第14讲　大数定律与中心极限定理

一、背景

二、大数定律

1.切比雪夫不等式

2.切比雪夫大数定律

3.贝努力大数定律

4.辛钦大数定律

三、中心极限定理

1.独立同分布中心极限定理

2. Lyapunov中心极限定理

3. De-Moivere-Laplace中心极限定理

第14讲　大数定律与中心极限定理

一、背景

1.为何能以某事件发生的频率作为该事件的概率的估计？

2.为何能以样本均值作为总体期望的估计？

3.为何正态分布在概率论中占有极其重要的地位？

4.大样本统计推断的理论基础是什么？

大数定律　　中心极限定理

第14讲 大数定律与中心极限定理

二、大数定律

1.切比雪夫不等式

设随机变量X的数学期望$E(X)=\mu$,方差$D(X)=\sigma^2$,则对任意的正数ε，不等式

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

或

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

成立.

第14讲 大数定律与中心极限定理

二、大数定律

2.切比雪夫大数定理

若X_1，X_2，…，X_n，…，为独立同分布随机变量序列，$E(X_k)=\mu$ $D(X_k)=\sigma^2$ $(k=1，2，…)$，则对任意的正数$\varepsilon>0$,有

$$\lim_{n\to+\infty} P\{|\frac{1}{n}\sum_{k=1}^{n} X_k - \mu| < \varepsilon\} = 1$$

或

$$\lim_{n\to+\infty} P\{|\frac{1}{n}\sum_{k=1}^{n} X_k - \mu| \geq \varepsilon\} = 0$$


切比雪夫

证 根据已知条件

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

$$D\left(\frac{1}{n}\sum_{k=1}^{n}X_k\right)=\frac{1}{n^2}D\left(\sum_{k=1}^{n}X_k\right)=\frac{1}{n^2}\sum_{k=1}^{n}D(X_k)=\frac{\sigma^2}{n}$$

由切比雪夫不等式，有

$$1-\frac{\sigma^2}{n\varepsilon^2}\le P\{|\frac{1}{n}\sum_{k=1}^{n}X_k-\mu|<\varepsilon)\le 1$$

又

$$\lim_{n\to\infty}\left(1-\frac{\sigma^2}{n\varepsilon^2}\right)=1$$

所以

$$\lim_{n\to+\infty}P\{|\frac{1}{n}\sum_{k=1}^{n}X_k-\mu|<\varepsilon\}=1$$

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

第14讲　大数定律与中心极限定理

二、大数定律

3.伯努例大数定理

设n_A为是n次独立重复试验中事件A发生的次数，p是事件A在每次试验中发生的概率,则对任意的正数ε>0,有

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

或

$$\lim_{n\to+\infty} P\{|\frac{n_A}{n}-p|\geq\varepsilon\}=0$$

伯努利

证 设

$$X_k=\begin{cases}0, & A\text{在第}k\text{次试验中发生},\\ 1, & A\text{在第}k\text{次试验中未发生}.\end{cases}$$

那么 $X_1,X_2,\cdots,X_n$ 相互独立，且服从参数为p的0—1分布，$E(X_k)=p$，$D(X_k)=p(1-p)$.

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

由切比雪夫大数定理,有

$$\lim_{n\to+\infty}P\{|\frac{1}{n}\sum_{k=1}^{n}X_k-\mu|<\varepsilon\}=1$$

即

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

第14讲　大数定律与中心极限定理

二、大数定律

4.辛钦大数定理

若X_1，X_2，…，X_n，…，为独立同分布随机变量序列，$E(X_k)=\mu$（k=1，2，…），则对任意的正数ε>0,有

$$\lim_{n\to+\infty} P\{|\frac{1}{n}\sum_{k=1}^{n} X_k - \mu| < \varepsilon\} = 1$$

或

$$\lim_{n\to+\infty} P\{|\frac{1}{n}\sum_{k=1}^{n} X_k - \mu| \geq \varepsilon\} = 0$$

辛钦

第14讲 大数定律与中心极限定理

二、大数定律

定义1 设Y_1，Y_2，…，Y_n，…，是一随机变量序列，a为一常数．若对任意给定正数ε>0，有

$$\lim_{n\to+\infty} P\{|Y_n-\mu|<\varepsilon\}=1$$

则称随机变量序列$Y_1,Y_2,\dots,Y_n,\dots$,依概率收敛于a.

定义2 设X_1，X_2，…，X_n，… 是一随机变量序列$Y_n=\dfrac{1}{n}\sum_{k=1}^{n}X_k$.若存在常数列$\{a_n\}$使对任意给定的正数ε，恒有$\lim\limits_{n\to+\infty} P\{|Y_n-a_n|<\varepsilon\}=1$，则称随机变量序列$\{Y_n\}$服从大数定律.

第14讲　大数定律与中心极限定理

三、中心极限定理

1.独立同分布中心极限定理

若X_1，X_2，…，X_n，…，为独立同分布随机变量序列，$E(X_k)=\mu$ $D(X_k)=\sigma^2$ (k=1，2,…)，则随机变量标准化量

$$Y_n=\frac{\sum_{k=1}^{n}X_k-n\mu}{\sqrt{n}\sigma}$$

的分布函数$F_n(x)$对于任意x满足

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

$$\frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\sigma} \underset{近似}{\sim} N(0,1) \Rightarrow \sum_{k=1}^{n} X_k \underset{近似}{\sim} N(n\mu, n\sigma^2)$$

$$\Downarrow$$

$$\frac{\frac{1}{n}\sum_{k=1}^{n} X_k - \mu}{\sigma/\sqrt{n}} \underset{近似}{\sim} N(0,1) \Rightarrow \bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \underset{近似}{\sim} N(\mu, \frac{\sigma^2}{n})$$

$$\lim_{n\to+\infty} F_n(x) = \lim_{n\to+\infty} P\{\frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\sigma} \le x\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

例1　一个加法器同时收到20个噪声电压V_k,(k=1,2,…,20),设他们是相互独立的随机变量,且都在区间(0,10)内服从均匀分布．记$V=\sum_{k=1}^{20}V_k$，求$P\{V>105\}$.

解　由题意　$E(V_k)=5,\ D(V_k)=\dfrac{100}{12}\quad(k=1,2,\cdots,20)$

随机变量

$$Z=\frac{\sum_{k=1}^{n}V_k-n\mu}{\sqrt{n}\sigma}=\frac{\sum_{k=1}^{20}V_k-20\times5}{\sqrt{20}\sqrt{100/12}}\underset{\text{近似}}{\sim}N(0,1)$$

于是

$$P(V>105)=P\{\frac{V-100}{\sqrt{20}\sqrt{100/12}}>\frac{105-100}{\sqrt{20}\sqrt{100/12}}\}$$

$$\approx P\{\frac{V-100}{\sqrt{20}\sqrt{100/12}} > 0.387\}$$

$$\approx 1-P\{\frac{V-100}{\sqrt{20}\sqrt{100/12}} \le 0.387\}$$

$$= 1-\Phi(0.387) \approx 0.384$$

所以

$$P(V>105) \approx 0.384$$

$$P(V>105) = P\{\frac{V-100}{\sqrt{20}\sqrt{100/12}} > \frac{105-100}{\sqrt{20}\sqrt{100/12}}\}$$

第14讲 大数定律与中心极限定理

三、中心极限定理

2.李雅普诺夫中心极限定理

若X_1，X_2，…，X_n，…，为

列，$E(X_k)=\mu_k$，$D(X_k)=\sigma_k^2>0$，

δ，使当$n\to\infty$时，

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

$$Z_n=\frac{\sum_{k=1}^{n}X_k-\sum_{k=1}^{n}\mu_i}{\sqrt{\sum_{k=1}^{n}\sigma_k^2}}$$

则随机变量标准化量Z_n的分布函数$F_n(x)$对于任意x满足

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

$$Z_n = \frac{\sum_{k=1}^{n} X_k - \sum_{k=1}^{n} \mu_i}{\sqrt{\sum_{k=1}^{n} \sigma_k^2}} \underset{\text{近似}}{\sim} N(0,1)$$

$$\Downarrow$$

$$\sum_{k=1}^{n} X_k \underset{\text{近似}}{\sim} N\left(\sum_{k=1}^{n} \mu_i, \sum_{k=1}^{n} \sigma_k^2\right)$$

李雅普诺夫

说明：无论各随机变量$X_k(k=1,2,\ldots)$服从什么分布，只要满足定理的条件，那么他们的和当n很大时，就近似服从正态分布，这就是为什么正态随机变量在概率论中占有非常重要地位的一个基本原因.

第14讲　大数定律与中心极限定理

三、中心极限定理

3. 棣莫弗——拉普拉斯中心极限定理

设随机变量服 $\eta_n (n=1,2,\cdots)$ 从参数为n, p的二项分布，则对任意x，有

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

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

证 设$X_k(k=1,2,\ldots)$服从参数为p的0-1分布，那么

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

$$E(X_k) = p, \quad D(X_k) = p(1-p)$$

于是由独立同分布中心极限定理，得

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

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

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

棣莫弗

例2　某单位内部有260部电话分机，每部分机有4%的时间使用外线与外界通话，可以认为每部电话分机使用不同的外线是相互独立的，问总机需备多少条外线才能95%满足每部分机在使用外线时不用等候?

分析　我们将每使用一部电话分机看作是一次试验，那么260部电话分机同时使用外线通话的分机数记X,则X是一个随机变量，且有X服从参数为260，0.04的二项分布.总机需备m条外线才能95%满足每部分机在使用外线时不用等候,即$P\{X\leq m\}\geq 0.95$.

例2　某单位内部有260部电话分机，每部分机有4%的时间使用外线与外界通话，可以认为每部电话分机使用不同的外线是相互独立的，问总机需备多少条外线才能95%满足每部分机在使用外线时不用等候?

解　设260部电话分机同时使用外线通话的分机数为X,则X服从参数为260，0.04的二项分布．总机需备m条外线才能95%满足每部分机在使用外线时不用等候．由已知条件及拉普拉斯中心极限定理，有

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

$$= P\left\{\frac{X-260\times 0.04}{\sqrt{260\times 0.04\times 0.96}} \le \frac{m-260\times 0.04}{\sqrt{260\times 0.04\times 0.96}}\right\}$$

$$P\{X\le m\}=P\{\frac{X-np}{\sqrt{np(1-p)}}\le\frac{m-np}{\sqrt{np(1-p)}}\}$$

$$=P\{\frac{X-260\times0.04}{\sqrt{260\times0.04\times0.96}}\le\frac{m-260\times0.04}{\sqrt{260\times0.04\times0.96}}\}$$

$$=\Phi(\frac{m-260\times0.04}{\sqrt{260\times0.04\times0.96}})\ge 0.95$$

查表 $\Phi(1.64)=0.9495,\ \Phi(1.65)=0.9505$.不妨取

$$\frac{m-260\times0.04}{\sqrt{260\times0.04\times0.96}}\ge 1.65$$

$$m\ge 260\times0.04+1.65\sqrt{260\times0.04\times0.96}\approx 15.61$$

这就是说，总机至少需备16条外线才能95%满足每部分机在使用外线时不用等候.

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，且服从统一分布.(1) 求参加会议的家长人数X超过450的概率；(2) 求有1名家长来参加会议的学生人数不多于340的概率.

解　（1）以X_k记第k个学生来参加会议的家长人数，则由已知条件X_k的分布率为

X_k	0	1	2
P	0.05	0.8	0.15

可以计算$E(X_k)=1.1$，$D(X_k)=0.19$，$k=1$，2，…，400.由独立同分布中心极限定理，得

$$P(X>450)=P(\sum_{k=1}^{400}X_k>450)$$

$$=P(\frac{\sum_{k=1}^{400}X_k-400\times 1.1}{\sqrt{400}\sqrt{0.19}}>\frac{450-400\times 1.1}{\sqrt{400}\sqrt{0.19}})$$

$$=1-P(\frac{450-400\times 1.1}{\sqrt{400}\sqrt{0.19}})\approx 1-P(1.147)\approx 0.1251$$

解　(2) 以Y记由一名家长参加会议的学生人数，则Y服从参数为400，0.8的二项分布．于是由棣莫弗——拉普拉斯中心极限定理，得

$$P(Y \le 340) = P\left(\frac{Y - np}{\sqrt{np(1-p)}} \le \frac{340 - np}{\sqrt{np(1-p)}}\right)$$

$$\approx \Phi\left(\frac{340 - np}{\sqrt{np(1-p)}}\right)$$

$$= \Phi\left(\frac{340 - 400 \times 0.8}{\sqrt{400 \times 0.8 \times 0.2}}\right)$$

$$= \Phi(2.5) \approx 0.9938$$

从而有1名家长来参加会议的学生人数不多于340的概率约为0.9938.

查表 $\Phi(3.01)=0.999$ ，所以

$$\frac{x/15-200\times 0.7}{\sqrt{200\times 0.7\times 0.3}}\geq 3.01 \quad \Rightarrow x\geq 2392.6$$

即供电所至少要供给这个车间2392.6千瓦的电力.

解　供电所至少要供给这个车间x千瓦的电力，才能以99.9%的概率保证这个车间不会因供电不足而影响生产.以X记200台车床在同一时间段内开动的台数，则由已知条件X服从参数为200，0.7的二项分布，于是由棣莫弗——拉普拉斯中心极限定理有

$$P(15X\leq x)=P\left(\frac{X-200\times 0.7}{\sqrt{200\times 0.7\times 0.3}}\leq \frac{x/15-200\times 0.7}{\sqrt{200\times 0.7\times 0.3}}\right)$$

$$\approx \Phi\left(\frac{x/15-200\times 0.7}{\sqrt{200\times 0.7\times 0.3}}\right)\geq 0.999$$

第12周　问 题

电视台需作节目A 收视率的调查.每天在播电视的同时，随机地向当地居民打电话询问是否在看电视．若在看电视，再问是否在看节目A．设回答看电视的居民户数为n．若要保证以95%的概率使调查误差在10%之内,n应取多大？

每晚节目A 播出一小时，调查需同时进行，设每小时每人能调查20户,每户居民每晚看电视的概率为70%，电视台需安排多少人作调查,又若使调查误差在1之内，n取多大？

第12周　问 题

一本书有1000000个印刷符号，排版时每个符号被排错的概率为千分之一．校对时，每个排版错误被改正的概率为0.99．求在校对后错误不多于15个的概率．

中心极限定理的意义

在第二章曾讲过有许多随机现象服从正态分布 是由于许多彼次没有什么相依关系、对随机现象谁也不能起突出影响，而均匀地起到微小作用的随机因素共同作用(即这些因素的叠加)的结果.

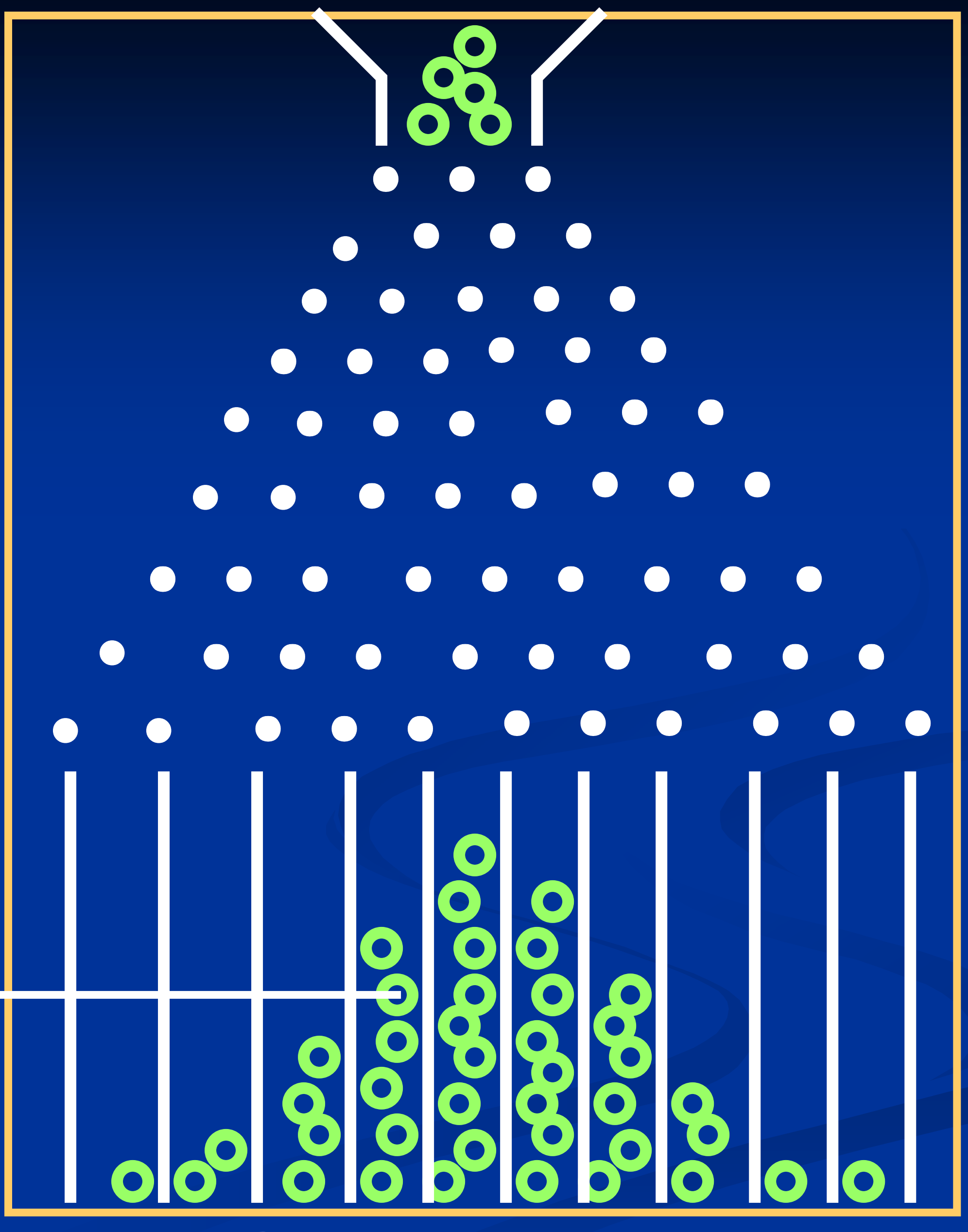
若联系于此随机现象的随机变量为X，则它可被看成为许多相互独立的起微小作用的因素X_k的总和 $\sum_k X_k$，而这个总和服从或近似服从正态分布.

对此现象还可举个有趣的例子——

高尔顿钉板试验—— 加以说明.

$N(0,\sqrt{n})$

n — 钉子层数



中心极限定理的应用

例1 炮火轰击敌方防御工事 100 次, 每轰击命中的炮弹数服从同一分布, 其数学期望为 2 , 均方差为1.5. 若各次轰击命中的炮弹数是相互独立的, 求100 次轰击

(1) 至少命中180发炮弹的概率;
(2) 命中的炮弹数不到200发的概率.

解 设 X_k 表示第 k 次轰击命中的炮弹数

$$E(X_k)=2,\quad D(X_k)=1.5^2,\quad k=1,2,\cdots,100$$

$X_1, X_2, \cdots, X_{100}$ 相互独立,

设 X 表示100次轰击命中的炮弹数, 则

$$X=\sum_{k=1}^{100} X_k,\ E(X)=200,\ D(X)=225,$$

由独立同分布中心极限定理, 有

$$X \overset{\text{近似}}{\sim} N(200, 225)$$

$$(1)\quad P(X \geq 180) \approx 1 - \Phi\left(\frac{180-200}{15}\right)$$

$$= 1 - \Phi(-1.3) = \Phi(1.3) = 0.91$$

$$(2)\quad P(0 \leq X < 200) \approx \Phi\left(\frac{200-200}{15}\right) - \Phi\left(\frac{0-200}{15}\right)$$

$$= \Phi(0) - \Phi(-13.33) = 0.5$$

例2 售报员在报摊上卖报，已知每个过路人在报摊上买报的概率为1/3. 令X是出售了100份报时过路人的数目，求

$P(280 \le X \le 320)$.

解 令X_i为售出了第 $i-1$ 份报纸后到售出第i 份报纸时的过路人数，$i = 1,2,\ldots,100$

$$P(X_i = k) = p(1-p)^{k-1}\Big|_{p=1/3}, \quad k = 1,2,\cdots$$

(几何分布)

$$E(X_i) = \frac{1}{p}\Big|_{p=1/3} = 3, \quad D(X_i) = \frac{1-p}{p^2}\Big|_{p=1/3} = 6$$

$X_1, X_2, \cdots, X_{100}$ 相互独立, $X = \sum_{k=1}^{100} X_k$

$$E(X) = 300, \quad D(X) = 600$$

由独立同分布中心极限定理, 有

$$X \sim N(300, 600) \text{（近似）}$$

$$P(280 \le X \le 320) \approx \Phi\left(\frac{320-300}{\sqrt{600}}\right) - \Phi\left(\frac{280-300}{\sqrt{600}}\right)$$

$$\approx 2\Phi\left(\frac{20}{\sqrt{600}}\right) - 1 \approx 2\Phi(0.8165) - 1 \approx 0.5878$$

例3 检验员逐个检查某产品,每查一个需用10秒钟. 但有的产品需重复检查一次,再用去10秒钟. 若产品需重复检查的概率为 0.5, 求检验员在 8 小时内检查的产品多于1900个的概率.

解 若在 8 小时内检查的产品多于1900个,即检查1900个产品所用的时间小于 8 小时.

设 X 为检查1900 个产品所用的时间(秒)

设 X_k 为检查第 k 个产品所用的时间(单位：秒), $k = 1,2,\ldots,1900$

X_k	10	20
P	0.5	0.5

$$E(X_k)=15,\quad D(X_k)=25$$

$X_1, X_2, \cdots, X_{1900}$ 相互独立同分布, $X=\sum_{k=1}^{1900} X_k$

$$E(X)=1900\times 15=28500$$

$$D(X)=1900\times 25=47500$$

$$X \overset{\text{近似}}{\sim} N(28500,47500)$$

$$P(10\times 1900 \le X \le 3600\times 8)$$

$$= p(19000 \le X \le 28800)$$

$$\approx \Phi\left(\frac{28800-28500}{\sqrt{47500}}\right)-\Phi\left(\frac{19000-28500}{\sqrt{47500}}\right)$$

$$\approx \varPhi(1.376)-\varPhi(-43.589)$$

$$\approx 0.9162$$

例5 设有一批种子，其中良种占1/6. 试估计在任选的6000粒种子中，良种比例与 1/6 比较上下不超过1%的概率.

解 设 X 表示6000粒种子中的良种数，

则 $X \sim B(6000, 1/6)$

由德莫佛—拉普拉斯中心极限定理，

有 $X \overset{\text{近似}}{\sim} N\left(1000, \frac{5000}{6}\right)$

$$P\left(\left|\frac{X}{6000}-\frac{1}{6}\right|<0.01\right)=P\left(\left|X-1000\right|<60\right)$$

$$\approx \Phi\left(\frac{1060-1000}{\sqrt{5000/6}}\right)-\Phi\left(\frac{940-1000}{\sqrt{5000/6}}\right)$$

$$=\Phi\left(\frac{60}{\sqrt{5000/6}}\right)-\Phi\left(\frac{-60}{\sqrt{5000/6}}\right)$$

$$=2\Phi\left(\frac{60}{\sqrt{5000/6}}\right)-1\approx 0.9624$$

比较几个近似计算的结果

二项分布(精确结果) $P\left(\left|\frac{X}{6000}-\frac{1}{6}\right|<0.01\right)\approx 0.9590$

中心极限定理 $P\left(\left|\frac{X}{6000}-\frac{1}{6}\right|<0.01\right)\approx 0.9624$

Poisson 分布 $P\left(\left|\frac{X}{6000}-\frac{1}{6}\right|<0.01\right)\approx 0.9379$

Chebyshev 不等式 $P\left(\left|\frac{X}{6000}-\frac{1}{6}\right|<0.01\right)\geq 0.7685$

*补充作业

设某农贸市场某种商品每日的价格的变化是个相互独立且均值为0，方差为$\sigma^2=2$的随机变量 Y_n，并满足

$$X_n = X_{n-1} + Y_n \quad (n \geq 1)$$

其中X_n是第n天该商品的价格.如果今天的价格为100，求18天后该商品的价格在 96 与 104 之间的概率.

解 设 X_0 表示今天该商品的价格, X_{18}为18天后该商品的价格, 则

$$X_{18}=X_{17}+Y_{18}=X_{16}+Y_{17}+Y_{18}=X_0+\sum_{i=1}^{18}Y_i$$

得 $P(96\le X_{18}\le 104)=P(-4\le \sum_{i=1}^{18}Y_i\le 4)$

$$=P(-\frac{4}{\sqrt{36}}\le \frac{1}{\sqrt{36}}\sum_{i=1}^{18}Y_i\le \frac{4}{\sqrt{36}})$$

$$\approx \Phi(2/3)-\Phi(-2/3)=2\Phi(2/3)-1$$

$$=2\times 0.747-1=0.494.$$