大数定律和中心极限定理

李孟棠，朱彬

中山大学智能工程学院

limt29@mail.sysu.edu.cn, zhub26@mail.sysu.edu.cn

2021 年 x 月 x 日

主要内容

大数定律 (law of large numbers)：描述随机变量序列的算术均值在某些条件下收敛到总体均值，数学表达类似

$$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{\text{某种意义}} \mu, \quad n \to \infty;$$

中心极限定理 (central limit theorem)：在一些条件之下，随机变量序列的算术均值近似满足正态分布，是大数定律某种意义上的强化版本，数学表达类似

$$\frac{1}{n}\sum_{k=1}^{n} X_k \overset{\text{近似}}{\sim} N(\mu, \sigma^2/n), \quad n \to \infty;$$

泛指一类定理。

主要内容

定理 (弱大数定律，又称 Khinchin 大数定律)

设随机变量序列 $X_1, X_2, \ldots$ 相互独立[a]且服从同一分布 *(independent and identically distributed, i.i.d.)*，且它们的数学期望满足 $E(X_k) = \mu$ $(k = 1, 2, \ldots)$。令 $\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ 为前 n 个随机变量的算术均值，则对于任意的 $\varepsilon > 0$，有

$$\lim_{n\to\infty} P\left\{\left|\bar{X}_n - \mu\right| < \varepsilon\right\} = 1.$$

[a]这里指对于任意整数 $n > 1$，随机变量 $X_1, X_2, \ldots, X_n$ 相互独立。

解读：$\bar{X}_n$ 是一个随机变量，$\left\{\left|\bar{X}_n - \mu\right| < \varepsilon\right\}$ 是一个事件。定理表明：对于任意 $\varepsilon > 0$，当 n 充分大时，不等式 $\left|\bar{X}_n - \mu\right| < \varepsilon$ 成立的概率接近 1，也就是算术均值 $\bar{X}_n$ 以大概率接近总体均值 μ。

定义（依概率收敛，convergence in probability）

设 $Y_1, Y_2, \ldots, Y_n, \ldots$ 为一个随机变量序列，Y 以一随机变量。如果对于任意正数 ε，有

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

或等价地，

$$\lim_{n\to\infty} P\{|Y_n - Y| \geq \varepsilon\} = 0,$$

则称序列 Y_n 依概率收敛于 Y，也记为 $Y_n \xrightarrow{P} Y$ 或 $Y_n \xrightarrow{p} Y$。

用 $\varepsilon - \delta$ 语言：令 $P_n(\varepsilon)$ 表示 Y_n 位于以 Y 为中心、ε 为半径的球之外的概率。那么 Y_n 依概率收敛于 Y 的含义为：对于任意的 $\varepsilon > 0$ 和 $\delta > 0$，存在整数 N（依赖于 ε 和 δ）使得对于任意的 $n \geq N$，有 $P_n(\varepsilon) < \delta$（数列极限的定义）。

依概率收敛的序列有以下性质：
（连续映射定理，continuous mapping theorem, Mann-Wald 1943）设 $X_n \xrightarrow{P} X$，那么对于连续函数 $g: \mathbb{R} \to \mathbb{R}$，有 $g(X_n) \xrightarrow{P} g(X)$。

回忆连续函数的 Heine 定义：一个连续函数把收敛序列映射到收敛序列，即如果 $x_n \to x$，那么 $g(x_n) \to g(x)$。

Khinchin 定理可以被表述为：对于期望为 μ 的 i.i.d. 随机变量序列 $\{X_k\}$，均值序列 $\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ 在 $n \to \infty$ 时依概率收敛到常数 μ。

例：考虑如下身高估计实验：首先在街上随机选择一个人，令 X 表示他（她）的身高。那么在采取测量措施之前，X 可以被视为随机变量。然后请别人去目测被选中人的身高。令 X_k 表示第 k 个人的估计，假设每个 X_k 都是和 X 同分布的随机变量（身高估计没有系统误差）且序列 $\{X_k\}_{k=1}^{n}$ 相互独立（称为随机变量 X 的 i.i.d. 样本）。那么根据大数定律，均值序列 $\bar{X}_n$ 将依概率收敛到 $E(X)$。

Khinchin 大数定律的证明.

假设随机变量序列的方差存在，即 $D(X_k) = \sigma^2 \ (k = 1, 2, \dots)$。一般情形的证明更为复杂，此处从略。

易知

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

又由独立性得

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

由切比雪夫不等式得

$$1 \geq P\left\{\left|\bar{X}_n - \mu\right| < \varepsilon\right\} \geq 1 - \frac{\sigma^2/n}{\varepsilon^2}.$$

在上式中令 $n \to \infty$ 即得到定理结论。 □

定理 (伯努利大数定律)

设 n_A 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。那么对于任意 $\varepsilon > 0$，有

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

或等价地，

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

证明.

因为 $n_A \sim b(n, p)$，由第四章例子知，

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

其中，$X_1, \ldots, X_n$ 相互独立且都服从以 p 为参数的 $(0-1)$ 分布，所以 $E(X_k) = p$ $(k = 1, \ldots, n)$。于是定理的结论由 Khinchin 大数定律直接导出。

□

伯努利大数定律的解读：概率的相对频率定义

对于任意 $\varepsilon > 0$，当独立重复试验的次数 n 充分大时，$\left\{\left|\frac{n_A}{n} - p\right| < \varepsilon\right\}$ 事件几乎必然发生。在这个意义下，我们说事件发生的相对频率趋向于该事件的概率。

定理 (强大数定律，又称 Kolmogorov 大数定律)

给定 *i.i.d.* 随机变量序列 $X_1, X_2, \ldots$，设它们的数学期望 $E(X_k) = \mu < \infty$ $(k = 1, 2, \ldots)$。令 $\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$，则有

$$P\left\{\lim_{n\to\infty} \bar{X}_n = \mu\right\} = 1.$$

证明比弱大数定律复杂，此处从略。

解读：随机变量 $\bar{X}_n : S \to \mathbb{R}$ 是一个函数，给定样本点 $e \in S$，$\{\bar{X}_n(e)\}_{n\geq 1}$ 是一个实数序列。定理表明：考察所有使得数列 $\bar{X}_n(e) \to \mu$ 的样本点 e 构成的事件，其概率为 1。

Definition (几乎肯定收敛，almost sure convergence)

随机变量序列 $\{X_n\}_{n\geq 1}$ 几乎肯定 (或几乎处处、以概率 1、强) 收敛到 X 的含义为

$$P\left\{\lim_{n\to\infty} X_n = X\right\} = 1,$$

也常记作 $X_n \xrightarrow{\text{a.s.}} X$。

定义表明随机变量 X_n 的值不收敛到 X 的值所构成的事件拥有的概率为 0。使用概率空间 $(S, \boldsymbol{F}, P)$ 和随机变量（从样本空间到实数轴的函数）的概念，上述定义等价于

$$P\left\{e \in S : \lim_{n\to\infty} X_n(e) = X(e)\right\} = 1.$$

例：考察某种短寿物种（例如蜉蝣）中的一个个体，记录它每天的进食数量。该记录构成一个不可预测 (unpredictable) 的数列，但我们**十分肯定**：在某一天该记录会归零，并且在那之后永远保持为零。

几乎肯定收敛强于依概率收敛：$X_n \xrightarrow{\text{a.s.}} X \implies X_n \xrightarrow{P} X$.

强大数定律表明：在 i.i.d. 假设下，$\bar{X}_n \xrightarrow{\text{a.s.}} \mu$。

在相同的假设下，强、弱大数定律都表明样本均值 $\bar{X}_n$ 收敛到总体均值 μ，区别在于收敛的方式 (modes of convergence) 不同。具体地，

- **弱定律**指出，指定一个充分大的 n，$\bar{X}_n$ 很可能接近 μ。但是事件 $|\bar{X}_n - \mu| > \varepsilon$ 仍有可能发生无穷多次，尽管发生的频率不高。

- **强定律**表明上述情形几乎肯定不会发生。特别地，强定律指出以下事件以概率 1 发生：对于任意的 $\varepsilon > 0$，不等式 $|\bar{X}_n - \mu| < \varepsilon$ 对于所有充分大的 n 都成立。

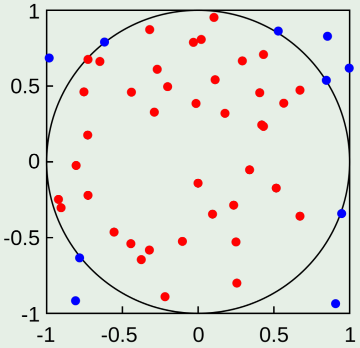
- 在一些期望 μ 不存在的情形（例如非 Lebesgue 绝对可积），强定律不再成立，但弱定律仍然适用。

利用随机采样的方法计算多维（即多变量）数值积分。

例 (估计 π 的数值)

如图所示，在 2×2 的正方形区域上随机（均匀）采样，样本点落入内接圆的概率为 $\pi/4$。由大数定律，我们知道后者约等于采样点落入圆域的相对频率，这里为 $40/50 = 0.8$。因此

$$\frac{\pi}{4} \approx 0.8 \implies \pi \approx 4 \times 0.8 = 3.2.$$



问题：计算多维定积分

$$I = \int_\Omega f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

其中 Ω 为 $\mathbb{R}^m$ 的子集，其体积为

$$V = \int_\Omega \mathrm{d}\boldsymbol{x}.$$

解：物理中计算物体体积中心 (centroid of the volume，即密度为常数时的质心) 的公式可以被解读为数学期望的计算

$$\frac{1}{V}\int_\Omega \boldsymbol{x}\mathrm{d}\boldsymbol{x} = \mathbb{E}(\boldsymbol{X}),$$

其中随机变量 $\boldsymbol{X}$ 服从在区域 Ω 上的均匀分布，即在 Ω 上的概率密度为常数 $1/V$，其余地方为零。类似地，

$$\frac{I}{V} = \frac{1}{V}\int_\Omega f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \mathbb{E}(f(\boldsymbol{X})).$$

朴素 (naive) Monte-Carlo 方法：在 Ω 上进行（独立）均匀采样，得到样本点

$$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \Omega.$$

再利用样本均值近似总体均值，即

$$\frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{x}_i) \approx \mathbb{E}(f(\boldsymbol{X})) = \frac{I}{V}.$$

所以待求积分

$$I \approx \frac{V}{N}\sum_{i=1}^{N} f(\boldsymbol{x}_i) := Q_N.$$

由样本的独立性可得估计误差（一个随机数）

$$\delta Q_N \approx \sqrt{\mathrm{Var}(Q_N)} = \frac{V\sigma}{\sqrt{N}}$$

以速率 $\frac{1}{\sqrt{N}}$ 衰减，其中 $\sigma^2 := \mathrm{Var}(f(\boldsymbol{X}_i))$。该近似误差不依赖于积分域的维数。

主要内容

定理 (经典 CLT，又称 Lindeberg-Lévy CLT)

给定 *i.i.d.* 随机变量序列 $\{X_k\}_{k\geq 1}$，设它们具有有限的数学期望和方差为：$E(X_k)=\mu$，$D(X_k)=\sigma^2>0$。那么均值序列 $\bar{X}_n=\frac{1}{n}\sum_{k=1}^n X_k$ 近似服从正态分布 $N(\mu,\sigma^2/n)$，或等价地，其标准化变量

$$\frac{\bar{X}_n-E(\bar{X}_n)}{[D(\bar{X}_n)]^{1/2}}=\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}$$

的分布函数 $F_n(x)$ 对于任意 x 满足

$$\begin{aligned}\lim_{n\to\infty}F_n(x)&=\lim_{n\to\infty}P\left\{\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}\leq x\right\}\\&=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}\mathrm{d}t=\Phi(x),\end{aligned}$$

也就是

$$\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}\overset{\text{近似}}{\sim}N(0,1).$$

证明略。定理解读：一般而言，均值随机变量 $\bar{X}_n$ 的分布函数较难求得。但是，上述中心极限定理表明，当 n 充分大时，$\bar{X}_n$ 近似为正态分布。因此，我们可以利用 $\Phi(x)$ 的信息对 $\bar{X}_n$ 做理论分析或计算。这也是数理统计中**大样本**统计推断的基础。

Definition (依分布收敛)

设随机变量 X 和随机变量序列 $\{X_n\}_{n\geq 1}$ 的分布函数分别为 $F(x)$ 和 $\{F_n(x)\}_{n\geq 1}$。如果对于任意 $F(x)$ 的连续点 $x \in \mathbb{R}$ 都有

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

那么称 X_n 依分布（或弱）收敛到 X，通常记作 $X_n \xrightarrow{d} X$ 或 $X_n \xrightarrow{\mathcal{D}} X$。

例（骰子工厂）：一个骰子工厂刚落成。由于生产过程的缺陷，前几颗产出的骰子不够均匀。投掷它们中任何一颗的结果所服从的分布将明显不同于理想的均匀分布。随着生产水平的提高，工厂产出的骰子将变得越来越均匀。投掷这样一颗新骰子的结果将越来越接近均匀分布。

依分布收敛是随机变量最弱的收敛方式之一，弱于依概率收敛：

$$X_n \xrightarrow{\text{a.s.}} X \implies X_n \xrightarrow{P} X \implies X_n \xrightarrow{d} X.$$

但在实际中应用非常广泛。

经典 CLT 的等价陈述：给定 i.i.d. 随机变量序列 $X_1, X_2, \ldots$，那么当 n 趋向于无穷大时，随机变量 $\sqrt{n}(\bar{X}_n - \mu)$ 依分布收敛到正态随机变量 $N(0, \sigma^2)$，即

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} N\left(0, \sigma^2\right).$$

定理 (Lyapunov CLT，不要求同分布)

设随机变量序列 $\{X_k\}_{k\geq 1}$ 相互独立且具有数学期望和方差

$$E(X_k) = \mu_k, \quad D(X_k) = \sigma_k^2 > 0, \quad k = 1, 2, \ldots.$$

记 $B_n^2 = \sum_{k=1}^n \sigma_k^2$。若存在正数 δ 使得当 $n \to \infty$ 时，

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

则随机变量之和 $\sum_{k=1}^n X_k$ 的标准化变量

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

的分布函数 $F_n(x)$ 对于任意 x 满足

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} P\{Z_n \leq x\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \mathrm{d}t = \Phi(x).$$

定理 (De Moivre-Laplace CLT，离散随机变量)

设随机变量 η_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

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

证明.

因为 $\eta_n \sim b(n,p)$，由第四章例子知 $\eta_n = \sum_{k=1}^{n} X_k$，其中，$X_1,\ldots,X_n$ 相互独立且都服从以 p 为参数的 $(0-1)$ 分布，所以 $E(X_k)=p$，$D(X_k)=p(1-p)$ $(k=1,\ldots,n)$。由 i.i.d. 随机变量序列 CLT 可直接导出本定理结论。 □

定理解读：正态分布是二项分布的极限分布。当 n 充分大时，可以利用(1)来计算二项分布的概率。

思考：第二章讲过泊松分布也是某种意义下二项分布的极限，和上述定理中的情况有何不同？

例：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，且都在区间 $[0,10]$ 上服从均匀分布。记 $V=\sum_{k=1}^{20} V_k$，求 $P\{V>105\}$ 的近似值。

解：易知 $\mu := E(V_k)=5$，$\sigma^2 := D(V_k)=100/12$。记 $n=20$，由 i.i.d. 随机变量序列 CLT，随机变量

$$Z_n=\frac{V/n-5}{\sigma/\sqrt{n}}$$

近似服从正态分布 $N(0,1)$，于是

$$\begin{aligned}
P\{V>105\} &= P\left\{\frac{V/n-5}{\sigma/\sqrt{n}}>\frac{105/n-5}{\sigma/\sqrt{n}}\right\}\\
&= 1-P\left\{\frac{V/n-5}{\sigma/\sqrt{n}}\le 0.387\right\}\\
&\approx 1-\Phi(0.387)=0.348.
\end{aligned}$$

例：一船舶在某海区航行，已知每遭受一次波浪的冲击，纵摇角大于 3° 的概率为 $p=1/3$。若船舶遭受了 90000 次海浪冲击，问其中有 29500 至 30500 次纵摇角大于 3° 的概率是多少？

解：将船舶每遭受一次波浪冲击看作一次试验，并假设各次试验相互独立。在 90000 次试验中纵摇角大于 3° 的次数记为 X，则有 $X\sim b(90000,1/3)$。所求概率即为

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

下面用 De Moivre-Laplace CLT 对上式近似求解：

$$\begin{aligned}P(29500\le X\le 30500)&=P\left\{\frac{29500-np}{\sqrt{np(1-p)}}\le\frac{X-np}{\sqrt{np(1-p)}}\le\frac{30500-np}{\sqrt{np(1-p)}}\right\}\\&\approx\Phi(5\sqrt{2}/2)-\Phi(-5\sqrt{2}/2)=0.9995.\end{aligned}$$

例：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、有一名家长、有两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满足 i.i.d. 条件。问：(1) 参加会议的家长人数 X 超过 450 的概率；(2) 有一名家长来参会的学生人数不多于 340 的概率。

解：(1) 以 X_k $(k = 1, \ldots, 400)$ 表示第 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

易知 $\mu := E(X_k) = 1.1$，$\sigma^2 := D(X_k) = 0.19$。用 $n = 400$ 表示学生总数，家长总数 $X = \sum_{k=1}^{n} X_k$。由 i.i.d. 随机变量序列 CLT 知：

$$P\{X > 450\} = P\{\bar{X}_n > 450/n\} = P\left\{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} > \frac{450/n - \mu}{\sigma/\sqrt{n}}\right\}$$
$$\approx 1 - \Phi(1.147) = 0.1251.$$

(2) 用 Y 表示有一名家长参加会议的学生人数，则 $Y \sim b(n, p)$，其中 $p = 0.8$。由 De Moivre-Laplace CLT，

$$P\{Y \leq 340\} = P\left\{\frac{Y - np}{\sqrt{np(1-p)}} \leq \frac{340 - np}{\sqrt{np(1-p)}}\right\} \approx \Phi(2.5) = 0.9938.$$

The End