-0.1n}{\sqrt{0.1\times 0.9n}}\right)-\Phi\left(\dfrac{10-0.1}{\sqrt{0.1\times 0.9n}}\right)$

$$= \Phi(3\sqrt{n})-\Phi\left(\frac{10-0.1n}{0.3\sqrt{n}}\right)$$

$$\approx 1-\Phi\left(\frac{10-0.1n}{0.3\sqrt{n}}\right) = \Phi\left(\frac{0.1n-10}{0.3\sqrt{n}}\right) \ge 0.9$$

$$\Rightarrow \frac{0.1n-10}{0.3\sqrt{n}} \ge 1.28$$

求解得 $n \ge 146.8$ 或 $n \le -68.3$,

所以至少取 $n=147$ 能够保证要求。

*60. 用概率论方法证明: 当 $n \to \infty$ 时, $e^{-n}\sum_{k=0}^{n}\frac{n^k}{k!} \to \frac{1}{2}$.

推广：设$a>0$是正数，$\lfloor\ \rfloor$表示取整函数，计算 $\lim_{n\to\infty} e^{-n}\sum_{k=0}^{\lfloor an\rfloor}\frac{n^k}{k!}=?$

解：令$X\sim P(n)$，则$e^{-n}\sum_{k=0}^{\lfloor an\rfloor}\frac{n^k}{k!}=P(X\le\lfloor an\rfloor)$

设$X_i\sim P(1)$，且X_i相互独立，则$\sum_{k=1}^{n}X_i$与X同分布

由中心极限定理有

$$P(X\le\lfloor an\rfloor)=P\left(\sum_{k=1}^{n}X_i\le\lfloor an\rfloor\right)=P\left(\frac{\sum_{k=1}^{n}X_i-n}{\sqrt{n}}\le\frac{\lfloor an\rfloor-n}{\sqrt{n}}\right)\approx\Phi\left(\frac{\lfloor an\rfloor-n}{\sqrt{n}}\right)$$

$$\lim_{n\to\infty}\Phi\left(\frac{\lfloor an\rfloor-n}{\sqrt{n}}\right)=\begin{cases}0, & a<1\\ \frac{1}{2}, & a=1\\ 1, & a>1\end{cases}$$