第五章　大数定律与中心极限定理

大数定律和中心极限定理的研究，在概率论的发展中占有重要地位，是概率论成为一门成熟的数学学科的重要标志之一，而且仍然是现代概率论的重要研究方向之一.

大数定律有重要的理论意义，是概率稳定性和大量观测结果平均水平稳定性的的数学定理；中心极限定理是在一定条件下关于“大量随机变量之和的极限分布是正态分布”的一系列定理的总称。我们这里只介绍实际中最常用的极限定理，其中大数定律包括贝努里大数定律、切比雪夫大数定律和辛钦大数定律；中心极限定理包括德莫佛—拉普拉斯定理和林德伯格—列维定理。

§5.1 大数定律

在第一章中，我们已经指出，人们在长期实践中发现，虽然个别随机事件在某次试验中可能出现也可能不出现，但是在大量重复试验中却呈现出明显的规律性，即一个随机事件出现的频率在某个固定数的附近变动，这就是所谓“频率的稳定性”，对于这点，至今为止，我们没有给出理论上的说明.

数学上怎样来描述在一定条件下的大量重复试验呢？我们在前面的内容中已经建立了贝努里试验这一概率模型，并指出它可以作为一定条件下的重复试验的数学模型. 在贝努里试验中，各次试验是相互独立的；在每次试验中，我们所关心的事件 A 出现的概率 $P(A) = p$ 保持不变. 这些特征可以看作是从数学角度把“在一定条件下”、“重复试验”等等用语的涵义加以明确化.

在贝努里试验中，若以 μ_n 记 n 次试验中 A 出现的次数，则 $\frac{\mu_n}{n}$ 便是在这 n 次试验中事件 A 出现的频率，所谓频率的稳定性无非是指当试验次数 n 增大时，频率 $\frac{\mu_n}{n}$ 接近于某个固定的常数.

这个固定的常数就是事件 A 在一次试验中发生的概率. 由此可见，讨论频率 $\frac{\mu_n}{n}$ 的极限行为是理解概率论中最基本的概念—概率所不可缺少的.正是这个缘故，在概率论的发展史上，极限定理的研究一直占有重要地位，而它的发源地就是贝努里试验这个概型.

从前几章的讨论中我们知道，μ_n 是随机变量，它服从二项分布

$$P\{\mu_n = k\} = C_n^k p^k q^{n-k} \qquad k = 0,1,2,\cdots,n$$

其数学期望 $E\mu_n = np$,方差 $D\mu_n = npq$ ，这在一定程度上帮助我们进一步了解了频率 $\frac{\mu_n}{n}$ 的性质，但是我们更需要知道的是，n 很大时，μ_n 或 $\frac{\mu_n}{n}$ 的性质。

显然，当 n 很大时，μ_n 一般也很大，所以直接研究 μ_n 不是很恰当，还是研究频率 $\frac{\mu_n}{n}$ 为宜。因为 $E(\frac{\mu_n}{n}) = p$, $D(\frac{\mu_n}{n}) = \frac{pq}{n}$ ，所以当 $n \to \infty$ 时，频率的数学期望保持不变，而方差则趋于零。我们知道方差为零的随机变量是常数，于是我们自然预期频率将趋于常数 p (事件 A 发生的概率)，但是频率 μ_n/n 是随机变量，关于它的极限又将如何表示呢?

一种提法是，当 n 足够大时，频率 $\frac{\mu_n}{n}$ 与概率 p 有较大偏差的概率很小，用数学语言来讲，就是要证明，对于任给的 $\varepsilon > 0$，

$$\lim_{n\to\infty} P\{|\mu_n / n - p| \geq \varepsilon\} = 0 \tag{5.1}$$

或它的等价式

$$\lim_{n\to\infty} P\{|\mu_n / n - p| < \varepsilon\} = 1$$

成立。

历史上，贝努里第一个研究了这种类型的极限定理. 在 1713 年发表的论文中，他建立了等式(5.1), 这是一大类概率论极限定理中的第一个（也是概率论的第一篇论文）.

定理 1 （贝努里大数定律）设 μ_n 是 n 重贝努里试验中事件 A 出现的次数，又 A 在每次试验中出现的概率为 p ($0 < p < 1$)，则对任意的 $\varepsilon > 0$，有

$$\lim_{n\to\infty} P\{|\mu_n / n - p| < \varepsilon\} = 1 \tag{5.2}$$

证明：令 $\xi_i = \begin{cases} 0 & 在第i次试验中A出现 \\ 1 & 在第i次实验中A不出现 \end{cases} \qquad (1 \leq i \leq n)$

则 $\xi_1, \xi_2, \cdots, \xi_n$ 是 n 个相互独立的随机变量，且

$$E\xi_i = p, D\xi_i = p(1-p) = pq,$$

而

$$\mu_n = \sum_{i=1}^{n}\xi_i,$$

所以

$$\frac{\mu_n}{n} - p = \frac{\mu_n - np}{n} = \frac{\sum_{i=1}^{n}\xi_i - E(\sum_{i=1}^{n}\xi_i)}{n},$$

由切比雪夫不等式

$$P\{|\mu_n/n - p| \geq \varepsilon\} = P\{|\frac{\sum_{i=1}^{n}\xi_i - E(\sum_{i=1}^{n}\xi_i)}{n}| \geq \varepsilon\} = P\{\left|\sum\xi_i - E(\sum_{i=1}^{n}\xi_i)\right| \geq n\varepsilon\}$$

$$\leq \frac{D(\sum_{i=1}^{n}\xi_i)}{n^2\varepsilon^2} \qquad \text{（由独立性 } D(\sum_{i=1}^{n}\xi_i) = \sum_{i=1}^{n}D\xi_i = np \text{）}$$

$$= \frac{npq}{n^2\varepsilon^2} = \frac{1}{n}\frac{pq}{\varepsilon^2} \to 0 \quad (n \to \infty)$$

这个定理以严格的数学形式表达了概率的稳定性。就是说当n很大时，事件发生的频率与概率有较大偏差的可能性很小，由实际推断原理，在实际应用中，当试验次数很大时，便可以用事件的频率来代替事件的概率。

下面给出的是比贝努里大数定律更广泛一些的切比雪夫大数定律。

定理 2 （切比雪夫大数定律）设$\xi_1, \xi_2, \cdots, \xi_n, \ldots$ 是一列两两不相关的随机变量，又设它们的方差有界，即存在常数 C >0 ，使有$D\xi_i \leq C \quad i = 1,2,\cdots$，则对任意的$\varepsilon > 0$，有

$$\lim_{n\to\infty} P\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i - \frac{1}{n}\sum_{i=1}^{n}E\xi_i\right| < \varepsilon\} = 1 \tag{5.3}$$

证明：由切比雪夫不等式，有

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

因为$\{\xi_i\}$两两不相关，且由它们的方差有界即可得到

$$D(\sum_{i=1}^{n}\xi_i)=\sum_{i=1}^{n}D\xi_i \le nC$$

从而有

$$P\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i-\frac{1}{n}\sum_{i=1}^{n}E\xi_i\right|\ge\varepsilon\}\le\frac{C}{n\varepsilon^2}\to 0, n\to\infty$$

可以看出贝努里大数定律是切比雪夫大数定律的特例，在它们的证明中，都是以切比雪夫不等式为基础的，所以要求随机变量具有方差，但是进一步的研究表明，方差存在这个条件并不是必要的，下面我们介绍一个满足独立同分布条件时的辛钦大数定律。

定理 3（辛钦大数定律）设$\xi_1,\xi_2,\cdots,\xi_n,\ldots$是一列独立同分布的随机变量，且数学期望存在，$E\xi_i=a \quad i=1,2,\cdots$，则对任意的$\varepsilon>0$，有

$$\lim_{n\to\infty}P\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i-a\right|<\varepsilon\}=1. \tag{5.4}$$

证明略。

显然，辛钦大数定律也是贝努里大数定律的一种推广。辛钦大数定律表明，当n很大时，随机变量在n次观测中的算术平均值$\frac{1}{n}\sum_{i=1}^{n}\xi_i$会“靠近”它的期望值，这就为寻找随机变量的期望值提供了一条实际可行的途径.

n个随机变量的算术平均值会“靠近”它的期望值，这种“靠近”是在概率意义下的接近。通俗地说，在定理的条件下，n个随机变量的算术平均，当n无限增加时将几乎变成一个常数。

设$\xi_1,\xi_2,\cdots,\xi_n,\cdots$是一列随机变量，$a$是一个常数，若对于任意的$\varepsilon>0$，有

$$\lim_{n\to\infty}P\{\left|\xi_n-a\right|<\varepsilon\}=1$$

则称序列$\xi_1,\xi_2,\cdots,\xi_n,\cdots$依概率收敛到$a$，记作$\lim_{n\to\infty}\xi_n\overset{p}{=}a$，或者$\xi_n\xrightarrow{P}a\ (n\to\infty)$。

贝努里大数定律表明了频率$\frac{\mu_n}{n}$依概率收敛于p，即$\frac{\mu_n}{n}\xrightarrow{P}p \quad (n\to\infty)$

关于依概率收敛有下面的一些性质。

若$\{\xi_n\}$、$\{\mu_n\}$是两个随机变量序列，并且当$n\to\infty$时，有$\xi_n \xrightarrow{P} a$，$\eta_n \xrightarrow{P} b$，其中a和b是常数，则有

（1）$\xi_n+\eta_n \xrightarrow{P} a+b$，$(n\to\infty)$

（2）$\xi_n-\eta_n \xrightarrow{P} a-b$，$(n\to\infty)$

（3）$\xi_n\cdot\eta_n \xrightarrow{P} ab$，$(n\to\infty)$

（4）若$b\neq 0$，$\xi_n/\eta_n \xrightarrow{P} a/b$，$(n\to\infty)$

上述结论的证明是容易的，在此略去。

§5.2　中心极限定理

自从高斯提出测量误差服从正态分布以后，人们发现，正态分布在自然界中极为常见，如炮弹的落点、身高、体重等，观察表明，如果一个量是由大量相互独立的随机因素的影响所造成，而每一个因素在总因素中所起的作用不是很大，则这种量通常都服从或近似服从正态分布.

如果一个量的形成受到众多的随机因素的影响，而其中任一单个因素所起的作用很有限，则这个量的概率分布，必然（近似地）用正曲线去描述，此结果在概率论上叫做“中心极限定理”。概率论中凡是关于“在一定条件下，随机变量之和的分布是正态分布”的定理，统称为中心极限定理。

中心极限定理在概率论和统计学中有极广泛的应用，它揭示了正态分布的源泉。中心极限定理的内容非常丰富，它有多种形式，我们只介绍常用的两种：德莫佛-拉普拉斯（De Moivre-Laplace）中心极限定理和林德伯格-列维（Lindeberg-Levy）中心极限定理。德莫佛-拉普拉斯中心极限定理说明二项分布的极限分布是正态分布，而林德伯格-列维中心极限定理说明“独立随机变量之和或算术平均值”的极限分布是正态分布。

定理 1（德莫佛-拉普拉斯中心极限定理）在n重贝努里试验中，事件A在每次试验中出现的概率为$p(0<p<1)$，μ_n为n次试验中事件A出现的次数，则

$$\lim_{n\to\infty} P\{\frac{\mu_n - np}{\sqrt{npq}} < x\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \qquad (5.5)$$

证明略。

该定理表明，正态分布是二项分布的极限分布。当 n 充分大时，我们可以用下式来计算二项分布的概率

$$\lim_{n\to\infty} P\{a \le \frac{\mu_n - np}{\sqrt{npq}} \le b\} = \frac{1}{\sqrt{2\pi}}\int_{a}^{b} e^{-\frac{t^2}{2}} dt$$

下一定理是定理德莫佛-拉普拉斯中心极限定理的推广，又把它叫做独立同分布的中心极限定理。

定理 2（林德伯格-列维中心极限定理）设$\xi_1,\xi_2,\cdots,\xi_n,\ldots$是一列独立同分布的随机变量，$E\xi_i = a \quad i=1,2,\cdots$，$D\xi_i = \sigma^2$（ $\sigma > 0$ ），则有

$$\lim_{n\to\infty} P\{\frac{\sum_{i=1}^{n}\xi_i - na}{\sigma\sqrt{n}} < x\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \qquad (5.6)$$

证明略。

定理 2 表明，均值为a，方差为σ^2的独立同分布的随机变量$\xi_1,\xi_2,\cdots,\xi_n$之和$\sum_{i=1}^{n}\xi_i$的标准化变量，在n充分大时，近似地服从标准正态分布。

将(5.6)式左端改写，则(5.6)式变为

$$\lim_{n\to\infty} P\{\frac{\frac{1}{n}\sum_{i=1}^{n}\xi_i - a}{\sigma/\sqrt{n}} < x\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \qquad (5.7)$$

(5.7)式表明，均值为a，方差为σ^2的独立同分布的随机变量$\xi_1,\xi_2,\cdots,\xi_n$的算术平均$\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$，在$n$充分大时近似地服从均值为$a$、方差为$\sigma^2/n$的正态分布。这一结果是数理统计中大样本统计推断的基础。

下面举几个关于中心极限定理应用的例子。

例 1 某单位内部有 260 部电话分机，每部分机有 4%的时间要使用外线通话，若各电话分机是否使用外线是相互独立的，问总机要配备多少条外线方可以 95%的把握保证每部分

机在使用外线时不必等候?

解： 令 $X_i=\begin{cases}1, & 第i个分机要使用外线\\ 0 & 第i个分机不使用外线\end{cases}$，$i$=1,2,…,260

$P\{X_k=1\}=0.04, P\{X_k=0\}=0.96$，则任一时刻 260 部分机中要使用外线的分机数为 $Y_n=\sum_{k=1}^{260}X_k$，且服从参数为 260，0.04 的二项分布，由极限定理

$$P\{Y_{260}<x\}=P\left\{\frac{Y_{260}-260\times0.04}{\sqrt{260\times0.04\times0.96}}\le\frac{x-260\times0.04}{\sqrt{260\times0.04\times0.96}}\right\}$$

$$\approx\Phi\left(\frac{x-260\times0.04}{\sqrt{260\times0.04\times0.96}}\right)=\Phi(b)\ge0.95$$

由正态分布表，要使 $\Phi(b)\ge0.95$，必需 b=1.65,

即 $x=1.65\times\sqrt{260\times0.04\times0.96}+260\times0.04\approx15.61$.

取最接近的整数 $x=16$，故总机至少要配备 16 条外线，才能保证有 95%的把握使各个分机在使用外线时不占线.

例 2 设男孩出生率为 0.515，求在 10000 个新生儿中女孩不少于男孩的概率.

解：用 X 表示 10000 个新生儿中男孩的个数，则 $X\sim B(n,p)$，其中，n=10000，p=0.515。

要求女孩数不少于男孩个数的概率,即求 $P\{X\le5000\}$。由德莫佛-拉普拉斯极限定理，有

$$\{X\le5000\}=\left\{\frac{X-np}{\sqrt{npq}}\le\frac{5000-np}{\sqrt{npq}}\right\}$$

令 $Y=\dfrac{X-np}{\sqrt{npq}}$，则 $Y\sim N(0,1)$ 于是有

$$P\{X\le5000\}=P\left\{Y\le\frac{5000-10000\times0.515}{\sqrt{10000\times0.515\times0.485}}\right\}\approx\Phi(-3)=1-\Phi(3)=0.00135$$

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人数相互独立，且服从统一分布。（1）求参加会议的家长数超过

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）以用 $X_k(k=1,2,\cdots,400)$ 表示第 k 个学生来参加会议的家长　　数，则 $X_k(k=1,2,\cdots,400)$ 的分布律为

X_k	0	1	2
p_k	0.05	0.8	0.15

易知 $E(X_k)=1.1, D(X_k)=0.19\ (k=1,2,\cdots,400)$。而 $X=\sum_{k=1}^{400} X_k$，由定理 2

$$P\{X>450\}=P\left\{\frac{X-400\times 1.1}{\sqrt{400}\sqrt{0.19}}>\frac{450-400\times 1.1}{\sqrt{400}\sqrt{0.19}}\right\}$$

$$=1-P\left\{\frac{X-400\times 1.1}{\sqrt{400}\sqrt{0.19}}\le 1.147\right\}$$

$$\approx 1-\Phi(1.147)=0.1257$$

(2)以 Y 记有一名家长来参加会议的学生数，则 $Y\sim b(400,0.8)$，则由定理 1 得

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

$$=P\left\{\frac{Y-400\times 0.8}{\sqrt{400\times 0.8\times 0.2}}\le 2.5\right\}$$

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