第 5 章

大数定律与中心极限定理

大数定律和中心极限定理是概率论和统计学的基础理论，在理论研究和应用实践中都有着重要应用。大数定律描述了独立同分布随机变量序列的算术平均值依概率收敛到分布的数学期望；中心极限定理描述了独立同分布随机变量序列之和的分布逼近于正态分布。在很多场合中都能见到被冠以“大数定律”和“中心极限定理”的各类结论，实际上这两大类定理有很多版本，如果读者对此有兴趣，则可以阅读专门的概率论著作。本书介绍其中常用的一些，大数定律包括切比雪夫(Chebyshev)大数定律、伯努利(Bernoulli)大数定律和辛钦(Khinchine)大数定律；中心极限定理包括棣莫弗-拉普拉斯(De Moivre-Laplace)中心极限定理，列维-林德伯格(Levy-Lindberg)中心极限定理。

本章重点内容：

(1) 切比雪夫不等式和切比雪夫大数定律。

(2) 伯努利大数定律。

(3) 辛钦大数定律。

(4) 棣莫弗-拉普拉斯中心极限定理。

(5) 列维-林德伯格中心极限定理。

5.1 大数定律

前文提到过，在大量随机试验中，某事件出现的频率 $f_n(A)$ 具有稳定性，当重复试验的次数 n 趋于无穷大时，频率趋向于一个特定的常数。这也是概率论的客观基础，本节将对此作一些理论解释。

5.1.1 切比雪夫不等式

切比雪夫不等式。设随机变量 X 的数学期望 $E(X)$ 和方差 $D(X)$ 都存在，则对任意的 $\varepsilon>0$，总有

$$P(|X-E(X)|\geqslant\varepsilon)\leqslant\frac{D(X)}{\varepsilon^2} \tag{5-1}$$

证明：只证明连续随机变量的情形。设 X 的概率密度为 $f(x)$，则有

$$P(|X-E(X)|\geqslant\varepsilon)=\int_{|X-E(X)|\geqslant\varepsilon}f(x)\mathrm{d}x=\int_{|X-E(X)|\geqslant\varepsilon}\frac{|X-E(X)|^2}{\varepsilon^2}f(x)\mathrm{d}x$$

$$\leqslant\frac{1}{\varepsilon^2}\int_{-\infty}^{\infty}(x-E(X))^2f(x)\mathrm{d}x=\frac{D(X)}{\varepsilon^2}\tag{5-2}$$

证毕。

切比雪夫不等式有时也可写成

$$P(|X-E(X)|<\varepsilon)\geqslant1-\frac{D(X)}{\varepsilon^2}\tag{5-3}$$

切比雪夫不等式给出了当随机变量的分布未知，只知道数学期望 $E(X)$ 和方差 $D(X)$ 时，估计概率 $P(|X-E(X)|\geqslant\varepsilon)$ 的上界，这个估计是比较粗糙的。例如取 $\varepsilon=2\sqrt{D(X)}$ 和 $3\sqrt{D(X)}$，可以得到

$$P(|X-E(X)|<2\sqrt{D(X)})\geqslant1-\frac{D(X)}{4D(X)}=0.75$$

$$P(|X-E(X)|<2\sqrt{D(X)})\geqslant1-\frac{D(X)}{9D(X)}=\frac{8}{9}$$

显然，如果已知随机变量的分布，则所求概率 $P(|X-E(X)|\geqslant\varepsilon)$ 可以明确地计算出来，也就没必要用切比雪夫不等式估计了，切比雪夫不等式只适用于分布未知的情形。

【例 5-1】 设随机变量 X 的概率密度为

$$f(x)=\begin{cases}2\mathrm{e}^{-2x}, & x>0\\ 0, & x\leqslant0\end{cases}$$

(1) 根据切比雪夫不等式估计 $P(X\geqslant3/2)\leqslant A$，求 A 的值。

(2) 直接计算 $P(X\geqslant3/2)$ 的值。

解：(1) 随机变量 X 实际上服从参数为 2 的指数分布，因此 $E(X)=1/2$，$D(X)=1/4$。根据切比雪夫不等式

$$P(X\geqslant3/2)=P(X-1/2\geqslant1)=P(X-1/2\geqslant1)+P(X-1/2\leqslant-1)$$

$$=P(|X-1/2|\geqslant1)=P(|X-E(X)|\geqslant1)$$

$$\leqslant\frac{D(X)}{1}=\frac{1}{4}$$

因此 $A=1/4$。

(2) 根据指数分布的性质

$$P(X>t)=\mathrm{e}^{-2t},\quad t>0$$

所以 $P(X\geqslant3/2)=\mathrm{e}^{-3}$。

代码如下：

```
#第5章/5-1.py
from sympy import symbols, exp, oo, Rational, integrate
x = symbols('x')
f = lambda x : 2 * exp(-2 * x)
p = integrate(f(x), (x,Rational(3, 2), oo))
print('所求的概率为', p)
```

输出如下：

```
所求的概率为 exp(-3)
```

【例 5-2】 设随机变量 X 的密度为 $f(x)$，$D(X)=1$，随机变量 Y 的密度为 $f(-y)$，并且 X 与 Y 的相关系数为 $-1/4$，用切比雪夫不等式估计 $P(|X+Y|\geqslant 2)$ 的上界。

解：随机变量 Y 的期望

$$E(Y)=\int_{-\infty}^{+\infty} yf(-y)\mathrm{d}y=\int_{+\infty}^{-\infty} -tf(t)\mathrm{d}(-t)=-\int_{-\infty}^{+\infty} tf(t)\mathrm{d}t=-E(X)$$

即

$$E(X+Y)=0$$

随机变量 Y 的方差

$$D(Y)=E(Y^2)-E(Y)^2=\int_{-\infty}^{+\infty} y^2 f(-y)\mathrm{d}y-(-E(X))^2$$

$$=\int_{-\infty}^{\infty} y^2 f(y)\mathrm{d}y-E(X)^2=D(X)$$

根据切比雪夫不等式

$$P(|X+Y|\geqslant 2)=P(|X+Y-E(X+Y)|\geqslant 2)$$

$$\leqslant \frac{D(X+Y)}{2^2}=\frac{D(X+Y)}{4}$$

对于 $D(X+Y)$ 有

$$D(X+Y)=D(X)+D(Y)+2\mathrm{cov}(X,Y)$$

$$=D(X)+D(Y)+2\rho_{XY}\sqrt{D(X)}\sqrt{D(Y)}$$

$$=1+1-\frac{1}{2}=\frac{3}{2}$$

综上有

$$P(|X+Y|\geqslant 2)\leqslant \frac{D(X+Y)}{4}=\frac{3}{8}$$

5.1.2 依概率收敛

下面给出依概率收敛的定义。设 $X_1,X_2,\cdots,X_n$ 是一个随机变量序列，A 是一个常数，如果对任意的 $\varepsilon>0$，有

$$\lim_{n\to\infty} P(|X_n-A|<\varepsilon)=1 \tag{5-4}$$

则称随机变量序列 $X_1,X_2,\cdots,X_n$ 依概率收敛于常数 A，也记作 $X_n\xrightarrow{P}A$。

依概率收敛的序列有以下性质。设 $X_n\xrightarrow{P}a$，$Y_n\xrightarrow{P}b$，并且函数 $g(x,y)$ 在点 (a,b) 连续，则

$$g(X_n,Y_n)\xrightarrow{P}g(a,b) \tag{5-5}$$

证明从略。

5.1.3 切比雪夫大数定律

下面不加证明地给出切比雪夫大数定律。切比雪夫大数定律。设 $X_1,X_2,\cdots,X_n$ 是

一个两两不相关的随机变量序列，存在常数 C 使 $D(X_i)\leqslant C(i=1,2,3,\cdots)$，则对任意的 $\varepsilon>0$ 有

$$\lim_{n\to\infty}P\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}E(X_i)\right|<\varepsilon\right)=1 \tag{5-6}$$

证明从略。

【例 5-3】 设 $X_1,X_2,\cdots,X_n$ 是相互独立的随机变量序列，X_n 服从参数为 n 的指数分布，$n\geqslant 1$，则下列随机变量序列中不服从切比雪夫大数定律的是(　　)。

A. $X_1,\frac{1}{2}X_2,\cdots,\frac{1}{n}X_n$　　B. $X_1,X_2,\cdots,X_n$

C. $X_1,2X_2,\cdots,nX_n$　　D. $X_1,2^2X_2,\cdots,n^2X_n$

解：根据切比雪夫大数定律的条件，要求方差存在且一致有界，即 $D(X_n)\leqslant C$，其中 C 是常数。因为 X_n 服从参数为 n 的指数分布，故 $D(X_n)=1/n^2$。检查 4 个选项，D 项不满足方差一致有界，因此，本题选 D。

5.1.4　辛钦大数定律

辛钦大数定律也称弱大数定律。设 $X_1,X_2,\cdots,X_n$ 是独立同分布的随机变量序列，具有数学期望 $E(X_i)=\mu$，则对任意的 $\varepsilon>0$ 有

$$\lim_{n\to\infty}P\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right|<\varepsilon\right)=1 \tag{5-7}$$

证明：只证明方差有限的情形。设 $D(X_i)=\sigma^2$，由于

$$E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\frac{1}{n}n\mu=\mu \tag{5-8}$$

且由独立性可得

$$D\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}D(X_i)=\frac{1}{n^2}n\sigma^2=\frac{\sigma^2}{n} \tag{5-9}$$

则由切比雪夫不等式得

$$1\geqslant P\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right|<\varepsilon\right)\geqslant 1-\frac{\frac{\sigma^2}{n}}{\varepsilon^2} \tag{5-10}$$

根据夹逼定理，令 $n\to\infty$ 有

$$\lim_{n\to\infty}P\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right|<\varepsilon\right)=1 \tag{5-11}$$

证毕。

通俗地讲，辛钦大数定律保证了独立同分布的随机变量序列，当 n 很大时它们的算术平均值很可能接近于数学期望。

5.1.5　伯努利大数定律

伯努利大数定律是辛钦大数定律的一个重要推论。设随机变量 $X_n\sim B(n,p)$，则对任意的 $\varepsilon>0$ 有

$$\lim_{n\to\infty}P\left(\left|\frac{X_n}{n}-p\right|<\varepsilon\right)=1 \tag{5-12}$$

证明：二项分布 $X_n \sim B(n,p)$ 可以写成 n 个独立的符合 $B(1,p)$ 分布的随机变量之和，即

$$X_n = Y_1 + Y_2 + \cdots + Y_n \tag{5-13}$$

其中 $Y_1, Y_2, \cdots, Y_n$ 独立同分布，$Y_i \sim B(1,p)$。又因为 $E(Y_i)=p$，则由辛钦大数定律可得

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i - p\right| < \varepsilon\right) = \lim_{n\to\infty} P\left(\left|\frac{X_n}{n} - p\right| < \varepsilon\right) = 1 \tag{5-14}$$

证毕。

伯努利大数定律表明，只要试验次数足够多，事件 $\{|X_n/n - p| < \varepsilon\}$ 是一个小概率事件，而小概率事件在实际中是几乎不发生的，这就是频率稳定性的真正含义。在实际应用中，当试验次数很大时，就可以用事件的频率来代替事件的概率。

【例 5-4】 设随机变量 $X_1, X_2, \cdots, X_n$ 相互独立，均服从分布函数

$$F(x;\theta) = \begin{cases} 1 - \exp\left\{-\dfrac{x^2}{\theta}\right\}, & x \geqslant 0 \\ 0, & x < 0 \end{cases}$$

(1) 是否存在实数 a，使对任意的 $\varepsilon > 0$ 都有

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i^2 - a\right| \geqslant \varepsilon\right) = 0$$

(2) 求 a 的取值。

解：(1) 记

$$f(x;\theta) = F'(x;\theta) = \begin{cases} \dfrac{2x}{\theta}\exp\left(-\dfrac{x^2}{\theta}\right), & x \geqslant 0 \\ 0, & x < 0 \end{cases}$$

求得

$$E(X_i^2) = \int_{-\infty}^{\infty} x^2 f(x;\theta)\,\mathrm{d}x = \int_0^{\infty} x^2 \frac{2x}{\theta}\exp\left(-\frac{x^2}{\theta}\right)\mathrm{d}x = \theta\int_0^{\infty} t\exp(-t)\,\mathrm{d}t = \theta$$

因为 $X_1, X_2, \cdots, X_n$ 独立同分布，所以 $X_1^2, X_2^2, \cdots, X_n^2$ 也相互独立，并且同分布。数学期望 $E(X_i^2)$ 存在，根据辛钦大数定律，对任意的 $\varepsilon > 0$ 有

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \theta\right| \geqslant \varepsilon\right) = 0$$

(2) a 的取值为 θ。

代码如下：

```
#第5章/5-2.py
from sympy import *
theta, x = symbols('theta, x')
Fx = 1 - exp(- x ** 2 / theta)
fx = Fx.diff(x)
#X^2的期望
ex2 = integrate(x ** 2 * fx, (x, 0, oo) )
print('X^2的期望为', ex2)
```

输出如下：

```
X^2 的期望为 Piecewise((theta, Abs(arg(theta)) < pi/2), (Integral(2 * x ** 3 * exp( - x ** 2/
theta)/theta, (x, 0, oo)), True))
```

5.2　中心极限定理

在客观实际中有很多随机变量，它们是由大量的相互独立的随机因素综合作用而成，其中每个因素在总的影响中所起的作用都是微小的，这种随机变量往往近似地服从正态分布。此现象就是中心极限定理的客观背景。

列维-林德伯格中心极限定理。设随机变量 $X_1, X_2, \cdots, X_n$ 独立同分布，存在数学期望和方差，$E(X_i)=\mu$，$D(X_i)=\sigma^2$，则对于任意实数，有

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \leqslant x\right) = \Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t \tag{5-15}$$

证明从略。

中心极限定理表明，当 n 充分大时，$\sum_{i=1}^{n} X_i$ 的标准化

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \tag{5-16}$$

近似服从标准正态分布 $N(0,1)$，或者说 $\sum_{i=1}^{n} X_i$ 近似服从 $N(n\mu, n\sigma^2)$，即

$$P\left(a < \sum_{i=1}^{n} X_i < b\right) \approx \Phi\left(\frac{b-n\mu}{\sqrt{n}\sigma}\right) - \Phi\left(\frac{a-n\mu}{\sqrt{n}\sigma}\right) \tag{5-17}$$

中心极限定理是数理统计中大样本统计推断的基础。注意，定理中独立同分布、数学期望存在、方差存在三者缺一不可。只要问题涉及独立同分布随机变量的和 $\sum_{i=1}^{n} X_i$，就可以考虑使用中心极限定理。

【例 5-5】 生产线生产的产品成箱包装，每箱质量是随机的。假如每箱平均重 50 千克，标准差为 5 千克，如果用载质量为 5 吨的汽车承运，试用中心极限定理说明每辆汽车最多可以装多少箱，才能保证不超载的概率大于 0.977（$\Phi(2)=0.977$）。

解：假设一辆车放了 n 箱产品。设 X_i 为第 i 箱产品的质量，根据题意可知，X_i 独立同分布，并且 $E(X_i)=50$，$\sqrt{D(X_i)}=5$。n 箱产品的总质量为 $T_n=\sum_{i=1}^{n} X_i$，$E(T_n)=50n$，$\sqrt{D(T_n)}=5\sqrt{n}$。由列维-林德伯格中心极限定理，T_n 近似地服从 $N(50n, 25n)$，故有

$$P(T_n \leqslant 5000) = P\left(\frac{T_n - 50n}{5\sqrt{n}} \leqslant \frac{5000-50n}{5\sqrt{n}}\right) \approx \Phi\left(\frac{1000-10n}{\sqrt{n}}\right) > 0.977$$

即

$$\frac{1000-10n}{\sqrt{n}}>2$$

解得 $n<98.02$，因此每辆汽车最多装 98 箱才能保证不超载的概率大于 0.977。

代码如下：

```
#第5章/5-3.py
from sympy import *
import numpy as np
t = symbols('t')
eq = 10 * t ** 2 + 2 * t - 1000
n_sqrt1, n_sqrt2 = solveset(eq, t)
#舍去负值
n = (n_sqrt1 ** 2).evalf()
print('最多{}箱才能保证不超载的概率大于0.977'.format(np.floor(n)) )
```

输出如下：

```
最多98箱才能保证不超载的概率大于0.977
```

【例 5-6】 设随机变量序列 $X_1,X_2,\cdots,X_n$ 独立同分布，都服从 $B(1,1/2)$，记 $\Phi(x)$ 为标准正态分布函数，则下面正确的是哪个(　　)。

A. $\lim\limits_{n\to\infty}P\left(\dfrac{\sum\limits_{i=1}^{n}X_i-2n}{2\sqrt{n}}\leqslant x\right)=\Phi(x)$　　B. $\lim\limits_{n\to\infty}P\left(\dfrac{\sum\limits_{i=1}^{n}X_i-2n}{\sqrt{2n}}\leqslant x\right)=\Phi(x)$

C. $\lim\limits_{n\to\infty}P\left(\dfrac{2\sum\limits_{i=1}^{n}X_i-n}{\sqrt{n}}\leqslant x\right)=\Phi(x)$　　D. $\lim\limits_{n\to\infty}P\left(\dfrac{\sum\limits_{i=1}^{n}X_i-n}{\sqrt{n}}\leqslant x\right)=\Phi(x)$

解：根据列维-林德伯格中心极限定理，选 C。

【例 5-7】 设随机变量序列 $X_1,X_2,\cdots,X_n$ 独立同分布，都服从参数为 1 的指数分布，求极限

$$\lim_{n\to\infty}P\left(\sum_{i=1}^{n}X_i\leqslant n\right)$$

解：因为 $X_1,X_2,\cdots,X_n$ 独立同分布，都服从参数为 1 的指数分布，所以 $E(X_i)=1$，$D(X_i)=1$。根据列维-林德伯格中心极限定理，$\sum\limits_{i=1}^{n}X_i$ 近似服从 $N(n,n)$，故有

$$\lim_{n\to\infty}P\left(\frac{\sum\limits_{i=1}^{n}X_i-n}{\sqrt{n}}\leqslant x\right)=\Phi(x)$$

因此

$$\lim_{n\to\infty}P\left(\sum_{i=1}^{n}X_i\leqslant n\right)=\lim_{n\to\infty}P\left(\frac{\sum\limits_{i=1}^{n}X_i-n}{\sqrt{n}}\leqslant 0\right)=\Phi(x)=\frac{1}{2}$$

【例 5-8】 一个加法器同时收到 20 个噪声电压 $V_k(k=1,2,\cdots,20)$，假设噪声电压是互相独立的随机变量，并且都在(0,10)区间上均匀分布，记 $V=\sum_{k=1}^{20}V_k$，求 $P(V>105)$ 的近似值。

解：因为噪声电压服从均匀分布，所以 $E(V_k)=5$，$D(V_k)=100/12$，根据列维-林德伯格中心极限定理，

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

近似服从标准正态分布 $N(0,1)$，则

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

即

$$P(V>105)\approx P(Z>0.387)=1-\Phi(0.387)\approx 0.348$$

代码如下：

```
#第 5 章/5 - 4.py
from scipy.stats import norm
import numpy as np
Z_score = (105 - 100) / (np.sqrt(20) * np.sqrt(100 / 12) )
Z = norm(loc = 0, scale = 1)
p = 1 - Z.cdf(Z_score)
print( 'P(V > 105) 的概率为', p)
```

输出如下：

```
P(V > 105) 的概率为 0.34926767915166934
```

棣莫弗-拉普拉斯中心极限定理是列维-林德伯格中心极限定理的重要推论。设随机变量 $X_n\sim B(n,p)$，则对任意的实数 $x>0$ 有

$$\lim_{n\to\infty}P\left(\frac{X_n-np}{\sqrt{np(1-p)}}\leqslant x\right)=\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{t^2}{2}\right)\mathrm{d}t \tag{5-18}$$

证明：二项分布 $X_n\sim B(n,p)$ 可以写成 n 个独立的符合 $B(1,p)$ 分布的随机变量之和，即

$$X_n=Y_1+Y_2+\cdots+Y_n$$

其中 $Y_1,Y_2,\cdots,Y_n$ 独立同分布，$Y_i\sim B(1,p)$。又因为 $E(Y_i)=p$，$D(Y_i)=p(1-p)$，则由列维-林德伯格中心极限定理可得

$$\lim_{n\to\infty}P\left(\frac{\sum_{i=1}^{n}Y_i-np}{\sqrt{np(1-p)}}\leqslant x\right)=\lim_{n\to\infty}P\left(\frac{X_n-np}{\sqrt{np(1-p)}}\leqslant x\right)$$

$$=\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{t^2}{2}\right)\mathrm{d}t \tag{5-19}$$

证毕。

中心极限定理表明，当 n 充分大时，$B(n,p)$ 的随机变量 X_n 的标准化

$$\frac{X_n-np}{\sqrt{np(1-p)}} \tag{5-20}$$

近似服从标准正态分布 $N(0,1)$，或者说 X_n 近似服从 $N(np,np(1-p))$，即

$$P(a<X_n<b)\approx\Phi\left(\frac{b-np}{\sqrt{np(1-p)}}\right)-\Phi\left(\frac{a-np}{\sqrt{np(1-p)}}\right) \tag{5-21}$$

【例 5-9】 一船舶在海上航行，已知每遭受一次海浪的冲击，纵摇角大于 3°的概率为 1/3。如果该船舶遭受了 90 000 次海浪的冲击，求其中有 29 500 到 30 500 次纵摇角大于 3°的概率是多少？

解：根据题意，可将海浪冲击看作伯努利试验，记 90 000 次海浪的冲击中纵摇角大于 3° 的次数为 X，则有 $X\sim B(90\,000,1/3)$，其分布律为

$$P(X=k)=\begin{bmatrix}90\,000\\k\end{bmatrix}\left(\frac{1}{3}\right)^k\left(\frac{2}{3}\right)^{90\,000-k},\quad k=0,1,2,\cdots,90\,000$$

所求概率为

$$P(29\,500\leqslant X\leqslant 30\,500)=\sum_{k=29\,500}^{30\,500}\begin{bmatrix}90\,000\\k\end{bmatrix}\left(\frac{1}{3}\right)^k\left(\frac{2}{3}\right)^{90\,000-k}$$

这个数字不容易直接计算，可以利用棣莫弗-拉普拉斯中心极限定理计算它的近似值，

$$P(29\,500\leqslant X\leqslant 30\,500)=P\left(\frac{29\,500-np}{\sqrt{np(1-p)}}\leqslant\frac{X-np}{\sqrt{np(1-p)}}\leqslant\frac{30\,500-np}{\sqrt{np(1-p)}}\right)$$

其中 $n=90\,000$，$p=1/3$，则有

$$\begin{aligned}P(29\,500\leqslant X\leqslant 30\,500)&=P\left(-\frac{5}{\sqrt{2}}\leqslant\frac{X-np}{\sqrt{np(1-p)}}\leqslant\frac{5}{\sqrt{2}}\right)\\&=\Phi\left(\frac{5}{\sqrt{2}}\right)-\Phi\left(-\frac{5}{\sqrt{2}}\right)\approx 0.9995\end{aligned}$$

代码如下：

```
#第 5 章/5 - 5.py
from scipy.stats import binom, norm
import numpy as np
#第 1 种方法
X = binom(n = 90000, p = 1 / 3)
p = X.cdf(30500) - X.cdf(29500)
print('第 1 种方法,纵摇角大于 3°的概率是:', p)
#第 2 种方法
n = 90000
p = 1 / 3
a = (29500 - n * p) / np.sqrt(n * p * (1 - p))
b = (30500 - n * p) / np.sqrt(n * p * (1 - p))
X = norm(loc = 0, scale = 1)
p = X.cdf(b) - X.cdf(a)
print('第 2 种方法,纵摇角大于 3°的概率是:', p)
```

输出如下：

```
第 1 种方法,纵摇角大于 3°的概率是: 0.999593113636761
第 2 种方法,纵摇角大于 3°的概率是: 0.999593047982555
```

【例 5-10】 每个学生来开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是 0.05、0.8 和 0.15。如果学校有 400 名学生，设各个学生参加会议的家长人数互相独立且服从同一分布，求

(1) 参加会议的家长总人数超过 450 人的概率。

(2) 有一名家长来参会的学生人数不多于 340 人的概率。

解：(1) 设每个学生来参会的家长人数为 $X_k(k=1,2,\cdots,400)$，X_k 独立同分布，X_k 的分布律见表 5-1。

表 5-1　X_k 的分布律

X_k	0	1	2
p_k	0.05	0.8	0.15

已知 $E(X_k)=1.1, D(X_k)=0.19$。设总家长人数为 $X=\sum_{k=1}^{400}X_k$，由列维-林德伯格中心极限定理，随机变量

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

近似服从正态分布 $N(0,1)$，故而

$$P(X>450)=P\left(\frac{\sum_{k=1}^{400}X_k-400\times 1.1}{\sqrt{400\times 0.19}}>\frac{450-400\times 1.1}{\sqrt{400\times 0.19}}\right)\approx 1-\Phi(1.147)\approx 0.1251$$

(2) 用 Y 表示一名家长参加会议的学生人数，则 $Y\sim B(400,0.8)$，由棣莫弗-拉普拉斯中心极限定理，

$$P(Y\leqslant 340)=P\left(\frac{Y-400\times 0.8}{\sqrt{400\times 0.8\times 0.2}}\leqslant\frac{340-400\times 0.8}{\sqrt{400\times 0.8\times 0.2}}\right)=P\left(\frac{Y-400\times 0.8}{\sqrt{400\times 0.8\times 0.2}}\leqslant 2.5\right)$$

故 $P(Y\leqslant 340)\approx\Phi(2.5)\approx 0.9938$

代码如下：

```
#第 5 章/5-6.py
from scipy.stats import binom, norm
import numpy as np
x = np.array([0, 1, 2])
p = np.array([0.05, 0.8, 0.15])
ex = (x * p).sum()
ex2 = (x ** 2 * p).sum()
dx = ex2 - ex ** 2
print('每个学生家长人数的期望为', ex)
```

```
print('每个学生家长人数的方差为', dx)
#(1) 求家长人数大于 450 的概率
a = (450 - 400 * ex) / np.sqrt(400 * dx)
X = norm(loc = 0, scale = 1)
p = 1 - X.cdf(a)
print('家长人数大于 450 的概率为', p)
#(2) 第 1 种方法
Y = binom(n = 400, p = 0.8)
p = Y.cdf(340)
print('第 1 种方法,有一名家长来参会的学生人数不多于 340 的概率为', p)
#(2) 第 2 种方法
a = (340 - 400 * 0.8) / np.sqrt(400 * 0.8 * 0.2)
X = norm(loc = 0, scale = 1)
p = X.cdf(a)
print('第 2 种方法,有一名家长来参会的学生人数不多于 340 的概率为', p)
```

输出如下：

```
每个学生家长人数的期望为 1.1
每个学生家长人数的方差为 0.18999999999999972
家长人数大于 450 的概率为 0.12567455440511255
第 1 种方法,有一名家长来参会的学生人数不多于 340 的概率为 0.9958883559149133
第 2 种方法,有一名家长来参会的学生人数不多于 340 的概率为 0.9937903346742238
```

5.3 本章习题

1. 一个保险公司有 10 000 个汽车投保人，每个投保人的索赔金额的数学期望为 280 元，标准差为 800 元，求索赔总金额超过 2 700 000 元的概率。

2. 假设各个零件的质量是独立同分布随机变量，其数学期望为 0.5kg，均方差为 0.1，求 5000 只零件的总质量超过 2510kg 的概率是多少？

3. 一批建筑木柱，其中有 80%的长度不小于 3m，现从这批木柱中随机抽取 100 根，求至少有 30 根短于 3m 的概率。

4. 一个食品店有 3 种蛋糕出售，并且出售哪一种是随机的，因而售出一个蛋糕的价格是一个随机变量，其取值为 1 元、1.2 元、1.5 元，概率分别是 0.3、0.2、0.5。如果已知售出 300 个蛋糕，求

(1) 收入至少为 400 元的概率。

(2) 售出价格为 1.2 元的蛋糕大于 60 个的概率。

5. 一栋楼有 200 住户，每个住户拥有的汽车数量 X 的分布律为

$$X=\begin{cases}0, & p=0.1\\ 1, & p=0.6\\ 2, & p=0.3\end{cases}$$

问需要多少车位才能使每辆汽车都具有一个车位的概率至少为 0.95？

6. 某制药厂断言，该厂生产的药品对于某种疾病的治愈率为 0.8。医院任意抽查 100 个服用此药品的病人，如果其中多于 75 人治愈，就接受此断言，否则就拒绝此断言。

(1) 如果实际上此药品的治愈率为 0.8，则接受这一断言的概率是多少？

(2) 如果实际上此药品的治愈率为 0.7，则接受这一断言的概率是多少？

5.4　常见考题解析：大数定律与中心极限定理

本章的主要考点是切比雪夫不等式和中心极限定理。切比雪夫不等式主要以客观题的形式出现，难度不大。计算题集中在中心极限定理部分。

【考题 5-1】 根据测试，某种畅销手机芯片的使用寿命服从均值为 100 周的指数分布，现在随机取出 16 只该芯片，假设它们的寿命是相互独立的。求这 16 只芯片的寿命之和大于 1920 周的概率。

解：设每个芯片的寿命为随机变量 X_i，X_i 服从均值为 100 周的指数分布，故有 $E(X_i)=100$，$\sigma^2(X_i)=10\,000$。根据中心极限定理，近似地有

$$X=\sum_{i=1}^{16}X_i \sim N(100\times 16,10\,000\times 16)=N(1600,400^2)$$

则问题转化为求概率 $P(X\geqslant 1920)$。

$$P(X\geqslant 1920)=P\left(\frac{X-1600}{400}\geqslant\frac{1920-1600}{400}\right)=P(Z\geqslant 0.8)=0.212$$

代码如下：

```
#第 5 章/5 - 7.py
from scipy.stats import norm
import numpy as np
#第 1 种方法
X = norm(loc = 1600, scale = 400)
p = 1 - X.cdf(1920)
print('第 1 种方法,寿命之和大于 1920 周的概率为', p)
#第 2 种方法
X = norm(loc = 0, scale = 1)
p = 1 - X.cdf(0.8)
print('第 2 种方法,寿命之和大于 1920 周的概率为', p)
```

输出如下：

```
第 1 种方法,寿命之和大于 1920 周的概率为 0.21185539858339664
第 2 种方法,寿命之和大于 1920 周的概率为 0.21185539858339664
```

【考题 5-2】 利用中心极限定理解决如下问题

(1) 某保险公司有 10 000 名投保人客户，每个投保人的理赔金额不等，它的数学期望是 280 元，标准差为 800 元，求理赔总金额超过 2 700 000 元的概率。

(2) 某保险公司有 50 张理赔单，金额不等。它们的数学期望是 5，方差是 6。求 50 张

理赔单赔付总金额大于300的概率(假设各个理赔单的理赔金额是相互独立的)。

解:(1) 设 X 为理赔总金额,由中心极限定理知道

$$P(X>2\,700\,000)=P\left(\frac{X-280\times10\,000}{800\times\sqrt{10\,000}}>\frac{2\,700\,000-280\times10\,000}{800\times\sqrt{10\,000}}\right)=P(Z>-1.25)$$

其中 Z 服从标准正态分布。可查表得 $P(Z>-1.25)=0.894$。

(2) 设 X 为理赔总金额,由中心极限定理知道

$$P(X>300)=P\left(\frac{X-5\times50}{\sqrt{6}\times\sqrt{50}}>\frac{300-250}{\sqrt{6}\times\sqrt{50}}\right)=P\left(Z>\frac{50}{\sqrt{6}\times\sqrt{50}}\right)$$

其中,Z 服从标准正态分布。可查表得概率值。

代码如下:

```
#第5章/5-8.py
from scipy.stats import norm
import numpy as np
#第1问,第1种方法
X = norm(loc = 280 * 10000, scale = np.sqrt(10000 * 640000))
p = 1 - X.cdf(2700000)
print('第1问,第1种方法,赔偿总金额超过2 700 000的概率为', p)
#第1问,第2种方法
X = norm(loc = 0, scale = 1)
p = 1 - X.cdf(-1.25)
print('第1问,第2种方法,赔偿总金额超过2 700 000的概率为', p)
#第2问,第1种方法
X = norm(loc = 250, scale = np.sqrt(300))
p = 1 - X.cdf(300)
print('第2问,第1种方法,赔偿合计超过300的概率为', p)

#第2问,第2种方法
X = norm(loc = 0, scale = 1)
p = 1 - X.cdf(50 / np.sqrt(300))
print('第2问,第2种方法,赔偿合计超过300的概率为', p)
```

输出如下:

```
第1问,第1种方法,赔偿总金额超过2 700 000的概率为 0.8943502263331446
第1问,第2种方法,赔偿总金额超过2 700 000的概率为 0.8943502263331446
第2问,第1种方法,赔偿合计超过300的概率为 0.0019462085613892732
第2问,第2种方法,赔偿合计超过300的概率为 0.0019462085613892732
```

【考题 5-3】 设随机变量服从均匀分布,即 $U\sim U(-0.5,0.5)$,解决下面的问题

(1) 将与 U 独立同分布的1 500个随机变量相加,求它们的和的绝对值超过15的概率。

(2) 最多可有多少个随机变量相加使总和的绝对值小于10的概率不小于0.9?

解:(1) 设1500个随机变量之和的随机变量为 X,由中心极限定理,近似地有 X 服从均值为0,方差为1500/12=125的正态分布,因此它们的和的绝对值不超过15的概率为

$$P(-15\leqslant X\leqslant15)=P\left(\frac{-15-0}{\sqrt{125}}\leqslant\frac{X-0}{\sqrt{125}}\leqslant\frac{15-0}{\sqrt{125}}\right)=P\left(\frac{-15}{\sqrt{125}}\leqslant Z\leqslant\frac{15}{\sqrt{125}}\right)$$

其中 Z 服从标准正态分布。故而可得绝对值超过 15 的概率为 $1-P(-15\leqslant X\leqslant 15)=0.1797$。

(2) 设有 n 个随机变量相加，设和为 X，那么由中心极限定理，近似地有 X 服从均值为 0，方差为 $n/12$ 的正态分布，要使

$$P(-10\leqslant X\leqslant 10)=P\left(\frac{-10-0}{\sqrt{n/12}}\leqslant Z\leqslant\frac{10-0}{\sqrt{n/12}}\right)=P\left(\frac{-10}{\sqrt{n/12}}\leqslant Z\leqslant\frac{10}{\sqrt{n/12}}\right)\geqslant 0.9$$

则需要

$$\frac{10}{\sqrt{n/12}}\geqslant Z_{0.95}$$

即该数字大于或等于标准正态分布的 0.95 分位点，查表可得 $Z_{0.95}=1.6449$，解得 $n\leqslant 443$。

代码如下：

```
#第 5 章/5 - 9.py
from scipy.stats import norm
import numpy as np
#第 1 问,第 1 种方法
X = norm(loc = 0, scale = np.sqrt(1500 / 12))
p = 1 - (X.cdf(15) - X.cdf( - 15))
print( '第 1 问,第 1 种方法,所求概率为', p)
#第 1 问,第 2 种方法
X = norm(loc = 0, scale = 1)
p = 1 - ( X.cdf(15 / np.sqrt(125)) - X.cdf( - 15 / np.sqrt(125) ) )
print( '第 1 问,第 2 种方法,所求概率为', p)
#第 2 问
a = norm.ppf(0.95)
n = (10 / a) ** 2 * 12 #n 越大,10/sqrt(n/12)越小
print( 'n 至多为', np.floor(n))
```

输出如下：

```
第 1 问,第 1 种方法,所求概率为 0.17971249487899987
第 1 问,第 2 种方法,所求概率为 0.17971249487899987
n 至多为 443.0
```

【考题 5-4】 设某种带包装的零食的质量是随机变量，它们相互独立且服从同样的分布。通过测定该分布的数学期望为 0.5kg，标准差为 0.1kg，求 5000 袋这种零食总质量超过 2510kg 的概率是多少？

解：已知随机变量的均值和方差，由中心极限定理知

$$\sum_{i=1}^{5000}X_i=X\sim N(0.5\times 5000,5000\times 0.1^2)=N(2500,50)$$

则可知 $P(X>2\,510)$ 等于 0.0786。

代码如下：

```
#第 5 章/5 - 10.py
from scipy.stats import norm
import numpy as np
#第 1 种方法
X = norm(loc = 2500, scale = np.sqrt(50))
p = 1 - X.cdf(2510)
print('第 1 种方法,总质量超过 2510kg 的概率为', p)
#第 2 种方法
X = norm(loc = 0, scale = 1)
p = 1 - X.cdf( (2510 - 2500) / np.sqrt(50) )
print('第 2 种方法,总质量超过 2510kg 的概率为', p)
```

输出如下：

```
第 1 种方法,总质量超过 2510kg 的概率为 0.07864960352514261
第 2 种方法,总质量超过 2510kg 的概率为 0.07864960352514261
```

【考题 5-5】 有一批钢材，其中有 80%的长度不小于 3m，现从这批钢材中抽取 100 根做检测，求其中至少有 30 件小于 3m 的概率。

解：以 3m 为标准，钢材的长度服从二项分布，即 $B(100,0.8)$，由二项分布知道所求概率为

$$p=\sum_{k=30}^{100}C_{100}^{k}0.2^{k}0.8^{100-k}$$

这个数值不容易笔算，可用中心极限定理来近似，即

$$P(X\geqslant 30)=P\left(\frac{X-np}{\sqrt{np(1-p)}}\geqslant\frac{30-np}{\sqrt{np(1-p)}}\right)=P\left(\frac{X-np}{\sqrt{np(1-p)}}\geqslant 2.5\right)=P(Z\geqslant 2.5)$$

其中，$n=100$，$p=0.2$，Z 服从标准正态分布。

代码如下：

```
#第 5 章/5 - 11.py
from scipy.stats import norm, binom
import numpy as np
#第 1 种方法,用二项分布,精确计算
X = binom(n = 100, p = 0.8)
p = X.cdf(70)
print('用二项分布,所求概率为', p)
#第 2 种方法,用中心极限定理,近似
X = norm(loc = 80, scale = 4)
p2 = X.cdf(70)
print('用中心极限定理,所求概率为', p2)
#或者
X = norm(loc = 0, scale = 1)
p3 = 1 - X.cdf(2.5)
print('用中心极限定理,化为标准正态分布,所求概率为', p3)
```

输出如下：

```
用二项分布,所求概率为 0.011248978720991605
用中心极限定理,所求概率为 0.006209665325776132
用中心极限定理,化为标准正态分布,所求概率为 0.006209665325776159
```

【考题 5-6】 加工某种特殊的零件需要两个阶段，第一阶段需要的时间（小时数）服从均值为 0.2 的指数分布，第二阶段所需时间服从均值为 0.3 的指数分布，并且这两个阶段相互独立。现在需要加工 20 个此类零件，求加工所需的总时间不超过 8h 的概率。

解：由指数分布的分布函数可知，若指数分布的均值为 θ，则它的方差为 θ^2。设第一阶段所用时间为随机变量 $X_1 \sim \exp(0.2)$，第二阶段所用时间为随机变量 $X_2 \sim \exp(0.3)$，则总时间 $X=X_1+X_2$ 的期望为 $E(X)=0.2+0.3=0.5$，方差 $D(X)=0.2^2+0.3^2=0.13$。设加工第 i 个零件所需时间为 $X^{(i)}$，根据中心极限定理，总的加工时间

$$X^* = \sum_{i=1}^{20} X^{(i)} \sim N(20\times 0.5, 20\times 0.13) = N(10, 2.6)$$

根据 $N(10,2.6)$的分布计算出概率为 0.1074。或者将该正态分布化为标准正态分布

$$P(X^* \leqslant 8) = P\left(\frac{X^*-10}{\sqrt{2.6}} \leqslant \frac{8-10}{\sqrt{2.6}}\right) = P\left(Z \leqslant \frac{-2}{\sqrt{2.6}}\right)$$

其中，Z 服从标准正态分布。查表可得所求概率为 0.1074。

代码如下：

```
#第 5 章/5 - 12.py
from scipy.stats import norm
import numpy as np
#第 1 种方法,标准正态分布
X = norm(loc = 0, scale = 1)
p = X.cdf( - 2 / np.sqrt(2.6))
print('第 1 种方法,标准正态分布,8h 内完成的概率为', p)
#第 2 种方法,正态分布
X = norm(loc = 10, scale = np.sqrt(2.6))
p = X.cdf(8)
print('第 2 种方法,正态分布,8h 内完成的概率为', p)
```

输出如下：

```
第 1 种方法,标准正态分布,8h 内完成的概率为 0.10742347370282451
第 2 种方法,正态分布,8h 内完成的概率为 0.10742347370282451
```

【考题 5-7】 某小商品超市有 3 种零食出售，由于这 3 种零食作为活动奖品随机出售，所以售出一袋零食的价格是一个随机变量，3 种零食的价格分别是 1 元、1.2 元和 1.5 元，取以上值的概率分别是 0.3、0.2 和 0.5。今假设共卖出 300 袋零食，

（1）求此超市收入超过 400 元的概率。

（2）求售出价格为 1.2 元的零食多于 60 袋的概率。

解：设每袋零食是随机变量 X，X 的期望为 $E(X)=1\times 0.3+1.2\times 0.2+1.5\times 0.5=$

1.29，X 的方差为 $D(X)=1\times0.3+1.44\times0.2+2.25\times0.5-E(X)^2=0.0489$。

(1) 根据中心极限定理，近似地有，卖出 300 袋零食的总收入

$$X=\sum_{i=1}^{300}X_i\sim N(300\times1.29,300\times0.0489)=N(387,14.67)$$

根据 $N(387,14.67)$的分布计算出概率为 0.0003。或者将该正态分布化为标准正态分布

$$P(X>400)=P\left(\frac{X-387}{\sqrt{14.67}}>\frac{400-387}{\sqrt{14.67}}\right)=P\left(Z>\frac{13}{\sqrt{14.67}}\right)$$

其中，Z 服从标准正态分布。查表可得所求概率为 0.0003。

(2) 这一问仅对价格为 1.2 元的零食感兴趣。将所有的零食分为两类，一类是 1.2 元单价的零食，另一类是其他，这是一个二项分布的模型。价格 1.2 元的零食服从 $B(300,0.2)$ 分布，因此多于 60 袋的概率为

$$P=\sum_{k=60}^{300}C_{300}^{k}0.2^k0.8^{300-k}$$

这个概率不容易笔算，使用中心极限定理将其化为正态分布

$$P(N\geqslant60)=P\left(\frac{N-60}{\sqrt{300\times0.2\times0.8}}\geqslant\frac{60-60}{\sqrt{300\times0.2\times0.8}}\right)=P(Z\geqslant0)$$

其中，Z 服从标准正态分布，容易计算这个概率为 0.5。

代码如下：

```
#第 5 章/5 - 13.py
from scipy.stats import norm, binom
import numpy as np
#第 1 问,第 1 种方法
S = norm(loc = 387, scale = np.sqrt(14.67))
p = 1 - S.cdf(400)
print('第 1 问,第 1 种方法,所求概率为', p)
#第 1 问,第 2 种方法
X = norm(loc = 0, scale = 1)
p = 1 - X.cdf(13 / np.sqrt(14.67))
print('第 1 问,第 2 种方法,所求概率为', p)
#第 2 问,用中心极限定理
S1 = norm(loc = 60, scale = np.sqrt(48))
print('第 2 问,用中心极限定理,概率为',1 - S1.cdf(60))
#第 2 问,用二项分布
S2 = binom(n = 300, p = 0.2)
p2 = 1 - S2.cdf(59)
print('第 2 问,用二项分布,概率为', p2)
```

输出如下：

```
第 1 问,第 1 种方法,所求概率为 0.0003442367509621791
第 1 问,第 2 种方法,所求概率为 0.0003442367509621791
第 2 问,用中心极限定理,概率为 0.5
第 2 问,用二项分布,概率为 0.5230223389695331
```

【考题 5-8】 某复杂输电设备由 100 个互相独立工作的零部件所组成，设备在正常工作时，每个零部件损坏的概率为 0.1，为了使设备正常维持工作，至少需要 85 个零部件正常运行，求整套设备正常工作的概率。

解：每个零部件是否损坏服从伯努利分布，损坏的概率为 0.1，整套设备 100 个零部件服从二项分布 $B(100,0.1)$。根据题意，设备正常工作，损坏的零件不能超过 15 个，这个概率可以表示为

$$P(N \leqslant 15)=\sum_{k=1}^{15} C_{100}^{k} 0.1^{k} 0.9^{300-k}$$

这个概率不容易笔算得出，使用中心极限定理将其化为正态分布

$$P(N \leqslant 15)=P\left(\frac{N-100\times 0.1}{\sqrt{100\times 0.1\times 0.9}} \leqslant \frac{15-100\times 0.1}{\sqrt{100\times 0.1\times 0.9}}\right)=P\left(Z \leqslant \frac{5}{3}\right)$$

查表或利用软件计算此概率为 0.9522。

代码如下：

```
#第 5 章/5 - 14.py
from scipy.stats import norm, binom
#用中心极限定理
S = norm(loc = 90, scale = 3)
p = 1 - S.cdf(85)
print('用中心极限定理,设备正常工作的概率为', p)
#用二项分布
s = 0
for k in range(85,101):
    s += binom(n = 100, p = 0.9).pmf(k)
print('用二项分布,设备正常工作的概率为', s)
#用二项分布,第二种方法
p2 = 1 - binom(n = 100, p = 0.9).cdf(84)
print('用二项分布,第二种方法,用二项分布积累函数,概率为', p2)
```

输出如下：

```
用中心极限定理,设备正常工作的概率为 0.9522096477271853
用二项分布,设备正常工作的概率为 0.9601094728889118
用二项分布,第二种方法,用二项分布积累函数,概率为 0.9601094728889168
```

【考题 5-9】 假设某随机变量 X，已知它的数学期望为 2.2，标准差为 1.4。

(1) 假设有一个与 X 独立同分布的样本$(X_1,X_2,\cdots,X_{52})$，其样本均值为 $\overline{X}$，试用中心极限定理求 $\overline{X}$ 的近似分布，并依此计算概率 $P(\overline{X}<2)$。

(2) 求样本$(X_1,X_2,\cdots,X_{52})$之和小于 100 的概率。

解：(1) 由中心极限定理知道样本$(X_1,X_2,\cdots,X_{52})$之和近似地服从正态分布

$$\sum_{k=1}^{52} X_k = X \sim N(52\times 2.2, 52\times 1.4^2)=N(114.4,101.92)$$

再由正态分布的性质，

$$\overline{X} \sim N(2.2,1.4^2/52)=N(2.2,1.4^2/52)$$

根据正态分布，可计算出概率 $P(\overline{X}<2)$为 0.1515。

(2) 由中心极限定理，已知样本$(X_1,X_2,\cdots,X_{52})$之和的分布

$$\sum_{k=1}^{52}X_k=X \sim N(52\times 2.2,52\times 1.4^2)=N(114.4,101.92)$$

根据分布 $N(114.4,101.92)$可计算小于 100 的概率。

代码如下：

```
#第 5 章/5 - 15.py
from scipy.stats import norm
import numpy as np
#第 1 问
X_ = norm(loc = 2.2, scale = 1.4 / np.sqrt(52))
p = X_.cdf(2)
print('第 1 问的概率为', p)
#第 2 问,第 1 种方法
S = norm(loc = 52 * 2.2, scale = 1.4 * np.sqrt(52))
p2 = S.cdf(100)
print('第 2 问,第 1 种方法,概率为', p2)
#第 2 问,第 2 种方法
X = norm(loc = 0, scale = 1)
p = X.cdf((100 - 114.4) / np.sqrt(101.92) )
print('第 2 问,第 2 种方法,概率为', p2)
```

输出如下：

```
第 1 问的概率为 0.15146803666167452
第 2 问,第 1 种方法,概率为 0.07688050541745406
第 2 问,第 2 种方法,概率为 0.07688050541745406
```

【考题 5-10】 有一款柴油内燃机一氧化碳的排放量的数学期望是 0.9，标准差是 1.9，某工厂有 100 台此种柴油内燃机，用 $\overline{X}$ 表示这些内燃机一氧化碳排放量的均值，求一个特殊的数字 M，使 $\overline{X}>M$ 的概率不超过 0.01。

解：此题的关键在于找到 $\overline{X}$ 的分布，由中心极限定理，近似地有

$$\overline{X} \sim N\left(0.9,\frac{1.9^2}{100}\right)=N(0.9,0.0361)$$

因此取 M 为正态分布 $N(0.9,0.0361)$的 0.99 分位点即可。

```
#第 5 章/5 - 16.py
from scipy.stats import norm
X_ = norm(loc = 0.9, scale = 0.19)
M = X_.ppf(0.99)
print('M 的最小值为', M)
```

输出如下：

```
M 的最小值为 1.3420060960677598
```

【考题 5-11】 有一位心理学家对学生做心理测试，他把学生分成两组，每组有 80 人，两组学生相互独立，测试指标值服从同一种分布。指标的数学期望是 5，方差为 0.3。用 $\overline{X}$ 和 $\overline{Y}$ 表示第 1 组和第 2 组指标的算术平均值。

(1) 求概率 $P(4.9<\overline{X}<5.1)$。

(2) 求概率 $P(-0.1<\overline{X}-\overline{Y}<0.1)$。

解：根据题意，指标的数学期望是 5，方差为 0.3，$\overline{X}$ 和 $\overline{Y}$ 的分布分别是

$$\overline{X} \sim N\left(5,\frac{0.3}{80}\right)=N(5,0.00375),\overline{Y} \sim N\left(5,\frac{0.3}{80}\right)=N(5,0.00375)$$

(1) 根据 $\overline{X}$ 的分布可计算出概率 $P(4.9<\overline{X}<5.1)$。

(2) 已知 $\overline{X}$ 和 $\overline{Y}$ 的分布，可知 $\overline{X}-\overline{Y}$ 的分布

$$\overline{X}-\overline{Y}\sim N\left(0,\frac{0.6}{80}\right)=N(0,0.0075)$$

从而可算出概率 $P(-0.1<\overline{X}-\overline{Y}<0.1)$。

代码如下：

```
#第 5 章/5 - 17.py
from scipy.stats import norm
import numpy as np
#第 1 问
X_ = norm(loc = 5, scale = np.sqrt(0.3 / 80) )
p = X_.cdf(5.1) - X_.cdf(4.9)
print('第 1 问的概率值为', p)
#第 2 问
XY_ = norm(loc = 0, scale = np.sqrt(0.6 / 80))
p2 = XY_.cdf(0.1) - XY_.cdf( - 0.1)
print('第 2 问的概率值为', p2)
```

输出如下：

```
第 1 问的概率值为 0.8975295651402493
第 2 问的概率值为 0.7517869210100763
```

【考题 5-12】 一高校计算机实验楼有 200 台服务器，每台服务器需要的管理员数量 X 满足以下分布：$P(X=0)=0.1$，$P(X=1)=0.6$，$P(X=2)=0.3$。求需要多少管理员才能使每台服务器有一个管理员的概率至少为 0.95？

解：根据随机变量 X 的分布，可以算出它的期望 $E(X)=1.2$，方差 $D(X)=0.36$。由中心极限定理，近似地有

$$Y=\sum_{i=1}^{200}X_i \sim N(200\times 1.2,200\times 0.36)=N(240,72)$$

显然管理员的数量应该是正态分布 $N(240,72)$的 0.95 分位点。也可将该正态分布转

化为标准正态分布求 0.95 分位点。设 N 为管理员数量，有

$$P(Y \leqslant N)=P\left(\frac{Y-240}{\sqrt{72}} \leqslant \frac{N-240}{\sqrt{72}}\right)=P\left(Z \leqslant \frac{N-240}{\sqrt{72}}\right)>0.05$$

则$(N-240)/\sqrt{72}$是标准正态分布的 0.95 分位点，解出 $N=254$。

代码如下：

```
#第 5 章/5 - 18.py
from scipy.stats import norm
import numpy as np
#第 1 种方法
k = norm(loc = 240, scale = np.sqrt(72)).ppf(0.95)
k = np.floor(k) + 1
print('第 1 种方法,至少需要{}个管理员 '.format(k))

#第 2 种方法,化为标准正态分布
k2 = 240 + np.sqrt(72) * norm.ppf(0.95)
k2 = np.floor(k2) + 1
print('第 2 种方法,化为标准正态分布,至少需要{}个管理员 '.format(k2))
```

输出如下：

```
第 1 种方法,至少需要 254.0 个管理员
第 2 种方法,化为标准正态分布,至少需要 254.0 个管理员
```

【考题 5-13】 某种芯片的使用寿命的数学期望 μ 未知，已知它的方差为 $\sigma^2=400$，为了估计 μ，随机独立抽取 n 只芯片进行独立测试。假设它们的寿命分别是 $X_1, X_2, \cdots, X_n$，用 $\overline{X}$ 表示它们的平均值，求 n 至少是多少才能使 $P(|\overline{X}-\mu|<1) \geqslant 0.95$。

解：此题需要算出 $\overline{X}$ 的分布，由中心极限定理，构造如下标准正态分布随机变量 Z

$$Z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

则显然有

$$P(|\overline{X}-\mu|<1)=P\left(\left|\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right|<\frac{\sqrt{n}}{\sigma}\right)=P\left(|Z|<\frac{\sqrt{n}}{\sigma}\right) \geqslant 0.95$$

这说明$\sqrt{n}/\sigma$ 是标准正态分布的 0.975 分位点。查表求出此分位点后可解出 $n=1537$。

代码如下：

```
#代码清单 5 - 19
#例 13
from scipy.stats import norm
import numpy as np
t = (20 * norm.ppf(0.975) ) ** 2
print('n 的最小值为', np.floor(t) + 1)
```

输出如下：

```
n 的最小值为 1537.0
```

【考题 5-14】 某制药厂宣称，该厂生产的某种药品对于治疗一种疾病的治愈率达到0.8，医院随机抽查了 100 个服用此药物的患者，如果其中有多于 75 人治愈，就接受该制药厂断言，否则就拒绝。

(1) 如果实际上此药品的治愈率真是 0.8，求接受这一断言的概率是多少？

(2) 如果实际上此药品的治愈率只有 0.7，求接受这一断言的概率是多少？

解：对于每个患者，要么治愈要么不治愈，这是一个伯努利分布随机变量，因此 100 个患者的和服从二项分布 $B(100,p)$，其中 p 是未知参数。那么 75 人治愈的概率为

$$P=\sum_{k=75}^{100}C_{100}^{k}p^{k}(1-p)^{100-k}$$

这个概率不好笔算，利用中心极限定理近似地有

$$P\left(\sum_{k=1}^{100}X_i\geqslant 75\right)=P\left(\frac{\sum_{k=1}^{100}X_i-100p}{\sqrt{100p(1-p)}}\geqslant\frac{75-100p}{\sqrt{100p(1-p)}}\right)=P\left(Z\geqslant\frac{75-100p}{\sqrt{100p(1-p)}}\right)$$

其中，Z 服从标准正态分布。

(1) 当 $p=0.8$ 时，

$$P\left(\sum_{k=1}^{100}X_i\geqslant 75\right)=P(Z\geqslant -1.25)$$

查表得概率为 0.8944。

(2) 当 $p=0.7$ 时

$$P\left(\sum_{k=1}^{100}X_i\geqslant 75\right)=P\left(Z\geqslant\frac{1}{2\sqrt{0.21}}\right)$$

查表得概率为 0.1376。

代码如下：

```
#第 5 章/5 - 20.py
from scipy.stats import binom, norm
import numpy as np
#第(1)问,中心极限定理
p = 1 - norm.cdf( - 1.25)
print( '第(1)问,中心极限定理,概率为', p)
#第(1)问,二项分布
p = 1 - binom(n = 100, p = 0.8).cdf(74)
print( '第(1)问,二项分布,概率为', p)

#第(2)问,中心极限定理
p = 1 - norm.cdf(1 / (np.sqrt(0.21) * 2) )
print( '中心极限定理:', p)
```

输出如下：

```
第(1)问,中心极限定理,概率为 0.8943502263331446
第(1)问,二项分布,概率为 0.9125246153564271
第(2)问,中心极限定理,概率为 0.13761676203741713
```

5.5 本章常用的 Python 函数总结

本章主要用到二项分布做精确计算,用正态分布做近似计算(利用中心极限定理)。导入函数的方式为 from scipy.stats import binom,norm。本章常用的 Python 函数见表 5-2。

表 5-2 本章常用的 Python 函数

函 数	代 码
二项分布的积累函数	binom(n = n,p = p).cdf(x) 其中 n 为试验总次数,p 为成功率,x 为分布函数中的自变量
正态分布的积累函数	norm(loc = loc,scale = scale).cdf(x) 其中 loc 为正态分布的均值,scale 为正态分布标准差,x 为分布函数中的自变量
正态分布的分位点函数	norm(loc = loc,scale = scale).ppf(q) 其中 loc 为正态分布的均值,scale 为正态分布标准差,q 为概率值,介于 0 和 1 之间

5.6 本章上机练习

实训环境:

(1) 使用 Python 3.x 版本。

(2) 使用 IPython 或 Jupyter Notebook 交互式编辑器,推荐使用 Anaconda 发行版中自带的 IPython 或 Jupyter Notebook。

【实训 5-1】 执行以下代码,解释所观察到的现象。

代码如下:

```
#第 5 章/5-21.py
import numpy as np
from scipy.stats import binom
import matplotlib.pyplot as plt
n = 10
p = 0.4
sample_size = 1500
expected_value = n * p
N_samples = range(1,sample_size,10)
for k in range(3):
    binom_rv = binom(n = n, p = p)
    X = binom_rv.rvs(size = sample_size)
    sample_average = [X[:i].mean() for i in N_samples]
    plt.plot(N_samples, sample_average, label = 'average of sample {}'.format(k))

plt.plot(N_samples, expected_value * np.ones_like(sample_average),
         ls = '--',label = 'true expected value: n*p = {}'.format(n*p),
```

```
        c = 'k')
plt.legend()
plt.grid(ls = '--')
plt.tick_params(direction = 'in')
plt.show()
```

输出结果如图 5-1 所示。

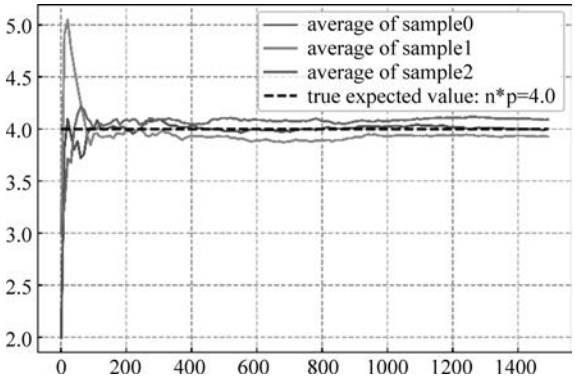

图 5-1　大数定律模拟,伯努利分布

在这个试验中,设置了 3 个实验组,分别用 3 种颜色表示。在每一组试验中,随着样本量的逐渐增大,样本均值越来越收敛于随机变量的期望。

【实训 5-2】 执行以下代码,解释所观察到的现象。

```
#第 5 章/5 - 22.py
import numpy as np
from scipy.stats import norm
import matplotlib.pyplot as plt
n = 100000

norm_rvs = norm(loc = 0,scale = 10).rvs(size = n)
_ = plt.hist(norm_rvs, density = True, alpha = 0.3, color = 'b',
         bins = 100, label = 'original')
mean_array = [ ]
n_samples = 5000
for i in range(n_samples):
    sample = np.random.choice(norm_rvs, size = 10, replace = False)
    mean_array.append(np.mean(sample))

plt.hist(mean_array,density = True, alpha = 0.3, color = 'r',
         bins = 100, label = 'sample size = 10')

for i in range(n_samples):
    sample = np.random.choice(norm_rvs, size = 50, replace = False)
    mean_array.append(np.mean(sample))
plt.hist(mean_array,density = True, alpha = 0.3, color = 'g',
         bins = 100, label = 'sample size = 50')
```

```
plt.gca().axes.set_xlim(-50,50)
plt.legend(loc = 'best')
plt.grid(ls = '--')
plt.tick_params(direction = 'in')
plt.show()
```

代码输出如图 5-2 所示。

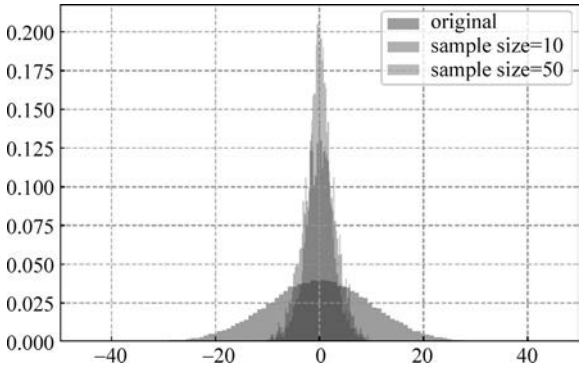

图 5-2　大数定律模拟，正态分布

在这个试验中，首先生成 100 000 个正态分布随机样本，分别从这 100 000 个样本中每次选出 10 个样本和 50 个样本，分别计算平均值，重复 5000 次，记录它们的平均值，画出平均值的直方图。从图 5-2 中发现，每次选出的样本数量越多，样本均值的分布越来越向期望值集中，这就是大数定律所要表达的意思。

【实训 5-3】 执行以下代码，解释所观察到的现象。

```
#第 5 章/5-23.py
import numpy as np
from scipy.stats import geom
import matplotlib.pyplot as plt
_, ax = plt.subplots(2,2)
p = 0.3
N = 1000000
geom_rvs = geom(p = p).rvs(size = N)
mean, var, skew, kurt = geom(p = p).stats(moments = 'mvsk')
ax[0,0].hist(geom_rvs, bins = 100, density = True)
ax[0,0].set_title('geometric distribution')
ax[0,0].grid(ls = '--')
ax[0,0].tick_params(direction= 'in')
n_array = [0,2,5,50]
for i in range(1,4):
    Z_array = [ ]
    n = n_array[i]
    for j in range(100000):
        sample = np.random.choice(geom_rvs,n)
        Z_array.append((sum(sample) - n * mean)/np.sqrt(n * var))
    ax[i//2, i%2].hist(Z_array, bins = 100, density = True)
```

```
    ax[i//2, i%2].set_title('n={}'.format(n))
    ax[i//2, i%2].set_xlim(-3,3)
ax[i//2, i%2].grid(ls = '--')
ax[i//2, i%2].tick_params(direction='in')

plt.show()
```

代码输出结果如图 5-3 所示。

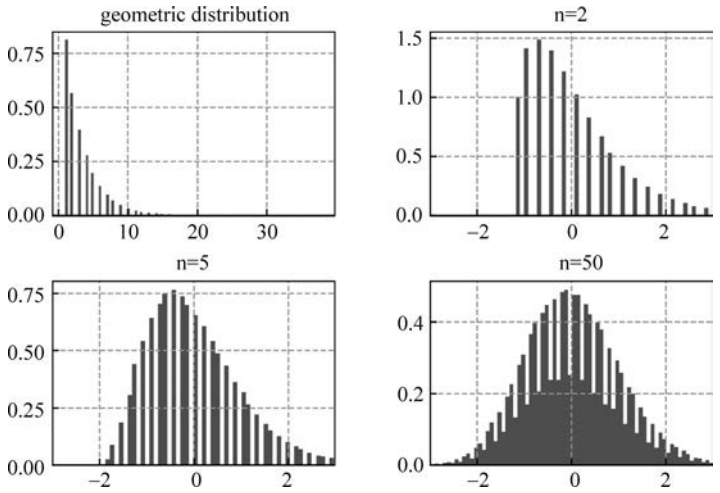

图 5-3　中心极限定理模拟

在这个试验中，可以发现随着采样数量的增加，随机变量和的标准化越来越接近于标准正态分布。

【实训 5-4】 执行以下代码，解释所观察到的现象。

```
#第5章/5-24.py
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.patches import Circle
from scipy.stats import uniform

n = 100000
r = 1.0
o_x, o_y = (0.0, 0.0)
uniform_x = uniform(o_x - r, 2 * r).rvs(n)
uniform_y = uniform(o_y - r, 2 * r).rvs(n)
d_array = np.sqrt((uniform_x - o_x) ** 2 + (uniform_y - o_y) ** 2 )
res = sum(np.where(d_array < r,1,0))
pi = (res/n)/(r**2) * (2*r) ** 2
fig, ax = plt.subplots(1,1)
ax.plot(uniform_x, uniform_y,'ro',alpha = 0.2, markersize = 0.3)
plt.axis('equal')
Circle = Circle(xy = (o_x,o_y),radius = r, alpha = 0.5)
ax.add_patch(Circle)
plt.grid(ls = '--')
```

代码输出如图 5-4 所示。

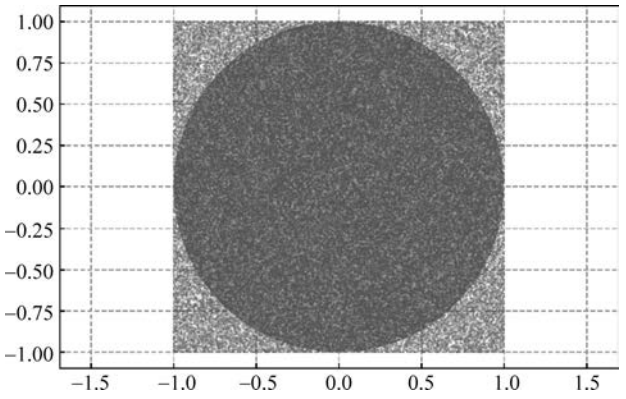

图 5-4　大数定律与数值模拟

【实训 5-5】 设某供电电网有 10 000 盏灯，夜晚每一盏灯开灯的概率都是 0.7，而所有电灯开或关是相互独立的，试估计开灯数量在 7000 至 8000 的概率。

代码如下：

```
#第 5 章/5 - 25.py
from scipy.stats import binom, norm
import numpy as np
#用二项分布
p = 0.7
n = 10000
X = binom(n = n, p = p)
p = X.cdf(8000) - X.cdf(7000)
print('用二项分布,概率为', p)
#用中心极限定理
p = 0.7
n = 10000
np_ = n * p
npq_ = n * p * (1 - p)
left = (7000 - np_) / np.sqrt(npq_)
right = (8000 - np_) / np.sqrt(npq_)
X = norm(loc = 0, scale = 1)
p = X.cdf(right) - X.cdf(left)
print('用中心极限定理,概率为', p)
```

输出如下：

```
用二项分布,概率为 0.4962276224574802
用中心极限定理,概率为 0.5
```

【实训 5-6】 某台服务器有 120 个终端，每个终端在 1h 内平均有 3min 使用打印机，假设各终端使用打印机与否是互相独立的，求至少有 10 个终端同时使用打印机的概率。

代码如下：

```
#第 5 章/5 - 26.py
from scipy.stats import binom, norm
import numpy as np
#用二项分布
n = 120
p = 3 / 60
X = binom(n = n, p = p)
m = 9
q = 1 - X.cdf(m)
print('用二项分布,概率为', q)
#用中心极限定理
c = (10 - n * p) / np.sqrt(n * p * (1 - p))
X = norm(loc = 0, scale = 1)
q = 1 - X.cdf(c)
print('用中心极限定理,概率为', q)
```

输出如下：

```
用二项分布,概率为 0.07862994670181611
用中心极限定理,概率为 0.04692635571997261
```

【实训 5-7】 有某种仪表 200 台，调整无误的概率为 0，调整过大或过小的概率都是 0.5。求调整过大的仪表在 95 台到 105 台的概率是多少？

代码如下：

```
#第 5 章/5 - 27.py
from scipy.stats import binom, norm
import numpy as np
#用二项分布
n = 200
p = 0.5
X = binom(n = n, p = p)
m1, m2 = 95, 105
q = X.cdf(m2) - X.cdf(m1)
print('用二项分布,概率为', q)
#用中心极限定理
left = (m1 - n * p) / np.sqrt(n * p * (1 - p))
right = (m2 - n * p) / np.sqrt(n * p * (1 - p))
X = norm(loc = 0, scale = 1)
q = X.cdf(right) - X.cdf(left)
print('用中心极限定理,概率为', q)
```

输出如下：

```
用二项分布,概率为 0.5193120773631319
用中心极限定理,概率为 0.5204998778130465
```