第五节 · 大样本理论

司继春

上海对外经贸大学统计与信息学院

大样本理论（Large sample theory）在现代统计理论中占据着核心位置。一般而言，大样本理论告诉我们当样本容量趋向于无穷时统计量的分布特征，而当数据足够多时，统计量的分布特征与这个极限特征仅有很小的误差。由于有限样本的统计性质通常难以计算，因而多数时候我们需要使用大样本理论对有限样本进行近似计算。下面我们从最简单的数列极限开始入手，进而介绍常用的大样本理论。

1 收敛的概念

首先我们先回顾一下数列极限的概念：

定义 1. 若 $\{a_n, n = 1, 2, ...\}$ 为实数序列，如果对于任意的 $\epsilon > 0$，存在 $n_0 = n_0(\epsilon)$ 使得：

$$|a_n - a| < \epsilon, \forall n > n_0$$

那么我们称数列 $\{a_n\}$ 的极限为 a，或者 $\{a_n\}$ **收敛到**（**converges to**）a，记为

$$\lim_{n\to\infty} a_n = a$$

或者 $a_n \to a \; as \; n \to \infty$。

数列极限有一些简单的性质，比如，如果 $a_n \to a, b_n \to b$，那么 $(c \cdot a_n + d \cdot b_n) \to ca + db$，$a_n \cdot b_n \to ab$，如果 $b \neq 0$，那么 $\frac{a_n}{b_n} \to \frac{a}{b}$。

如一个经常使用的数列极限：

$$\lim_{n\to\infty} \left(1 + \frac{c}{n}\right)^n = e^c$$

而：

$$\lim_{n\to\infty} \frac{\left(1 + \frac{c}{n}\right)^n}{\left(1 + \frac{d}{n}\right)^n} = e^{c-d}$$

此外，数列可能随着 $n \to \infty$ 时，并不趋向于一个常数，而是趋向于 ∞。更严谨的，我们可以定义 $a_n \to \infty$ 如下：

表 1: $\frac{1}{n}$ 和 $\frac{1}{n^2}$ 的收敛速度

n	1	2	5	10	100	1000	10000
$\frac{1}{n}$	1	0.5	0.2	0.1	1×10^{-2}	1×10^{-3}	1×10^{-4}
$\frac{1}{n^2}$	1	0.25	0.04	0.01	1×10^{-4}	1×10^{-6}	1×10^{-8}

定义 2. 若 $\{a_n, n=1,2,...\}$ 为实数序列，如果对于任意的 M，存在一个 $n_0 = n_0(M)$，使得：

$$a_n > M, \forall n > n_0$$

那么我们称 $\{a_n\}$ 趋向于 ∞，记为

$$\lim_{n\to\infty} a_n = \infty$$

或者 $a_n \to \infty \; as \; n \to \infty$。

如经常遇到的数列：$\ln n, n^k, e^n$ 等都趋向于正无穷。相反，如果数列不趋向于无穷，那么我们称其为有界的。其定义如下：

定义 3. 若 $\{a_n, n=1,2,...\}$ 为实数序列，如果存在常数 $b<\infty$，使得 $|a_n|<b$，那么我们称数列 $\{a_n\}$ 为**有界**的（**bounded**），否则称之为**无界**的（**unbounded**）。

显然，如果一个数列收敛到一个常数，那么其必然是有界的，而一个趋向于 ∞ 的数列必定是无界的。而反过来则不成立，一个有界的数列并不一定是收敛的，例如 $a_n = (-1)^n$，虽然是有界的，但其极限并不存在。同时，无界的序列也不一定趋向于 ∞，例如 $a_n = n\cdot[1+(-1)^n]$，当 n 为奇数时 $a_n = 0$，因而这个数列并不趋向于 ∞。

对于两个序列 $\{a_n\}, \{b_n\}$，我们经常会关心两个序列收敛的速度问题。比如，如果令 $a_n = \frac{1}{n^2}, b_n = \frac{1}{n}$，我们有 $a_n \to 0, b_n \to 0$，然而两个序列收敛到 0 的速度是不一样的。表 (1) 列出了随着 n 的增大，两个序列趋向于 0 的速度，可以看到 $\frac{1}{n^2}$ 比 $\frac{1}{n}$ 以更快的速度趋向于 0。

一般的，为了比较两个序列收敛速度的问题，我们做如下定义：

定义 4. 对于两个序列 $\{a_n\}, \{b_n\}$，如果随着 $n\to\infty$，有：

$$\frac{a_n}{b_n} \to 0$$

那么我们记为 $a_n = o(b_n)$。特别的，如果令 $b_n = 1$，那么 $a_n = o(1)$ 等价为 $a_n \to 0$。

如在上例中，

$$\frac{a_n}{b_n} = \frac{\frac{1}{n^2}}{\frac{1}{n}} = \frac{1}{n} \to 0$$

因而 $\frac{1}{n^2} = o\left(\frac{1}{n}\right)$，即 $\frac{1}{n^2}$ 以更快的速度收敛到 0。如果两个序列 $a_n \to 0, b_n \to 0$，且 $a_n = o(b_n)$，那么我们称 a_n 为比 b_n 高阶的无穷小量。

假设有两个数列，

$$a_n = \frac{1}{n} + \frac{6}{n^2} - \frac{8}{n^3}$$

而另外一个序列：

$$b_n = \frac{1}{n}$$

如果定义 $R_n = \frac{6}{n^2} - \frac{8}{n^3}$，显然 $R_n = o\left(\frac{1}{n}\right)$，因而 $a_n = b_n + o\left(\frac{1}{n}\right)$，即：

$$\frac{a_n}{b_n} = \frac{b_n + o\left(\frac{1}{n}\right)}{b_n} \to 1$$

因而尽管两个序列 a_n 和 b_n 并不相等，但是当 $n \to \infty$ 时，两者误差趋向于 0，因而我们可以舍去无穷小量 R_n，使用更简单的序列 b_n 去逼近 a_n。

一般的，如果两个序列 a_n, b_n 随着 $n \to \infty$ 满足：

$$\frac{a_n}{b_n} \to 1$$

那么我们称这两个序列是**渐进等价**（**asymptotically equivalent**）的，记为 $a_n \sim b_n$。渐进意味着随着 $n \to \infty$，而等价意味着两个序列的误差很小。实际上，如果我们把相对误差写为：

$$\left|\frac{a_n - b_n}{b_n}\right| = \left|\frac{a_n}{b_n} - 1\right|$$

那么渐进等价意味着随着 n 的增大，相对误差趋向于 0。实际上，如果 $a_n = o(b_n)$，那么 $b_n + a_n = b_n + o(b_n) \sim b_n$，即一个序列加上这个序列的无穷小量，渐进等价于这个序列本身。

当然，由于 $a_n = \frac{1}{n} + \frac{6}{n^2} + o\left(\frac{1}{n^2}\right)$，我们也可以使用 $\frac{1}{n} + \frac{6}{n^2}$ 作为 a_n 的更加精确的逼近。

这种逼近比较常用的即**泰勒级数**（**Taylor series**）。当 $x \to a$ 时，$(x-a) = o(1)$，同时我们有 $(x-a)^{k+1} = o\left((x-a)^k\right)$，即当 $x \to a$ 时，$(x-a)$ 的高阶幂是低阶幂的无穷小量。对于一个单变量实值函数 $f(x)$ 且 k 阶可微，那么有：

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots \frac{f^{(k)}(a)}{k!}(x-a)^k + o\left(|x-a|^k\right)$$

因而对于一个难以计算的函数 f，我们经常使用其前 k 阶泰勒多项式对其进行逼近。

例 1. 函数 $f(x) = \ln(1+x)$ 在 $x = 0$ 处的泰勒展开为：

$$f(x) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n} = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots$$

因而当 x 充分靠近 0 时，我们可以使用前 k 阶泰勒展开对其进行逼近。特别的，

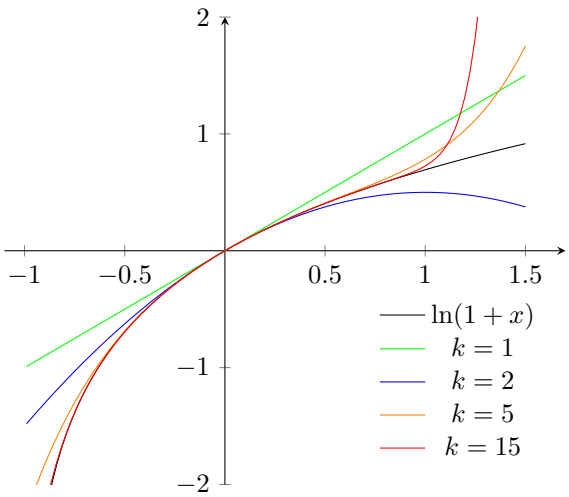

图 1: $\ln(1+x)$ 的泰勒展开

如果令 $k=1$，$\ln(1+x)=x+o(x)\approx x$。图 (1) 展示了使用不同阶数的多项式逼近 $\ln(1+x)$ 的结果。

值得注意的是，在图 (1) 中，只有当 x 在 $(-1,1)$ 区间之内，随着 k 的增加多项式逐渐逼近 $\ln(1+x)$。注意我们使用泰勒级数进行逼近的前提条件是 x 充分的接近于 a，因而如果 x 不在泰勒级数的收敛半径之内，泰勒级数不能用于逼近原始函数。

更一般的，如果 $f:\mathbb{R}^n\to\mathbb{R}$ 为多元实值函数，那么其泰勒级数为:

$$f(x)=f(a)+\frac{\partial f}{\partial x'}(a)(x-a)+\frac{1}{2!}(x-a)'\frac{\partial^2 f}{\partial x\partial x'}(a)(x-a)+o\left(\|x-a\|^2\right)$$

其中 x 和 a 为 $n\times 1$ 向量。

例 2. 令 $f(x)=e^{x_1}\ln(1+x_2)$，其中 $x=(x_1,x_2)'$。那么:

$$\frac{\partial f}{\partial x}=\begin{bmatrix} e^{x_1}\ln(1+x_2) \\ \frac{e^{x_1}}{1+x_2}\end{bmatrix},\frac{\partial^2 f}{\partial x\partial x'}=\begin{bmatrix} e^{x_1}\ln(1+x_2) & \frac{e^{x_1}}{1+x_2} \\ \frac{e^{x_1}}{1+x_2} & -\frac{e^{x_1}}{(1+x_2)^2}\end{bmatrix}$$

那么其在 $a=(0,0)'$ 处的二阶泰勒展开:

$$\begin{aligned} p_2(x)&=[0,1]\begin{bmatrix}x_1\\x_2\end{bmatrix}+\frac{1}{2}[x_1,x_2]\begin{bmatrix}0&1\\1&-1\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix} \\ &=x_2+\frac{1}{2}\left(2x_1x_2-x_2^2\right)\end{aligned}$$

练习 1. 写程序近似逼近正态分布的分布函数并画图。

如果令 $a_n = n^\alpha$，$b_n = n^\beta$，其中 α, β 为常数，那么：

$$\frac{a_n}{b_n} = \frac{n^\alpha}{n^\beta} = \frac{1}{n^{\beta-\alpha}}$$

当 $\beta > \alpha$ 时，以上极限趋向于 0，即 $a_n = o(b_n)$，a_n 相比于 b_n 以更慢的速度趋向于 ∞。

例 3. （指数增长）对于任意的 $k > 0$ 以及 $a > 1$，有 $n^k = o(a^n)$。为了证明，取 $c = a - 1 > 0$。考虑 $k = 2$ 的情形，欲证明

$$\frac{n^k}{a^n} \to 0$$

等价于证明

$$\frac{a^n}{n^k} \to \infty$$

由于：

$$a^n = (1+c)^n = 1 + nc + \binom{n}{2} c^2 + \binom{n}{3} c^3 + \cdots > \binom{n}{3} c^3$$

因而

$$\frac{a^n}{n^2} > \frac{\binom{n}{3} c^3}{n^2} = \frac{c^3}{6}\left(n - 3 + \frac{2}{n}\right) \to \infty$$

对于其他整数 k 可以类似证明。如果 k 不为整数，则取 k' 为比 k 大的最小整数进行证明。

类似的，还可以证明 $n^k = o(\log n)$。这三个序列，$a^n, n^k, \log n$ 是经常使用的三种增长速度，三者趋向于无穷的增长速度逐渐递减，分别代表着快速、适当和慢速的增长。相应的，其倒数，$a^{-n}, n^{-k}, 1/\log n$ 趋向于 0 的速度逐渐递减。

练习 2. 请确定 $a_n = \sqrt{\log n}$ 与 $b_n = \log(\sqrt{n})$ 的阶。

与无穷小的逼近类似，如果 $a_n = a^n + n^k$，$b_n = a^n$，那么 $a_n = b_n + o(b_n)$，因而我们同样可以使用 b_n 对 a_n 进行近似逼近。

练习 3. 请找出一项表达式使其与如下序列渐进等价：

1. $\ln n + \frac{1}{2}n$
2. $\ln n + \ln(\ln n)$
3. $n^2 + e^n$

对于小 o 符号，有如下性质：

定理 1. （小 o 的性质）

1. 若 $a_n = o(b_n), b_n = o(c_n)$，那么 $a_n = o(c_n)$
2. 对于任意的常数 $c \neq 0$，及 $a_n = o(b_n)$，有 $ca_n = o(b_n)$
3. 对于任意的数列 $c_n \neq 0$，及 $a_n = o(b_n)$，有 $c_n a_n = o(c_n b_n)$
4. 如果 $d_n = o(b_n), e_n = o(c_n)$，那么 $d_n e_n = o(b_n c_n)$
5. 如果 $a_n, b_n > 0, c_n, d_n > 0$，$a_n = o(b_n), c_n = o(d_n)$，那么 $a_n + c_n = o(b_n + d_n)$。

练习 4. 请问如下命题是否成立？若成立，请给出证明，若不成立，请给出反例：

1. $a_n = o(b_n), c_n = o(b_n)$，那么 $a_n + c_n = o(b_n)$。
2. $a_n = o(b_n), c_n = o(d_n)$，那么 $a_n + c_n = o(b_n + d_n)$。

与小 o 符号相对应，我们还可以定义大 O 符号：

定义 5. 对于两个序列 $\{a_n\}, \{b_n\}$，如果随着 $n \to \infty$，$\left|\frac{a_n}{b_n}\right|$ 是有界的，即存在一个 M 使得：

$$\left|\frac{a_n}{b_n}\right| < M$$

那么我们记为 $a_n = O(b_n)$。特别的，如果令 $b_n = 1$，那么 $a_n = O(1)$ 等价为 a_n 是有界的。

根据以上定义，如果 $a_n = o(b_n)$，那么必然有 $a_n = O(b_n)$。进而，我们可以使用大 O 符号定义序列同阶。

定义 6. 对于两个序列 $\{a_n\}, \{b_n\}$，如果 $a_n = O(b_n)$，且同时 $b_n = O(a_n)$，那么我们称两个序列是同阶的，简记为 $a_n \asymp b_n$。

下面的例子展示了 $\sim, \asymp, o, O$ 的区别：

例 4. 对于序列 $a_n = \frac{1}{n} + \frac{b}{n\sqrt{n}} + \frac{c}{n^2} + \frac{d}{n^2\sqrt{n}}$，同时定义 $R_n = \frac{b}{n\sqrt{n}} + \frac{c}{n^2} + \frac{d}{n^2\sqrt{n}}$ 那么：

1. $a_n \sim \frac{1}{n}$
2. 若 $b = 0$，$R_n = O\left(\frac{1}{n^2}\right)$
3. 若 $b = 0$，$R_n \asymp \frac{1}{n^2}$
4. 若 $b \neq 0$，$R_n \sim \frac{b}{n\sqrt{n}}$
5. 若 $b = c = 0$，$R_n = o\left(\frac{1}{n^2}\right)$

大 O 符号有如下性质：

定理 2. （大 O 的性质）

1. 若 $a_n = O(b_n), b_n = O(c_n)$，那么 $a_n = O(c_n)$
2. 对于任意的常数 $c \neq 0$，及 $a_n = O(b_n)$，有 $ca_n = O(b_n)$
3. 对于任意的数列 $c_n \neq 0$，及 $a_n = O(b_n)$，有 $c_n a_n = O(c_n b_n)$
4. 如果 $d_n = O(b_n), e_n = O(c_n)$，那么 $d_n e_n = O(b_n c_n)$
5. 如果 $a_n = o(b_n)$，$c_n = O(b_n)$，那么 $a_n c_n = o(b_n)$
6. 如果 $a_n = o(b_n)$，$c_n = O(b_n)$，那么 $a_n + c_n = O(b_n)$

例如，如果 $a_n = \frac{1}{n} + \frac{b}{n\sqrt{n}} + \frac{c}{n^2} + \frac{d}{n^2\sqrt{n}} = O\left(\frac{1}{n}\right)$，那么根据性质 (3)，$n \cdot a_n = \frac{b}{\sqrt{n}} + \frac{c}{n} + \frac{d}{n\sqrt{n}} = O\left(n \cdot \frac{1}{n}\right) = O(1)$。

例 5. 如果令 $a_n = c \cdot (nh)^{-1}, b_n = g \cdot h^4$，其中 c, g 为非零常数，$h = n^q, q < 0$，那么 $a_n = O\left((nh)^{-1}\right) = O\left(n^{-q-1}\right)$，$b_n = O\left(h^4\right) = O\left(n^{4q}\right)$，所以 $a_n + b_n = O\left(n^{-q-1} + n^{4q}\right)$。例如，当 $q = -\frac{1}{2}$ 时，$a_n + b_n = O\left(n^{-\frac{1}{2}} + n^{-2}\right)$，由于 $n^{-2} = o\left(n^{-\frac{1}{2}}\right)$，因而 $a_n + b_n = O\left(n^{-\frac{1}{2}}\right)$。类似的，当 $\frac{n^{-q-1}}{n^{4q}} = n^{-1-5q} \to 0$，即 $-1 - 5q < 0$ 时，$n^{-q-1} = o\left(n^{4q}\right)$，$a_n + b_n = O\left(n^{4q}\right)$；当 $-1 - 5q > 0$ 时，$n^{4q} = o\left(n^{-q-1}\right)$，$a_n + b_n = O\left(n^{-q-1}\right)$；当 $q = -\frac{1}{5}$ 时，$a_n + b_n = O\left(n^{-\frac{4}{5}}\right)$。因而可以证明，当 $q = -\frac{1}{5}$ 时，使得 $a_n + b_n$ 以最快的速度趋向于 0。

练习 5. 如果 $h = n^q, -1 < q < 0$，$a_n = \frac{1}{n^2 h^2} + \frac{10}{n^3 h}$，$b_n = 3h^3 + 10h^4$，求 q 使得 $a_n + b_n$ 以最快的速度趋向于 0。

2 概率收敛的概念

上面讨论了数列收敛的概念。在概率统计中，最经常使用的工具是随机变量，因而在这里我们将讨论随机变量的收敛。在这里我们将主要介绍四种收敛的概念：**几乎必然收敛**（**almost sure convergence**）、**依概率收敛**（**convergence in probability**）、**均方收敛**（**convergence in mean square** or **convergence in quadratic mean**）以及**依分布收敛**（**convergence in distribution** or **convergence in law**）。

2.1 几乎必然收敛

假设 $\{X_n\}$ 为在概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的一系列的随机变量。由于随机变量 X_i 是定义在样本空间 Ω 上的函数，因而定义随机变量收敛的一个最简单的想法是对于每一个 $\omega \in \Omega$，都有 $X_n(\omega) \to X(\omega)$，那么我们可以称随机变量序列 $\{X_n\}$ 收敛到随机变量 X。

然而我们不必要求如此严格，我们可以要求在某一些点 $\omega \in \Omega$ 处，$X_n(\omega) \nrightarrow X(\omega)$，只要这样的点「不多」就可以了。更进一步，我们可以使用概率来描述「不多」这一概念，这样就催生了第一个收敛的定义：

定义 7. （几乎必然收敛）如果概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的一系列随机变量 $\{X_n\}$ 满足：

$$\mathscr{P}\left(\left\{\lim_{n\to\infty} X_n(\omega) = X\right\}\right) = 1$$

那么我们称 X_n 几乎必然收敛于 X，记为 $X_n \xrightarrow{\text{a.s.}} X$，或者 $X_n \to X\ a.s.$。

回忆几乎处处（$a.e$）和几乎必然（$a.s.$）的概念，几乎必然收敛意味着的确存在某些情况使得 $\lim_{n\to\infty} X_n(\omega) = X$ 不成立，然而不成立的概率为 0。

特别的，如果令 $X = c$ 为常数，是一个退化的随机变量，那么当 $n \to \infty$ 时，随机变量趋向于一个非随机的常数。在统计中，如果我们的估计量作为一个随机变量趋向于一个常数（通常是真值），那么这个估计量被称为**一致的**（**consistent**）。接下来我们将会经常碰到随机变量收敛到常数的情况。

例 6. 令 $\Omega = \mathbb{R}$，分布函数为：

$$F(\omega) = \begin{cases} 0 & \omega < -1 \\ \frac{1}{2}\omega + \frac{1}{2} & -1 \le \omega \le 1 \\ 1 & \omega > 1 \end{cases}$$

即在 $[-1, 1]$ 上的均匀分布，进而使用此分布函数构建 $\mathbb{R}$ 上的概率测度 P。随机变量

$$X_n(\omega) = \begin{cases} 0 & if\ |\omega| > \frac{1}{n} \\ n & if\ |\omega| \le \frac{1}{n} \end{cases}$$

因而 $\{\lim_{n\to\infty} X_n(\omega) \neq 0\} = \{0\}$，即只有在 $\omega = 0$ 处 X_n 不收敛到 0，进而 $\mathscr{P}\{\lim_{n\to\infty} X_n(\omega) \neq 0\} = \mathscr{P}\{0\} = 0$，因而 $X_n \xrightarrow{\text{a.s.}} 0$。

对于几乎必然收敛，我们有如下命题。

定理 3. $X_n \xrightarrow{a.s.} X$ 等价于对于任意的 $\epsilon > 0$，当 $n \to \infty$ 时，对于任意的 $k > n$，有：

$$\mathscr{P}(\{|X_k - X| < \epsilon\}) \to 1$$

Proof. 令

$$A_{n,\epsilon} = \{\omega \in \Omega : |X_k - X| < \epsilon\ \forall k \ge n\}$$

那么根据几乎必然收敛的定义，$X_n \to X$ 即对于任意的 ϵ，存在一个 n 使得对于任意的 $k > n$ 有 $|X_k - X| < \epsilon$，因而收敛的点可以表述为：

$$\bigcap_{\epsilon>0} \bigcup_{n=1}^{\infty} A_{n,\epsilon}$$

因而证明 $X_n \xrightarrow{a.s.} X$ 等价于证明 $\mathscr{P}\left(\cup_{\epsilon>0}\cap_{n=1}^{\infty} A_{n,\epsilon}\right)\to 1$。对于 $0<\epsilon_1<\epsilon_2$，由于 $\bigcup_{n=1}^{\infty} A_{n,\epsilon_1}\subset\bigcup_{n=1}^{\infty} A_{n,\epsilon_2}$，因而随着 $\epsilon\to 0$，$\bigcap_{\epsilon>0}\bigcup_{n=1}^{\infty} A_{n,\epsilon}\downarrow\bigcup_{n=1}^{\infty} A_{n,\epsilon}$。而由于 $A_{n,\epsilon}\subset A_{n+1,\epsilon}$ 对于 n 是单调递增的，因而随着 $n\to\infty$，$A_{n,\epsilon}\uparrow\bigcup_{n=1}^{\infty} A_{n,\epsilon}$。因而 $\mathscr{P}\left(\cup_{\epsilon>0}\cap_{n=1}^{\infty} A_{n,\epsilon}\right)\to 1$ 等价于 $\mathscr{P}\left(A_{n,\epsilon}\right)\to 1$，即命题得证。 □

2.2 依概率收敛

几乎必然收敛关注的是点态收敛，只要不收敛的点不要太多即可。而换一种思路，我们也可以关注随机变量 $X_n - X$ 之间的误差，如果两者之间的误差趋向于 0，我们也可以定义其为收敛，这就诞生了依概率收敛的概念。

定义 8. （依概率收敛）如果对于任意的 $\epsilon>0$，概率空间 $(\Omega,\mathscr{F},\mathscr{P})$ 上的一系列随机变量序列 $\{X_n\}$ 满足：

$$\mathscr{P}\left(|X_n - X|>\epsilon\right)\to 0$$

那么我们称 X_n 依概率收敛于 X，记为 $X_n \xrightarrow{P} X$，或 $\mathrm{plim} X_n = X$。

即随着 $n\to\infty$，X_n 与 X 之间误差比较大的点的概率趋向于 0。

例 7. 在例 (7) 中，对于任意的 $\epsilon>0$，可以得到 $|X_n-0|<\epsilon$ 的点集为：$\left\{\omega:|\omega|\le\frac{1}{n}\right\}$，因而对于任意的 $\epsilon>0$，都有 $\mathscr{P}\left(|X_n-0|<\epsilon\right)=\frac{2}{n}\to 0$，因而 $X_n\xrightarrow{P}0$。

注意依概率收敛和几乎必然收敛是两个不同的概念，依概率收敛并不一定能得到几乎必然收敛。

例 8. 令概率空间 $(\Omega,\mathscr{F},\mathscr{P})$ 如例 (7) 中定义，定义随机变量 $X_{j,n}, 0\le j<n-1$ 为：

$$X_{j,n}=\begin{cases}1 & \omega\in[\frac{j}{n},\frac{j+1}{n})\\ 0 & else\end{cases}$$

对于任意的 i，令 $n=\sup_n\left\{i>\frac{n(n+1)}{2}\right\}$，$j=(i \mod n)+1$，随机变量 $X_i=X_{j,n}$。对于任意的 $0<\epsilon<1$：

$$\mathscr{P}\left(|X_i|\ge\epsilon\right)=\frac{1}{n}$$

因而随着 $i\to\infty$，$n\to\infty$，$\mathscr{P}\left(|X_i|\ge\epsilon\right)\to 0$，因而 $X_i\xrightarrow{P}0$。然而随着 $i\to\infty$，除了 $\omega=1$ 之外，没有任何一个点收敛于 0，因而 X_i 并不几乎必然收敛于 0。

而相反，观察依概率收敛的定义，根据定理 (3)，对于任意的 $\epsilon>0$，

$$\mathscr{P}\left(|X_n-X|<\epsilon\right)\ge\mathscr{P}\left(\{|X_k-X|<\epsilon,\forall k>n\}\right)\to 1$$

因而几乎必然收敛可以得到依概率收敛。因而几乎必然收敛是比依概率收敛更强的一个结论。

与数列极限相似，我们也可以在概率收敛的语境下定义小 o 符号。

定义 9. $\{X_n\}$ 与 $\{Y_n\}$ 为定义在概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的两个随机变量序列，如果

$$\frac{X_n}{Y_n} \xrightarrow{P} 0$$

那么我们记为 $X_n = o_p(Y_n)$。特别的，当 $Y_n = 1$ 时，即 $X_n = o_p(1)$，等价于 $X_n \xrightarrow{P} 0$。

小 o_p 符号是对小 o 符号的推广，他们的性质非常相似。比如，对于定义在同一概率空间上的三个随机变量序列 $\{X_n\}, \{Y_n\}, \{Z_n\}$，如果 $X_n = o_p\{Y_n\}$，那么 $X_n Z_n = o_p(Y_n Z_n)$。特别的，如果 $Z_n = a_n$ 为退化的随机变量序列，即实数序列，那么 $a_n X_n = o_p(a_n Y_n)$。例如，如果 $X_n = o_p(n)$，那么 $\frac{1}{n} X_n = o_p(1)$。

类似的，小 o_p 符号允许我们做近似的逼近。比如如果随机变量 $Z_n = X_n + o_p(1)$，那么我们可以使用 X_n 对 Z_n 进行近似，因为两者当 $n \to \infty$ 时是等价的。

类似的，我们还可以定义大 O_p 符号。

定义 10. $\{X_n\}$ 与 $\{Y_n\}$ 为定义在概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的两个随机变量序列，如果对于任意的 $\epsilon > 0$，存在一个 C_ϵ 使得：

$$\sup_n \mathscr{P}(|X_n| \geq C_\epsilon |Y_n|) < \epsilon$$

那么我们记 $X_n = O_p(Y_n)$。特别的，当 $Y_n = 1$ 时，我们称 X_n **依概率有界**（**bounded in probability**）。

同样，大 O_p 符号是对大 O 符号的推广。如果 $X_n = O_p(1)$，那么意味着 X_n 不能太大，尽管允许 $X_n(\omega)$ 的某些值趋向于 ∞，但是这样的点随着 $n \to \infty$，其概率也慢慢变为 0。例如，在例 (7) 中，尽管 $X_n(0) \to \infty$，但是对于任意的 ϵ，令 $C_\epsilon = \frac{2}{\epsilon} + 1$，那么上式成立，因而 $X_n = O_p(1)$。

注意根据切比雪夫不等式，如果 $\mathrm{Var}(X_n) < M$，即随机变量序列 $\{X_n\}$ 的方差有界，那么对于任意的 $\epsilon > 0$，取 $C_\epsilon = \sqrt{\mathbb{E}(X_n^2)/\epsilon + 1}$，那么：

$$\mathscr{P}(|X_n| \geq C_\epsilon) \leq \frac{\mathbb{E}(X_n^2)}{C_\epsilon^2} = \frac{\mathbb{E}(X_n^2)}{\mathbb{E}(X_n^2)/\epsilon + 1} < \epsilon$$

因而 $X_n = O_p(1)$。

与大 O 符号类似，大 O_p 符号有如下性质：

定理 4.（大 O_p 性质）如果 $X_n = o_p(1), Y_n = o_p(1), Z_n = O_p(1), W_n = O_p(1)$，那么：

1. $X_n + Y_n = o_p(1)$

2. $X_n + Z_n = O_p(1)$

3. $Z_n + W_n = O_p(1)$

4. $X_n Y_n = o_p(1)$

5. $X_n Z_n = o_p(1)$

6. $Z_n W_n = O_p(1)$

以上特征可以分别简记为：$o_p(1)+o_p(1) = o_p(1), O_p(1)+O_p(1) = O_p(1), O_p(1)+o_p(1) = O_p(1), o_p(1) \cdot o_p(1) = o_p(1), O_p(1) \cdot o_p(1) = o_p(1), O_p(1) \cdot O_p(1) = O_p(1)$。

2.3 均方收敛

依概率收敛是讨论当 $n \to \infty$ 时 X_n 与 X 的误差，而类似于条件期望的定义，我们也可以使用平方的期望作为误差的一个度量，这就催生了均方收敛的概念。

定义 11.（均方收敛）如果概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的一系列随机变量序列 $\{X_n\}$ 随着 $n \to \infty$ 满足：

$$\mathbb{E}\left(|X_n - X|^2\right) \to 0$$

那么我们称 X_n 均方收敛于 X，记为 $X_n \xrightarrow{L^2} X$。

均方收敛意味着，随着 $n \to \infty$，X_n 与 X 之间误差的平方的期望是趋向于 0 的。同样，均方收敛也是一个比依概率收敛更强的收敛。

定理 5. 如果随机变量序列 $X_n \xrightarrow{L^2} X$，那么 $X_n \xrightarrow{P} X$。

Proof. 根据切比雪夫不等式，对于任意的 $\epsilon > 0$，有：

$$\mathscr{P}(|X_n - X| > \epsilon) \leq \frac{\mathbb{E}\left(|X_n - X|^2\right)}{\epsilon^2} \to 0$$

□

实际上均方收敛的概念可以扩展到任意的 $r > 0$，如果 $\mathbb{E}(|X_n - X|^r) \to 0$，那么我们称 X_n 依 r 阶均值收敛于 X（**convergence in the rth mean**），记为 $X_n \xrightarrow{L^r} X$。可以证明，如果 $X_n \xrightarrow{L^r} X$，那么 $X_n \xrightarrow{P} X$。

然而相反则不成立，如在在例 (7) 中，$\mathbb{E}\left(|X_n|^2\right) = 2n$，并不趋向于 0，尽管 $X_n \xrightarrow{\text{a.s.}} 0$ 进而 $X_n \xrightarrow{P} 0$，然而 X_n 并不均方收敛于 0。因而均方收敛也是比依概率收敛更强的一个结论。而尽管均方收敛和几乎必然收敛都比依概率收敛要强，但是这两者之间并没有强弱的关系，也不等价，只有在一定条件下，几乎必然收敛才与均方收敛才同时成立。

2.4 依分布收敛

之前讨论的收敛分别从点态和误差的角度讨论了随机变量收敛的问题，接下来我们从随机变量的分布函数的角度讨论随机变量的收敛性。

如果 $\{X_n\}$ 为一系列随机变量，其对应的分布函数为 $F_n(x)$，那么其极限：

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

如果存在，那么一个自然的问题是，$F(x)$ 是否还能构成一个分布函数。极限的单调性显然成立，那么接下来需要验证：

$$\lim_{x\to-\infty} F(x) = 0, \lim_{x\to\infty} F(x) = 1 \tag{1}$$

以及 $F(x)$ 的右连续性。

首先，针对式 (1)，有以下定理：

定理 6. 对于随机变量序列 $\{X_n\}$ 及其对应的分布函数 $\{F_n(x)\}$，如果

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

那么式 *(1)* 成立的充分必要条件是 $X_n = O_p(1)$。

例 9. 如果令随机变量

$$X_n = \begin{cases} 0 & with\ prob\ 1-p \\ n & with\ prob\ p \end{cases}$$

可知 X_n 并不是有界的，而

$$F_n(x) = \begin{cases} 0 & x < 0 \\ 1-p & 0 \le x < n \\ 1 & x = n \end{cases}$$

那么显然 $\lim_{n\to\infty} F_n(x) = 0 = F(x)$，因而 $\lim_{x\to\infty} F(x) = 0 \neq 1$。

以上定理解决了极限函数 $F(x)$ 的极限问题，然而对于右连续的要求，却并不能保证。比如令 $X_n = X + \frac{1}{n}$，X 的分布函数为 $G(x)$，那么 $F_n(x) = G\left(x - \frac{1}{n}\right)$，因而 $F(x) = \lim_{n\to\infty} F_n(x) = G(x-)$，如果 $G(x)$ 在 x 处不连续，那么 $F(x)$ 在 x 处为左连续函数。

因而为了避免右连续的问题，依分布收敛的定义通常只考虑分布函数的连续处的点。正式的，依分布收敛的定义如下：

定义 12. （依分布收敛）令 F_n, F 为分布函数，如果对于每一个 $F(x)$ 连续的

点 x，有：

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

那么我们称 $F_n(x)$ 弱收敛于 $F(x)$，记为 $F_n \xrightarrow{w} F$。如果一系列随机变量 $\{X_n\}$ 的分布函数 $F_{X_n}(x) \xrightarrow{w} F_X$，我们称 X_n 依分布收敛于 X，记为 $X_n \xrightarrow{d} X$。

注意以上定义只要求随机变量 X_n 的分布函数收敛，而并没有对 X_n 和 X 之间的关系做出限定。实际上，根据依分布收敛的定义，$\{X_n\}$ 甚至不需要来自于同一个概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$。同时需要注意，弱收敛的定义中也没有要求对于每个 x 都有 $\lim_{n\to\infty} F_n(x) = F(x)$，而仅仅只要求了 $F(x)$ 在 x 处连续。

上面我们介绍了只有当 $X_n = O_p(1)$ 时，X_n 的分布函数的极限才可能是分布函数。而反过来依然成立，即如果 X_n 依分布收敛，那么 $X_n = O_p(1)$。由于依分布收敛于 O_p 符号的这种关系，依分布收敛的很多性质可以与 O_p 类比，比如：

定理 7. （*Slutsky*）如果随机变量 $X_n \xrightarrow{d} X$，$R_n = o_p(1)$，那么 $X_n + R_n \xrightarrow{d} X$。同时如果 $Y_n \xrightarrow{P} a \neq 0$，那么 $\frac{X_n}{Y_n} \xrightarrow{d} X/a$。如果 $Y_n \xrightarrow{P} a$，那么 $X_n Y_n \xrightarrow{d} aX$。

即以上可表述为 $O_p(1) + o_p(1) = O_p(1)$，$\frac{O_p(1)}{a+o_p(1)} = O_p(1)/a$，$O_p(1) \cdot (a + o_p(1)) = aO_p(1)$。比如令 $Y_n \xrightarrow{P} a, Z_n \xrightarrow{P} b$，即 $Y_n - a = o_p(1), Z_n - b = o_p(1)$，其中 a, b 为常数，那么：

$$Z_n X_n + Y_n \xrightarrow{d} bX + a$$

实际上依分布收敛是一个比依概率收敛还要弱的收敛。我们有如下结论：

定理 8. 如果 $X_n \xrightarrow{P} X$，那么 $X_n \xrightarrow{d} X$。

而相反，依分布收敛则不一定可以得到依概率收敛。我们有如下结论：

定理 9. 如果 $X_n \xrightarrow{d} c$，那么 $X_n \xrightarrow{P} c$。

此外，依分布收敛的概念与期望的收敛密不可分。我们有如下定理：

定理 10. $X_n \xrightarrow{d} X$ 的充分必要条件为，对于任意的有界连续函数 g，有：

$$\mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)]$$

注意以上定理成立的前提是函数 $g(x)$ 必须为连续且有界的函数。比如如下两个例子中，如果违背了两个假定，结论并不一定成立。

例 10. 令 $g(x) = x$，随机变量

$$X_n = \begin{cases} n & with\ prob\ \frac{1}{n} \\ 0 & with\ prob\ 1 - \frac{1}{n} \end{cases}$$

那么 $X_n \xrightarrow{d} 0$，然而 $\mathbb{E}\left[g\left(X_n\right)\right] = n \cdot \frac{1}{n} = 1 \neq 0 = \mathbb{E}\left(0\right)$。

例 11. 令 $X_n = \frac{1}{n}$ 为退化的随机变量，那么 $X_n \xrightarrow{d} 0$。令

$$g\left(x\right) = \begin{cases} 1 & if\ x > 0 \\ 0 & if\ x = 0 \end{cases}$$

那么 $\mathbb{E}\left[g\left(X_n\right)\right] = 1 \nrightarrow 0 = \mathbb{E}\left[g\left(0\right)\right]$。

2.5 几种收敛之间的关系

以上我们介绍了四种收敛的概念，下面我们将集中收敛之间的关系整理如下：

定理 11. （四种收敛之间的关系）$\{X_n\}$ 为概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的一系列随机变量，那么

1. $X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{P} X$
2. $X_n \xrightarrow{L^r} X \Rightarrow X_n \xrightarrow{P} X, r > 0$，特别的，$X_n \xrightarrow{L^2} X \Rightarrow X_n \xrightarrow{P} X$
3. $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{d} X$
4. 如果 $X_n \xrightarrow{P} X$，那么存在 X_n 的一个子列 $\left\{X_{n_j}\right\}$，当 $j \to \infty$ 时，$X_{n_j} \xrightarrow{a.s.} X$
5. 如果 $X_n \xrightarrow{d} X$ 且 $P\left(X = c\right) = 1$，那么 $X_n \xrightarrow{P} c$
6. 若 $X_n \xrightarrow{a.s.} X$，且对于 $r > 0$ 以及一个正的随机变量 Z，满足 $\mathbb{E}\left(Z\right) < \infty$，如果 $\left|X_n\right|^r \leq Z$，那么 $X_n \xrightarrow{L^r} X$
7. 若 $X_n \xrightarrow{a.s.} X$，$X_n \geq 0$，且 $\mathbb{E}\left(X_n\right) \to \mathbb{E}\left(X\right) < \infty$，那么 $X_n \xrightarrow{L^1} X$
8. 对于任意的 $\epsilon > 0$，如果 $\sum_{n=1}^{\infty} \mathscr{P}\left(\left|X_n - X\right| \geq \epsilon\right) < \infty$，那么 $X_n \xrightarrow{a.s.} X$。
9. $X_n \xrightarrow{P} 0 \Longleftrightarrow \mathbb{E}\left(\frac{\left|X_n\right|}{1+\left|X_n\right|}\right) \to 0$

此外，我们上面讨论的都是一元随机变量的收敛，所有概念都可以扩展到随机向量中，仅需要把所有的绝对值替换为欧几里得范数，即 $\|x\| = \sqrt{x'x}$。

3 大数定律

在实际应用中，我们经常关注一系列随机变量的和的极限情况，即：

$$S_n = \sum_{i=1}^{n} X_i$$

的极限情况。实际上我们下面要讨论的**大数定律**（**Law of Large Numbers, LLN**）即在讨论随机变量和的极限行为。特别的，在大数定律中，我们最为关注的是样本均值的极限情况，即：

$$\frac{S_n - \mathbb{E}(S_n)}{n} \xrightarrow{P} 0$$

是否成立。实际上根据上一节的讨论，如果均方收敛那么则必然有依概率收敛，所以我们不妨从均方收敛入手。

实际上，$S_n - \mathbb{E}(S_n) = \sum_{i=1}^{n} [X_i - \mathbb{E}(X_i)]$，因而：

$$\begin{aligned}
\mathbb{E}[S_n - \mathbb{E}(S_n)]^2 &= \mathbb{E}\left(\sum_{i=1}^{n}[X_i - \mathbb{E}(X_i)]\right)^2 \\
&= \mathbb{E}\left(\sum_{i=1}^{n}[X_i - \mathbb{E}(X_i)]^2 + 2\sum_{1\le j<i\le n}^{n}[X_i - \mathbb{E}(X_i)][X_j - \mathbb{E}(X_j)]\right) \\
&= \sum_{i=1}^{n}\mathrm{Var}(X_i) + 2\sum_{1\le j<i\le n}^{n}\mathrm{Cov}(X_i, X_j)
\end{aligned}$$

如果假设每个 X_i 都有有限的二阶矩，即存在一个 M 使得对于所有的 $i = 1, 2, \ldots$，都有 $\mathrm{Var}(X_i) < M$，那么 $\sum_{i=1}^{n}\mathrm{Var}(X_i) < nM$，所以 $\sum_{i=1}^{n}\mathrm{Var}(X_i) = O(n)$。而根据 Cauthy-Schwartz 不等式，如果二阶矩有限，任意两个随机变量的协方差也必然有界，所以 $\sum_{1\le j<i\le n}^{n}\mathrm{Cov}(X_i, X_j) = O(n^2)$。进而：

$$\mathbb{E}\left[\frac{S_n - \mathbb{E}(S_n)}{n}\right]^2 = \frac{1}{n^2}O(n) + \frac{1}{n^2}O(n^2) = o(1) + O(1)$$

因而当 $\mathrm{Cov}(X_i, X_j) = 0$，即 $\{X_i\}$ 之间两两不相关时，上式趋向于 0。因而我们有如下定理：

定理 12. 如果概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的一个随机变量序列 $\{X_i\}$ 两两不相关，且存在一个 M 使得对于所有的 $i = 1, 2, \ldots$，都有 $Var(X_i) < M$，那么：

$$\frac{S_n - \mathbb{E}(S_n)}{n} \xrightarrow{L^2} 0$$

从而

$$\frac{S_n - \mathbb{E}(S_n)}{n} \xrightarrow{P} 0$$

注意在以上定理中，我们并没有限定 X_i 具有相同的均值或者相同的方差，实际上，如果令

$$\bar{\mu}_n = \frac{\mathbb{E}(S_n)}{n} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) \tag{2}$$

那么可以得到均值 $S_n/n \xrightarrow{P} \bar{\mu}_n$，即样本均值收敛到期望的平均。

推论 1. 如果 $\{X_i\}$ 为两两独立且同分布的随机变量序列，且其方差存在，记 $\mu = \mathbb{E}(X_i)$，那么：

$$\frac{S_n - \mu}{n} \xrightarrow{L^2} 0$$

或者：

$$\frac{S_n}{n} \xrightarrow{L^2} \mu$$

即样本均值收敛到其期望。

实际上，定理 (12) 不仅在均方收敛和依概率收敛的意义下成立，在几乎必然收敛的意义下也成立，即满足定理 (12) 的假设下，有：

$$\frac{S_n}{n} \xrightarrow{\text{a.s.}} \mu$$

证明见 Ch. 5.1, Chung (2001)。

以上的大数定律建立在不相关和二阶矩有界的条件下，同时得到了依概率收敛和几乎必然收敛的结果。实际上，根据结论中收敛方式的不同，大数定律分为「**强大数定律（Strong Law of Large Numbers，SLLN）**」和「**弱大数定律（Weak Law of Large Numbers，WLLN）**」，两者的区别在于，SLLN 需要得到几乎必然收敛这一更强的结果，而弱大数定律只要求得到依概率收敛这一比较弱的结果。相应的，强大数定律需要的假设条件更强，而弱大数定律需要的假设更弱。

实际上，大数定律通常在几种不同的假设之间做权衡，比如随机变量之间的相关性（独立、两两独立、不相关）、是否同分布以及是否存在高阶矩。通常情况下，为了同样得到某种收敛，如果放松了某个假设，则必须在另外的假设上加强。而在这其中，**独立同分布（independent and identically distributed, $i.i.d$）**是相关性和是否同分布两个假设条件中最宽松的假设条件。

对于一个随机变量序列 $\{X_i, i = 1, 2, ...\}$，如果对于任意的 $(\alpha_1, ..., \alpha_k), k = 2, 3, ...$，$\{X_{\alpha_1}, X_{\alpha_2}, ...X_{\alpha_k}\}$ 之间相互独立，那么我们称随机变量序列 $\{X_i\}$ 相互独立。如果随机变量序列 $\{X_i, i = 1, 2, ...\}$ 中每个随机变量的分布都相同，那么我们称 $\{X_i\}$ 同分布。如果两者同时成立，那么我们称 $\{X_i\}$ 为独立同分布的随机变量序列，简称为 $i.i.d$。

注意 $\{X_i\}$ 相互独立是比两两独立更强的假设，$\{X_i\}$ 相互独立必须要求任意数量的随机变量组合挑出来都是独立的，而两两独立只要求任意两个随机变量 X_i 和 X_j 是独立的。

下面我们就分别探讨弱大数定律和强大数定律。

3.1 弱大数定律

虽然依概率收敛结论比较弱，但是在通常情况下，依概率收敛是最简单，而且在很多应用中也已经足够的收敛形式，而弱大数定律就是解决随机变量之和的依概率收敛的问题。正式的，我们定义如下：

定义 13. 概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的一个随机变量序列 $\{X_i\}_{i\geq 1}$，如果对于数列 $\{a_n\}_{n\geq 1}$ 和 $\{b_n\}_{n\geq 1}$，随着 $n \to \infty$，满足：

$$\frac{S_n - a_n}{b_n} \xrightarrow{P} 0$$

那么我们称 $\{X_i\}$ 服从弱大数定律。

下面的例子是在独立同分布和二阶矩有限的情况下得到的最简单的弱大数定律。

例 12. 令 $\{X_i\}$ 为一系列 $i.i.d$ 的的随机变量，且 $\mathbb{E}\left(X_i^2\right) < \infty$，令 $\mu = \mathbb{E}(X_i), \sigma^2 = \mathrm{Var}(X_i)$，那么根据切比雪夫不等式：

$$P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) \leq \frac{\mathrm{Var}\left(\frac{S_n}{n}\right)}{\epsilon^2} = \frac{1}{\epsilon^2}\frac{n\sigma^2}{n^2} = \frac{\sigma^2}{\epsilon^2}\frac{1}{n} = o(1)$$

从而 $S_n/n \xrightarrow{P} \mu$。实际上，直接使用定理 (12) 可以得到同样的结果。

例 13. 如果令 $\{X_i\}$ 为一系列 $i.i.d$ 的随机变量，且 $X_i \sim Ber(p)$，那么 $\mathbb{E}(X_i) = p, \mathrm{Var}(X_i) = p(1-p) < \infty$，定义：

$$\hat{p} = \frac{\sum_{i=1}^n X_i}{n}$$

即成功的比例，那么根据上例，可以得到 $\hat{p} \xrightarrow{P} p$。

练习 6. 程序题：给定一个 p 和一个 n，重复的生成 n 个服从伯努利分布的随机变量，并计算其均值 $\hat{p}_n$。对于 $n = 10, 20, 30, ..., 1000$ 重复以上过程，并将结果以 n 为 x 轴，将 $\hat{p}_n$ 画在一张图上观察其收敛性。将伯努利分布换成 Cauthy 分布，再次观察其收敛性。

以上的弱大数定律是在二阶矩（方差）有限、独立同分布的条件下得到的，而这些假设仍然可以放宽。例如，根据定理 (12)，在二阶矩有限的条件下，同分布的假设可以不用，而独立的条件可以放宽为两两不相关。而以下的定理放松了二阶矩有限的假定以及独立的假定，保留了独立同分布的假定：

定理 13. 令 $\{X_i\}$ 为概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的两两独立且同分布的随机变量序列，若 $\mathbb{E}|X_i| < \infty$，那么 $S_n/n \xrightarrow{P} \mu$，其中 $\mu = \mathbb{E}(X_i)$。

而以下的定理则同时放宽了同分布的假定以及二阶矩的假定。

定理 14. 令 $\{X_i\}$ 为概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的独立的随机变量序列，如果存在一个常数 $p \in [1, 2]$，随着 $n \to \infty$，使得：

$$\frac{1}{n^p}\sum_{i=1}^n \mathbb{E}|X_i|^p \to 0$$

那么 $S_n/n \xrightarrow{P} \mu_n$，其中 μ_n 根据式 *(2)* 定义。

3.2 强大数定律

以上讨论了弱大数定律，然而很多时候我们仍然需要更强的结论，如几乎必然收敛。因而引入强大数定律就非常必要了。

定义 14. 概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的一个随机变量序列 $\{X_i\}_{i\geq 1}$，如果对于数列 $\{a_n\}_{n\geq 1}$ 和 $\{b_n\}_{n\geq 1}$，随着 $n \to \infty$，满足：

$$\frac{S_n - a_n}{b_n} \xrightarrow{\text{a.s.}} 0$$

那么我们称 $\{X_i\}$ 服从强大数定律。

下面的例子是在独立同分布和四阶矩有限的情况下得到的最简单的强大数定律。

例 14. （Borel's SLLN）令 $\{X_i\}$ 为概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的 $i.i.d$ 的随机变量序列，且 $\mathbb{E}X_i^4 < \infty$。根据切比雪夫不等式：

$$\begin{aligned}
P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) &\leq \frac{\mathbb{E}\left[\left(\frac{S_n}{n} - \mu\right)^4\right]}{\epsilon^4} \\
&= \frac{\mathbb{E}\left[(S_n - n\mu)^4\right]}{n^4\epsilon^4} \\
&= \frac{\mathbb{E}\left[\left(\sum_{i=1}^n (X_i - \mu)\right)^4\right]}{n^4\epsilon^4} \\
&= \frac{n\mathbb{E}\left[(X_i - \mu)^4\right] + 3n(n-1)\left[\mathbb{E}(X_i - \mu)^2\right]^2}{n^4\epsilon^4} \\
&= O\left(\frac{1}{n^2}\right)
\end{aligned}$$

根据定理 (11.8)，可以得到 $S_n/n \xrightarrow{\text{a.s.}} \mu$，其中 $\mu = \mathbb{E}(X_i)$。

可以看到，为了得到更强的结论（几乎必然收敛），需要使用更强的假设（四阶矩有限而非二阶矩有限）。回忆一下定理 (12) 也可以得到几乎必然收敛的结论，然而其中的条件仍然可以继续放宽。比如下面的 SLLN 就放宽了矩的假设，然而将强了独立性的假设以及同分布的假设。

定理 15. （*Etemadi's SLLN*）令 $\{X_i\}$ 为概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的两两独立且同分布的随机变量序列，且 $\mathbb{E}(X_i) < \infty$，那么 $S_n/n \xrightarrow{a.s.} \mu$。

实际上，上述定理的条件与定理 (13) 中的条件是一样的，而几乎必然收敛可以推出依概率收敛，因而上述定理实际上可以导出定理 (13)。

而以下定理相对于定理 (12) 则放宽了同分布及二阶矩有限的假定，加强了独立性的假定。

定理 16. 令 $\{X_i\}$ 为概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的独立的随机变量序列，且 $\mathbb{E}(X_i) < \infty$，如果存在一个常数 $p \in [1,2]$，随着 $n \to \infty$，使得：

$$\sum_{i=1}^{\infty} \frac{\mathbb{E}|X_i|^p}{i^p} < \infty$$

那么 $S_n/n \xrightarrow{a.s.} \mu_n$。

例 15. 在例 (13) 中，使用定理 (15)，我们同样可以得到 $\hat{p} \xrightarrow{\text{a.s.}} p$。现在假设我们有一系列 $i.i.d$ 的随机变量 $\{X_i, i = 1, ..., n\}$，其分布函数为 $F(x)$，那么定义**经验分布函数**（**empirical distribution funciton**）为：

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{[X_i,\infty)}(x) = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \le x\}$$

可以得到随机变量 $1_{[X_i,\infty)}(x) \sim Ber(F(x))$，因而 $\hat{F}_n(x) \xrightarrow{\text{a.s.}} F(x)$。实际上，我们可以得到更强的结论，即

$$P\left\{\sup_x \left|\hat{F}_n(x) - F(x)\right| \to 0\right\} = 1$$

首先令 $\epsilon > 0$ 为任意小的常数，令整数 $k > \frac{1}{\epsilon}$，并令 $-\infty = x_0 < x_1 \le x_2 \le \cdots \le x_{k-1} < x_k = \infty$，使得对于 $j = 1, ..., k-1$，有 $F(x_j-) \le \frac{j}{k} \le F(x_j)$。注意如果 $x_{j-1} < x_j$，那么有 $F(x_j-) - F(x_{j-1}) \le \epsilon$。由于 $\hat{F}_n(x) \xrightarrow{\text{a.s.}} F(x), \hat{F}_n(x-) \xrightarrow{\text{a.s.}} F(x-)$，因而：

$$\Delta_n = \max\{|F_n(x_j) - F(x_j)|, |F_n(x_j-) - F(x_j-)|, j = 1, ..., k-1\} \xrightarrow{\text{a.s.}} 0$$

令任意的 $x \in (x_{j-1}, x_j]$，那么

$$\begin{aligned}\hat{F}_n(x) - F(x) &\le \hat{F}_n(x_j-) - F(x_{j-1}) \\ &\le \hat{F}_n(x_j-) - F(x_j-) + \epsilon\end{aligned}$$

以及

$$\begin{aligned}\hat{F}_n(x) - F(x) &\ge \hat{F}_n(x_{j-1}) - F(x_j-) \\ &\ge \hat{F}_n(x_{j-1}) - F(x_{j-1}) + \epsilon\end{aligned}$$

因而

$$\sup_x \left|\hat{F}_n(x) - F(x)\right| \le \Delta_n + \epsilon \xrightarrow{\text{a.s.}} \epsilon$$

由于对于任意 $\epsilon > 0$，上式都成立，因而

$$P\left\{\sup_x \left|\hat{F}_n(x) - F(x)\right| \to 0\right\} = 1$$

3.3 一致大数定律

以上大数定律讨论的是一些随机变量的和的收敛，而**一致大数定律**（**Uniform law of large numbers, ULLN**）讨论的则是函数的收敛。

现在假设有一个函数 $g(x,\theta), \theta \in \Theta$，如果我们有一系列 $i.i.d$ 的随机变量 $\{X_i\}$，那么根据 SLLN，在可积性的条件下可以得到，对于任意的 $\theta \in \Theta$，有：

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i,\theta) \xrightarrow{\text{a.s.}} \mathbb{E}g(X_i,\theta) \triangleq g(\theta)$$

然而这个结论是基于点态收敛，仍然不足以支撑更多更有用的结论，很多时候我们需要这个收敛对 θ 是一致（uniformly）收敛，即：

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i,\theta) - g(\theta)\right| \xrightarrow{\text{a.s.}} 0$$

对于以上结论，我们有以下定理可以使用：

定理 17. 对于 $i.i.d$ 的随机变量 $\{X_i\}$ 以及函数 $g(x,\theta)$，如果：

1. Θ 为紧集
2. 对于所有的 x，$g(x,\theta)$ 对 θ 都是连续的
3. 存在一个不依赖于 θ 的函数 $K(x)$ 满足 $\mathbb{E}(K(x)) < \infty$，使得对于所有的 x 和 θ，有：$|g(x,\theta)| \leq K(x)$

那么：

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i,\theta) - g(\theta)\right| \xrightarrow{a.s.} 0$$

4 中心极限定理

以上我们讨论了不同的大数定律。除了 S_n 的收敛特性之外，我们还关心 S_n 极限时的分布情况，此时我们需要使用**中心极限定理**（**Central limit theorem, CLT**）。

根据之前对依分布收敛的讨论，我们需要找到一个 $O_p(1)$，进而可以得到

依分布收敛。如果假设 $\{X_i\}$ 支架两两不相关，那么：

$$\begin{aligned}\mathrm{Var}(S_n) &= \mathbb{E}[S_n - \mathbb{E}(S_n)]^2 \\ &= \mathbb{E}\left[\sum_{i=1}^{n}(X_i - \mathbb{E}(X_i))\right]^2 \\ &= \sum_{i=1}^{n}\mathrm{Var}(X_i) \\ &= O(n)\end{aligned}$$

因而 $S_n/\sqrt{n} = O_p(1)$。如果记 $\bar{X}_n = S_n/n$，那么 $\sqrt{n}\bar{X}_n = O_p(1)$。

对于 $i.i.d$ 的随机变量 $\{X_i\}$，我们有如下定理：

定理 18. 令 $\{X_i\}$ 为概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上 $i.i.d$ 的随机变量序列，且 $\mathbb{E}(X_i) = \mu, Var(X_i) = \sigma^2$，那么：

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

或：

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{d} N(0,1)$$

以上定理意味着，只要 X_i 有有限的二阶矩，那么不管 X_i 服从何种分布，其均值在极限的条件下都服从正态分布。

例 16. 如果 $\{X_i\}$ 为 $i.i.d$ 的随机变量，且 $X_i \sim Ber(p)$，令 $\hat{p}_n$ 如前定义，那么：

$$\sqrt{n}(\hat{p}_n - p) \xrightarrow{d} N(0, p(1-p))$$

练习 7. 程序题：给定一个 p 和一个 n，重复的生成 n 个服从伯努利分布的随机变量，并计算 $\hat{p}_n$。对于每一个 n，重复计算出 500 个 $\hat{p}$。对于 $n = 3, 10, 30, 100$ 重复以上过程，并画出每个 n 的情况下 500 个 $\hat{p}$ 的直方图。

例 17. 如果 $\{X_i\}$ 为 $i.i.d$ 的随机变量，且 $X_i \sim N(0,1)$，那么可知 $\mathbb{E}\left(X_i^2\right) = 1, \mathbb{E}\left(X_i^4\right) = 3$，因而：

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - 1\right) \xrightarrow{d} N(0,2)$$

练习 8. 程序题：将上述练习中的伯努利分布换成正态分布的平方，重复上述练习的过程。换成 Cauthy 分布，继续重复上述练习的过程。

以上定理还可以对随机向量进行扩展。

定理 19. 令 $\{X_i\}$ 为概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上 $i.i.d$ 的随机向量序列，且 $\mathbb{E}(X_i) = \mu, Var(X_i) = \Sigma$，那么：

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

或：

$$\sqrt{n}\Sigma^{-\frac{1}{2}}\left(\bar{X}_n - \mu\right) \xrightarrow{d} N(0, I)$$

以上中心极限定理仅适用于独立同分布的情况，而如果扩展到独立但不同分布的情况，情况就稍微复杂一点。为此，我们先引入一个随机变量的三角向量，即形如：

$$\begin{array}{llllll} X_{11} & & & & \sum \Rightarrow & Z_1 \\ X_{21} & X_{22} & & & \sum \Rightarrow & Z_2 \\ X_{31} & X_{32} & X_{33} & & \sum \Rightarrow & Z_3 \\ X_{41} & X_{42} & X_{43} & X_{44} & \sum \Rightarrow & Z_4 \\ \cdots & & & & & \cdots \end{array}$$

其中每一行的随机变量 X_{nj} 假设为相互独立的。我们有如下定理：

定理 20. （*Lindeberg-Feller*）对于 $n = 1, 2, ...$，令 $X_{nj}, j = 1, 2, ..., n$ 为独立的随机变量，$\mathbb{E}(X_{nj}) = 0$，$Var(X_{nj}) = \sigma_{nj}^2$。令

$$Z_n = \sum_{j=1}^{n} X_{nj}$$

并令

$$\sigma_n^2 = \sum_{j=1}^{n} \sigma_{nj}^2$$

如果 ***Lindeberg*** 条件成立，即对于任意的 $\epsilon > 0$，随着 $n \to \infty$，有：

$$\frac{1}{\sigma_n^2} \sum_{j=1}^{n} \mathbb{E}\left[X_{nj}^2 \cdot 1\left\{|X_{nj}| \geq \epsilon \sigma_n^2\right\}\right] \to 0$$

那么有

$$\frac{Z_n}{B_n} \xrightarrow{d} N(0, 1)$$

以上定理被称为 Lindeberg-Feller CLT。其中 Lindeberg 条件的一个推论是，随着 $n \to \infty$，

$$\max_{j \leq n} \frac{\sigma_{nj}^2}{\sigma_n^2} \to 0$$

也就是当 n 趋向于无穷时，任何随机变量的方差都是小到可以忽略不计的，即在 Z_n 中，每个随机变量 X_{nj} 对 Z_n 的影响可以不一样，但是没有一个 X_{nj} 对 Z_n 有决定性的影响。

实际上由于 Lindeberg 条件比较难以验证，很多时候我们会直接使用其充分条件，即如果存在 $\delta > 0$，有：

$$\sum_{j=1}^{n} \mathbb{E} \left| X_{nj} - \mathbb{E} X_{nj} \right|^{2+\delta} = o\left(\sigma_n^{2+\delta}\right)$$

那么 Lindeberg 条件即满足。

5 变换的收敛

以上讨论了随机变量的收敛，很多时候我们会关心随机变量的变换的收敛情况。比如，我们知道在大样本条件下，样本均值渐进服从正态分布，那么样本均值的平方服从何种分布呢？为此我们引入以下定理：

定理 21. 令 $\{X_i\}$ 为 k 维随机向量，$g(x) : \mathbb{R}^k \to \mathbb{R}^l$ 为连续的函数函数，那么：

1. $X_n \xrightarrow{a.s.} X \Rightarrow g(X_n) \xrightarrow{a.s.} g(X)$
2. $X_n \xrightarrow{P} X \Rightarrow g(X_n) \xrightarrow{P} g(X)$
3. $X_n \xrightarrow{d} X \Rightarrow g(X_n) \xrightarrow{d} g(X)$

例 18. 根据中心极限定理，$i.i.d$ 的 k 维随机变量 $\{X_i\}$ 满足：

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

那么 $\sqrt{n}\Sigma^{-\frac{1}{2}}\left(\bar{X}_n - \mu\right) \xrightarrow{d} N(0, I)$，从而 $n\left(\bar{X}_n - \mu\right)' \Sigma^{-1}\left(\bar{X}_n - \mu\right) \xrightarrow{d} \chi_k$。

例 19. 对于二维随机向量 (X, Y)，其相关系数定义为

$$\operatorname{Corr}(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}} = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}}$$

令 (X_i, Y_i) 为 $i.i.d$ 的样本，那么在可积性条件下，

$$\begin{cases} \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mathbb{E}(X) \\ \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{P} \mathbb{E}(Y) \\ \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \xrightarrow{P} \mathbb{E}(XY) \\ \frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{P} \mathbb{E}\left(X^2\right) \\ \frac{1}{n}\sum_{i=1}^{n} Y_i^2 \xrightarrow{P} \mathbb{E}\left(Y^2\right) \end{cases}$$

从而

$$\frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \frac{1}{n}\sum_{i=1}^{n} X_i \frac{1}{n}\sum_{i=1}^{n} Y_i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2}} \xrightarrow{P} \operatorname{Corr}(X, Y)$$

例 20. 之前曾讨论过，如果 $(X_1,...,X_n)' \sim N\left(\mu\iota,\sigma^2 I\right)$，那么：

$$\frac{\sqrt{n}\left(\bar{X}-\mu\right)}{\sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1}}} \sim t_{n-1}$$

现在我们不假设 X_i 服从正态分布，而是假设其独立同分布且具有有限的二阶矩，那么我们有

$$\bar{X} \xrightarrow{P} \mathbb{E}(X), \frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{P} \mathbb{E}\left(X^2\right)$$

进而：

$$\begin{aligned}
\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1} &= \frac{\sum_{i=1}^{n}\left(X_i^2+\bar{X}^2-2\bar{X}X_i\right)}{n-1} \\
&= \frac{\sum_{i=1}^{n} X_i^2+n\bar{X}^2-2\bar{X}\sum_{i=1}^{n} X_i}{n-1} \\
&= \frac{\sum_{i=1}^{n} X_i^2+n\bar{X}^2-2n\bar{X}^2}{n-1} \\
&= \frac{\sum_{i=1}^{n} X_i^2+n\bar{X}^2-2n\bar{X}^2}{n-1} \\
&= \frac{\sum_{i=1}^{n} X_i^2-n\bar{X}^2}{n-1} \\
&= \frac{\sum_{i=1}^{n} X_i^2}{n-1}-\frac{n}{n-1}\bar{X}^2 \\
&\xrightarrow{P} \mathbb{E}\left(X^2\right)-[\mathbb{E}(X)]^2 = \mathrm{Var}(X)
\end{aligned}$$

进而：

$$\frac{\sqrt{n}\left(\bar{X}-\mu\right)}{\sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1}}} \xrightarrow{P} \frac{\sqrt{n}\left(\bar{X}-\mu\right)}{\sqrt{\mathrm{Var}(X)}} = \sqrt{n}\left(\frac{\bar{X}-\mu}{\sqrt{\mathrm{Var}(X)}}\right) \xrightarrow{d} N(0,1)$$

因而当样本足够大时，即使 X_i 不服从正态分布，以上的 $\frac{\sqrt{n}\left(\bar{X}-\mu\right)}{\sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1}}}$ 仍然服从正态分布。

现在假设 $a_n(X_n-c) \xrightarrow{d} Y$，$\lim_{a_n\to\infty} a_n=\infty$，那么 $X_n = O_p\left(\frac{1}{a_n}\right), X_n - c = o_p(1)$。对于任意的连续二阶可微的函数 $g(x)$，都可以对其泰勒展开：

$$\begin{aligned}
a_n\left[g(X_n)-g(c)\right] &= \frac{\partial g}{\partial x'}(c)\, a_n(X_n-c)+\frac{1}{2}a_n(X_n-c)'\frac{\partial^2 g}{\partial x\partial x'}(c)(X_n-c)+\cdots \\
&= \frac{\partial g}{\partial x'}(c)\, a_n(X_n-c)+o_p(1) \xrightarrow{d} \frac{\partial g}{\partial x'}(c)\, Y
\end{aligned}$$

因而 $a_n\left[g\left(X_n\right)-g\left(c\right)\right] \xrightarrow{d} \frac{\partial g}{\partial x'}\left(c\right)Y$。特别的，如果 $Y \sim N\left(0,\Sigma\right)$，那么：

$$a_n\left[g\left(X_n\right)-g\left(c\right)\right] \xrightarrow{d} N\left(0, \frac{\partial g}{\partial x'}\left(c\right)\Sigma\frac{\partial g}{\partial x}\left(c\right)\right)$$

以上过程我们称之为 **delta 方法**（**delta method**）。

参考文献

[1] Athreya, K.B., Lahiri, S.N., 2006. Measure Theory and Probability Theory. Springer, New York.

[2] Chung, K.L., 2001. A Course in Probability Theory, 3rd editio. ed. Elsevier Ltd., Singapore.

[3] Lehmann, E.L., 1999. Elements of Large-Sample Theory. Springer Science & Business Media, New York.

[4] Shao, J., 2007. Mathematical Statistics, 2nd ed. Springer, New York.

[5] Wooldridge, J.M., 2010. Econometric Analysis of Cross Sectional and Panel Data, 2nd ed. The MIT Press, Cambridge.