第五章 大数定律及中心极限定理

§1 大数定律

§2 中心极限定理

大数定律的一般形式

设$\{X_n, n = 1,2,\dots\}$是一个随机变量序列，而且对每个n，$E(X_n)$存在，如果对于任意给定的$\epsilon > 0$ ，有

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

则称随机变量序列$\{X_n\}$服从大数定律．

等价形式：

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

即$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{P} \frac{1}{n}\sum_{k=1}^{n} E(X_k)$

§2 中心极限定理

大数定律→收敛

→如何收敛？

→中心极限定理

问题的引出

考察射击命中点与靶心距离的偏差，偏差可能的因素

瞄准误差

子弹制造过程方面 (如外形、重量等) 的误差

射击时武器的振动

气象因素(如风速、风向、能见度、温度等)

……

问题： 某个随机变量是由大量相互独立且都很小的随机变量相加而成的, 研究其概率分布情况

一般地，若某项偶然因素对总和的影响是均匀的、微小的，即没有一项起特别突出的作用，则这些大量独立偶然因素总和的随机变量近似服从正态分布

——中心极限定理试图证明了这一假设的正确性

研究在什么条件下，大量独立随机变量和的分布以正态分布为极限，这一类定理称为中心极限定理。

定义：设$X_1, X_2, \dots, X_n, \dots$是独立的随机变量序列，若$E(X_k), D(X_k)\ (k = 1,2,\dots)$都存在，令

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

若对任意$x \in R$，有，

$$\lim_{n\to\infty} P\{Y_n \le x\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

则称该随机变量序列服从中心极限定理

等价地，若$\{X_n\}$服从中心极限定理，则当n足够大时Y_n近似服从标准正态分布

定理 1（独立同分布的中心极限定理）

设随机变量$X_1, X_2, \ldots, X_n, \ldots$独立同分布，且$E(X_k)=\mu,\ D(X_k)=\sigma^2>0,\ (k=1,2,\ldots)$，则随机变量之和$\sum_{k=1}^{n} X_k$的标准化变量

$$Y_n=\frac{\sum_{k=1}^{n} X_k-E\left(\sum_{k=1}^{n} X_k\right)}{\sqrt{D\left(\sum_{k=1}^{n} X_k\right)}}=\frac{\sum_{k=1}^{n} X_k-n\mu}{\sqrt{n}\sigma}$$

的分布函数$F_n(x)$对于任意的x满足

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

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

对中心极限定理的理解:

1. 当$n \to \infty$时，随机变量序列Y_n的分布函数收敛于标准正态分布
   1. 注意这与X_i的分布无关
   2. 大众取得共识呈正态分布
2. 等价表示——若$X_1, X_2, \dots, X_n, \dots$独立同分布，且$E(X_k) = \mu,\ D(X_k) = \sigma^2 > 0,\ (k = 1,2,\dots)$，则对充分大的n

   2.1 $\frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\sigma}$ (或等价地$\frac{\frac{1}{n}\sum_{k=1}^{n} X_k - \mu}{\sigma/\sqrt{n}}$)近似服从标准正态分布$N(0,1)$

   2.2 $\sum_{k=1}^{n} X_k$近似服从正态分布$N(n\mu, n\sigma^2)$

对中心极限定理的理解:

3. 提供了相应的计算方法

$$P\left\{a<\sum_{k=1}^{n}X_k<b\right\}$$

$$=P\left\{\frac{a-n\mu}{\sqrt{n}\sigma}<\frac{\sum_{k=1}^{n}X_k-n\mu}{\sqrt{n}\sigma}<\frac{b-n\mu}{\sqrt{n}\sigma}\right\}$$

$$\approx\Phi\left(\frac{b-n\mu}{\sqrt{n}\sigma}\right)-\Phi\left(\frac{a-n\mu}{\sqrt{n}\sigma}\right)$$

$$P\{a<\bar{X}<b\}=P\left\{\frac{a-\mu}{\sigma/\sqrt{n}}<\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}<\frac{b-\mu}{\sigma/\sqrt{n}}\right\}$$

$$\approx\Phi\left(\frac{b-\mu}{\sigma/\sqrt{n}}\right)-\Phi\left(\frac{a-\mu}{\sigma/\sqrt{n}}\right)$$

离散情形，X_i的分布为：

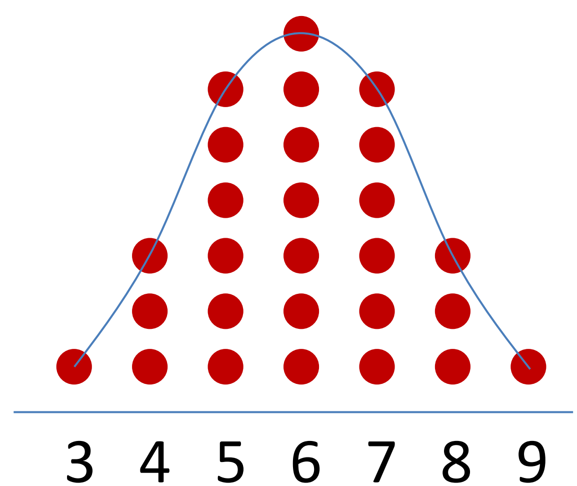
X_i	1	2	3
$P(X_i)$	1/3	1/3	1/3

三个独立同分布的和：

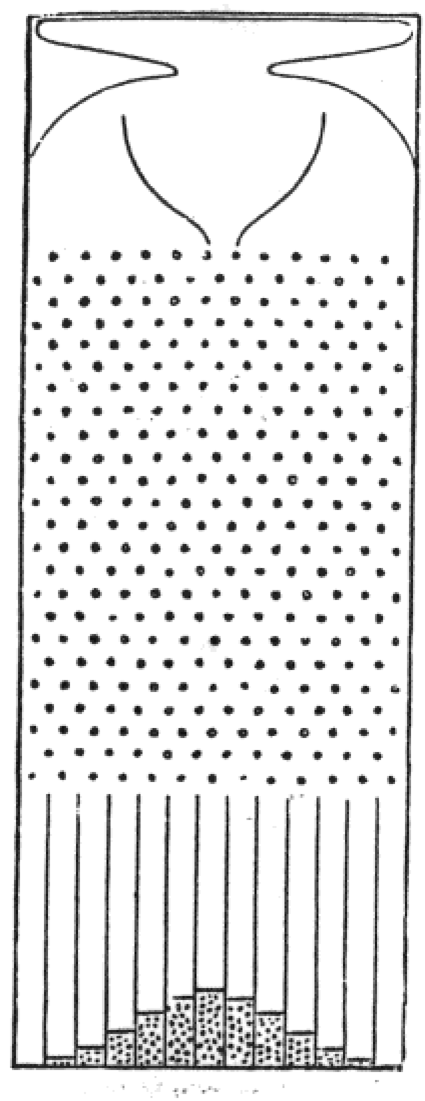
X_1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3
X_2	1	1	1	2	2	2	3	3	3	1	1	1	2	2	2	3	3	3	1	1	1	2	2	2	3	3	3
X_3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Σ	3	4	5	4	5	6	5	6	7	4	5	6	5	6	7	6	7	8	5	6	7	6	7	8	7	8	9

和的概率分布

$X_1+X_2+X_3$	3	4	5	6	7	8	9
$P(\sum X_i)$	$\frac{1}{27}$	$\frac{3}{27}$	$\frac{6}{27}$	$\frac{7}{27}$	$\frac{6}{27}$	$\frac{3}{27}$	$\frac{1}{27}$

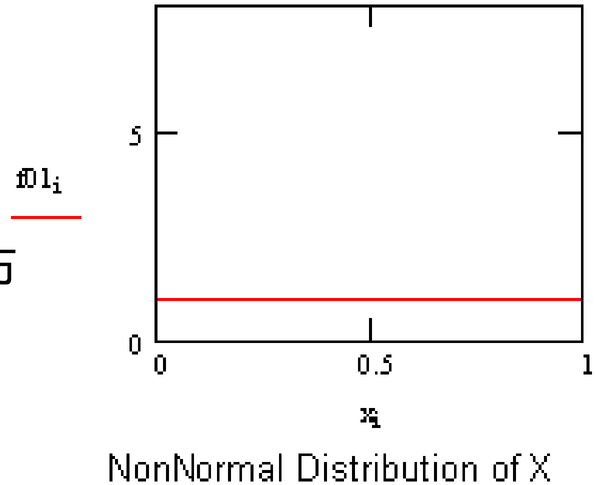


Bean Machine (Galton Board)

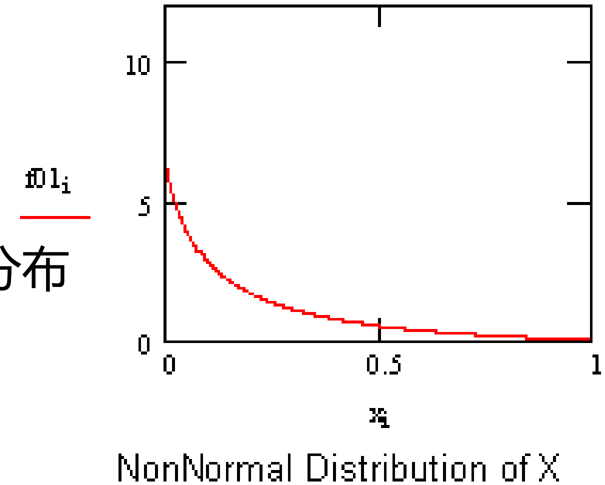


独立同分布的中心极限定理动画模拟

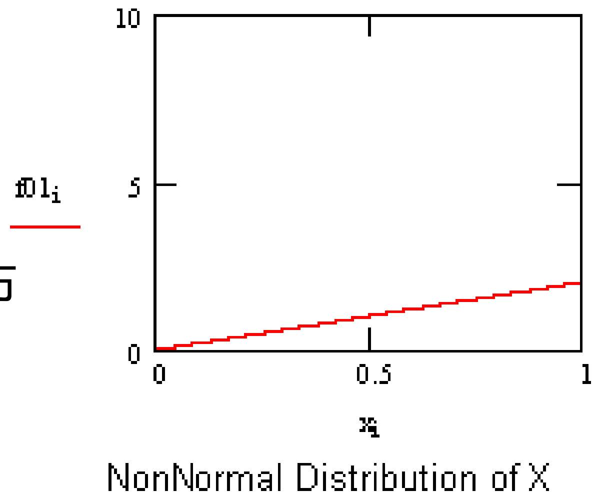
均匀分布



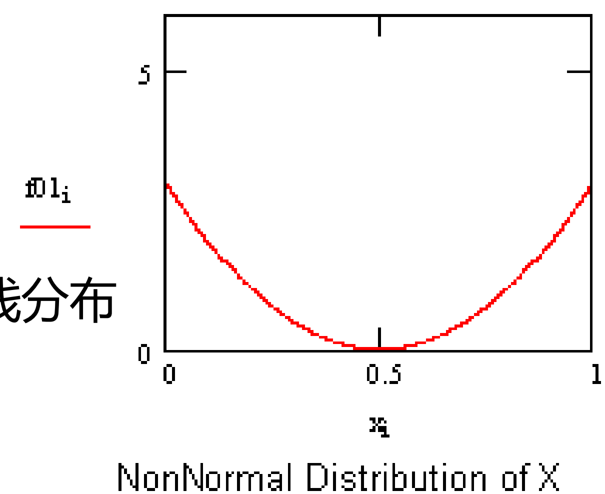
$1/x$分布



三角分布



抛物线分布



GIFs from
http://www.statisticalengineering.com/central_limit_theorem_(summary).html

定理 2 (Lyapunov定理)

设随机变量$X_1, X_2, \ldots, X_n, \ldots$相互独立，且$E(X_k) = \mu_k$，$D(X_k) = \sigma_k^2 > 0, (k = 1,2, \ldots)$，记$B_n^2 = \sum_{k=1}^n \sigma_k^2$，

若存在$\delta > 0$，使得当$n \to \infty$时

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

则随机变量之和$\sum_{k=1}^n X_k$的标准化变量

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

的分布函数$F_n(x)$对于任意的x满足

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} P\{\frac{\sum_{k=1}^n X_k - \sum_{k=1}^n \mu_k}{B_n} \le x\}$$

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

Lyapunov定理表明：

1. 只要独立性即可，不一定要同分布。
2. 无论$X_1, X_2, \ldots, X_n, \ldots$服从何种分布，只要满足定理的条件，其和$\sum x_k$在n足够大的时候都近似服从正态分布

定理3.(De Moivre – Laplace定理)

设随机变量$\eta_n(n=1,2,\ldots)$服从参数为$n,p\ (0<p<1)$的二项分布，则对于任意x，有

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

证明：令

$$X_k=\begin{cases}0, & \text{在第}k\text{次实验中}A\text{不发生}\\ 1, & \text{在第}k\text{次实验中}A\text{发生}\end{cases},\qquad k=1,2,\ldots$$

则$\{X_k\}$是相互独立的随机变量序列，于是

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

由独立同分布中心极限定理得

$$\lim_{n\to\infty} P\left\{\frac{\eta_n - np}{\sqrt{np(1-p)}} \le x\right\}$$

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

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

定理表明：

二项分布的极限分布是正态分布，即当n充分大时,
可以用正态分布逼近二项分布

由上面的定理知

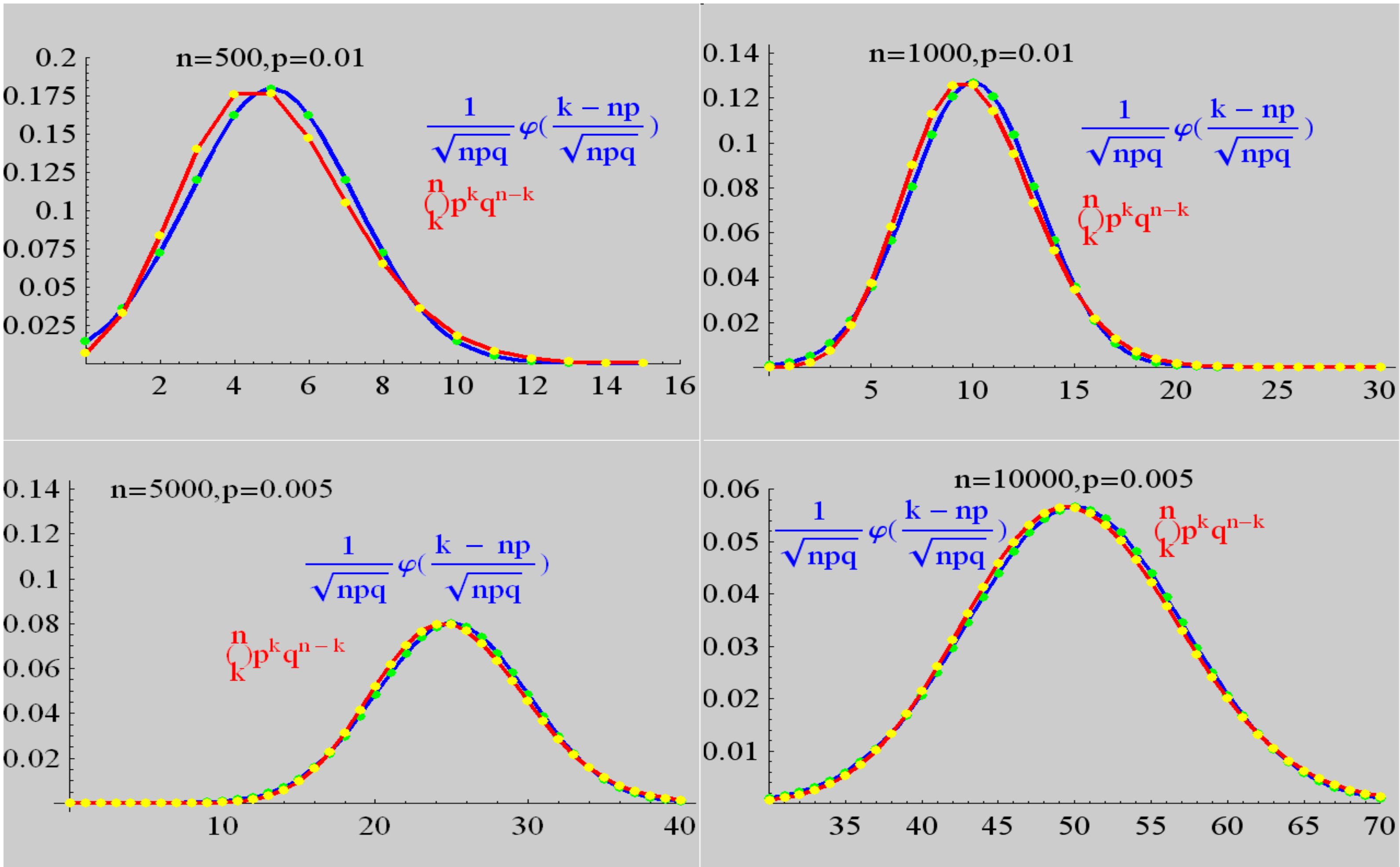
$$P\left\{\left|\frac{n_A}{n}-p\right| \le \epsilon\right\} = P\left\{\left|\frac{n_A-np}{n}\right| \le \epsilon\right\}$$

$$= P\left\{\left|\frac{n_A-np}{\sqrt{np(1-p)}}\right| \le \epsilon\sqrt{\frac{n}{p(1-p)}}\right\}$$

$$= \Phi\left(\epsilon\sqrt{\frac{n}{p(1-p)}}\right) - \Phi\left(-\epsilon\sqrt{\frac{n}{p(1-p)}}\right)$$

$$= 2\Phi\left(\epsilon\sqrt{\frac{n}{p(1-p)}}\right) - 1$$

下面的图形表明:正态分布是二项分布的逼近.



大数定律回答了用频率估计概率的合理性问题，De Moivre-Laplace中心极限定律回答了用频率估计概率的误差问题

中心极限定理阐明了在什么条件下，原本不属于正态分布的一些随机变量其总和分布渐近服从正态分布

大数定律是研究随机变量序列依概率收敛的极限问题

两者均是大量的随机变量之和的极限形式，当 $\mathrm{X}_1, X_2, \ldots, \mathrm{X}_n, \ldots$相互独立同分布且有大于零的有限方差时，两者同时成立，否则关系不确定

例1. 一个加法器同时收到20个噪声电压V_k $(k=1,2,\ldots,20)$，设它们是相互独立的随机变量，且服从(0,10)上的均匀分布，记$V=\sum_{k=1}^{20}V_k$，求$P\{V>105\}$的近似值。

解：$E(V_k)=5, D(V_k)=\frac{100}{12}, (k=1,2,\ldots,20)$，

令

$$Z=\frac{\sum_{k=1}^{20}V_k-20\times 5}{\sqrt{20}\sqrt{\frac{100}{12}}}=\frac{V-100}{10\sqrt{5}}\sqrt{3}$$

由独立同分布中心极限定理，随机变量Z近似服从$N(0,1)$

于是

$$P\{V>105\}=P\left\{\frac{V-100}{10\sqrt{5}}\sqrt{3}>\frac{105-100}{10\sqrt{5}}\sqrt{3}\right\}$$

$$=1-P\left\{\frac{V-100}{10\sqrt{5}}\sqrt{3}<\frac{105-100}{10\sqrt{5}}\sqrt{3}\right\}$$

$$=1-\int_{-\infty}^{0.387}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt=1-\Phi(0.387)$$

$$=0.348$$

例2. 一船舶在某海区航行, 已知每遭受一次海浪的冲击, 纵摇角大于 3° 的概率为1/3, 若船舶遭受了90,000次波浪冲击, 问其中有29,500 ~ 30,500次纵摇角大于 3° 的概率是多少？

解： 将船舶每遭受一次海浪的冲击看作一次试验, 并假设各次试验是独立的，在90 000次波浪冲击中纵摇角大于 3° 的次数为X，则X是一个随机变量，

且$X\sim b(90{,}000,\frac{1}{3})$。

分布律为：

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

$k=1,2,\dots,90000$

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

为避免直接计算，采用De Moivre-Laplace中心极限定律

$$P\{29500 \le X \le 30500\}$$

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

$$\approx \int_{\frac{29500-np}{\sqrt{np(1-p)}}}^{\frac{30500-np}{\sqrt{np(1-p)}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \int_{-\frac{5}{\sqrt{2}}}^{\frac{5}{\sqrt{2}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

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

例3. 对于一个学生而言, 来参加家长会的家长人数是一个随机变量. 设一个学生无家长、1名家长、 2名家长来参加会议的概率分别为0.05，0.8，0.15. 若学校共有400名学生, 设各学生参加会议的家长数相互独立, 且服从同一分布. (1) 求参加会议的家长数X 超过450的概率; (2) 求有1名家长来参加会议的学生数不多于340的概率。

解：(1) 以X_k $(k=1,2,\dots,400)$记第k个学生来参加会议的家长人数

则X_k的分布律为

X_k	0	1	2
p_k	0.05	0.8	0.15

于是$E(X_k)=1.1$， $D(X_k)=0.19$，$(k=1,2,\dots,400)$

参会人数为：$X=\sum_{k=1}^{400} X_k$，由独立同分布的中心极限定理，随机变量

$$Z=\frac{\sum_{k=1}^{400} X_k-400\times 1.1}{\sqrt{400}\sqrt{0.19}}=\frac{X-440}{\sqrt{400}\sqrt{0.19}}$$

近似服从正态分布$N(0,1)$

于是，$P\{X>450\}$

$$=P\left\{\frac{X-440}{\sqrt{400}\sqrt{0.19}}>\frac{450-440}{\sqrt{400}\sqrt{0.19}}\right\}$$

$$=1-P\left\{\frac{X-440}{\sqrt{400}\sqrt{0.19}}\leq 1.147\right\}$$

$$\approx 1-\Phi(1.147)$$

$$=0.1357$$

(2) 以Y记有一名家长来参加会议的学生数，

则$Y \sim b(400,0.8)$， De Moivre-Laplace定律知，

$$P\{X \le 340\}$$

$$= P\left\{\frac{Y - 400 \times 0.8}{\sqrt{400 \times 0.8 \times 0.2}} \le \frac{340 - 400 \times 0.8}{\sqrt{400 \times 0.8 \times 0.2}}\right\}$$

$$= P\left\{\frac{Y - 400 \times 0.8}{\sqrt{400 \times 0.8 \times 0.2}} \le 2.5\right\} \approx \Phi(2.5) = 0.9938$$

例4. 一个螺丝钉重量是一个随机变量，期望值是1两，标准差是0.1两。求一盒(100个)同型号螺丝钉的重量超过10.2斤的概率。

解：设第i个螺丝钉的重量为X_i，一盒的总重量为：$X=\sum_{i=1}^{100} X_i$

$X_1, X_2, \dots, X_{100}$相互独立，$E(X_i)=1$，$D(X_i)=0.1^2$

$$E(X)=\sum_{i=1}^{100} E(X_i)=100, D(X)=\sum_{i=1}^{100} D(X_i)=1$$

X近似服从正态分布$N(100,1)$

$$P(X>102)=P\left(\frac{X-100}{1}>\frac{102-100}{1}\right)$$

$$=1-P\left(\frac{X-100}{1}\le 2\right)\approx 1-\Phi(2)=0.02275$$

例5. 10部机器独立工作，每部停机的概率为0.2，求3部机器同时停机的概率。

解：设同时停机的数目为X，它服从二项分布

$n = 10, p = 0.2, np = 2, \sqrt{npq} = 1.265$

(1) 直接计算

$$P(X = 3) = C_{10}^{3} 0.2^{3} 0.8^{7} \approx 0.2013$$

(2) 用极限定理

$$P(X = 3) = \frac{1}{\sqrt{npq}} \phi_0 \left(\frac{k - np}{\sqrt{npq}} \right)$$

$$= \frac{1}{1.265} \phi_0 \left(\frac{3 - 2}{1.265} \right) = \frac{1}{1.265} \phi_0(0.79) \approx 0.2308$$

相差较大，这是因为n较小，一般要求$n \geq 30$

例6. 某单位有200台电话分机，每台大约有5%时间使用外线。若各分机是否使用外线是相互独立的，问总机至少要装多少条外线才能使打外线的接通率达到90%？

解：用X表示需使用外线的分机数量，服从二项分布。

$$n = 200\text{，}p = 0.05\text{，}np = 10\text{，}npq = 9.5$$

假设要装k条外线，则

$$P(X \le k) = P\left(\frac{X-10}{\sqrt{9.5}} \le \frac{k-10}{\sqrt{9.5}}\right) = \Phi\left(\frac{k-10}{\sqrt{9.5}}\right) \ge 0.9$$

故

$$\frac{k-10}{\sqrt{9.5}} \ge 1.30$$

于是，$k \ge 14$，至少要装14条外线

例7. 设随机变量$X_1, X_2, \dots, X_n$相互独立，且X_i在区间(-1,1)上服从均匀分布$(i = 1,2,\dots,n)$，试证当n充分大时，随机变量$Z_n = \frac{1}{n}\sum_{i=1}^{n} X_i^2$近似服从正态分布，并指出其分布参数。

证：记$Y_i = X_i^2$，$(i = 1,2,\dots,n)$

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

$$D(Y_i) = E\left(Y_i^2\right) - [E(Y_i)]^2 = E\left(X_i^4\right) - [E(Y_i)]^2$$

因为：

$$E\left(X_i^4\right) = \int_{-1}^{1} x_i^4 \cdot \frac{1}{2} dx_i = \frac{1}{5}\text{，}D(Y_i) = \frac{1}{5} - \left(\frac{1}{3}\right)^2 = \frac{4}{45}$$

因为$X_1, X_2, \dots, X_n$相互独立，根据独立同分布的中心极限定理，

$$nZ_n = \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i$$

近似服从正态分布

$$N\left(\frac{n}{3}, \frac{4n}{45}\right),$$

故Z近似地服从正态分布

$$N\left(\frac{1}{3}, \frac{4}{45n}\right)$$

本章小结

两类极限定理

- 大数定律——回答了收敛性的问题，研究了随机变量序列依概率收敛的极限问题，证明了用频率估计概率的合理性
- 中心极限定理——阐明了在什么条件下，原本不属于正态分布的一些随机变量其总和分布渐近服从正态分布

大数定律的关系

大数定律的一般形式

对随机变量序列$\{X_n\}$ $(n=1,2,\dots)$，$E(X_n)$存在，

$$\frac{1}{n}\sum_{k=1}^{n}X_k \xrightarrow{P} \frac{1}{n}\sum_{k=1}^{n}E(X_k)$$

不要求独立同分布

Khinchin大数定律

$X_1,X_2,\dots,X_n,\dots$ 独立同分布，

$E(X_k)=\mu,\ (k=1,2,\dots)$

$$\frac{1}{n}\sum_{k=1}^{n}X_k \xrightarrow{P} \mu$$

Markov大数定律

随机变量序列$\{X_n\}$，$E(X_n)$存在，且

$$\lim_{n\to\infty} D\left(\frac{1}{n}\sum_{k=1}^{n}X_k\right)=0\ (\text{Markov条件})$$

$$\frac{1}{n}\sum_{k=1}^{n}X_k \xrightarrow{P} \frac{1}{n}\sum_{k=1}^{n}E(X_k)$$

Bernoulli试验

Bernoulli大数定律

n次独立重复试验发生f_A次，p是每次试验中发生的概率

$$\frac{f_A}{n} \xrightarrow{P} p$$

Bernoulli试验

Poisson大数定律

n 次独立试验发生 f_A 次，第k次发生概率为p_k

$$\frac{n_A}{n} \xrightarrow{P} \frac{1}{n}\sum_{k=1}^{n}p_k$$

不要求独立性，但对方差有约束

不要求方差的存在性

0-1分布

Chebyshev大数定律

$X_1,X_2,\dots,X_n,\dots$两两相互独立

$E(X_n)$、$D(X_k)$存在$(k=1,2,\dots)$

$$\frac{1}{n}\sum_{k=1}^{n}X_k \xrightarrow{P} \frac{1}{n}\sum_{k=1}^{n}E(X_k)$$

独立同分布

独立同分布下的Chebyshev大数定律

$E(X_k)=\mu, D(X_k)$存在

$(k=1,2,\dots)$

$$\frac{1}{n}\sum_{k=1}^{n}X_k \xrightarrow{P} \mu$$

三个中心极限定理

独立同分布的中心极限定理

$\{X_n\}$独立同分布，且$E(X_k)=\mu$, $D(X_k)=\sigma^2>0, (k=1,2,\dots)$，

$$\lim_{n\to\infty} Y_n=\frac{\sum_{k=1}^{n}X_k-n\mu}{\sqrt{n}\sigma}\sim N(0,1)$$

不要求同分布，但要求方差满足收敛性

Lyapunov定理

$\{X_n\}$独立，且$E(X_k)=\mu_k$, $D(X_k)=\sigma_k^2>0$, $B_n^2=\sum_{k=1}^{n}\sigma_k^2$，

若存在$\delta>0$，使得当$n\to\infty$时

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

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

二项分布逼近

De Moivre – Laplace定理

$\eta_n\sim b(n,p)\ (0<p<1)$，则对于任意$x$，有

$$\lim_{n\to\infty}\frac{\eta_n-np}{\sqrt{np(1-p)}}\sim N(0,1)$$

作业

概率论与数理统计

pp. 125-126, #2，#4，#6，#12