中心极限定理

续本达

清华大学 工程物理系

2023-10-30 清华

复习

辛钦 Khinchin 大数定律

定理

设 $X_1, X_2, \cdots, X_n$ 是相互独立的随机变量序列，服从同一分布，且具有 $E(X_k) = \mu (k = 1, 2, \cdots, n)$。则对任意正数 ϵ，有

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

即

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{P} \mu$$

数学期望可以由 n 个独立同分布的随机变量的算术平均值近似。

伯努利 Bernoulli 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 是每次试验中 A 发生的概率，则 $\forall \epsilon > 0$，有

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

因为 $n_A = X_1 + X_2 + \cdots + X_n$，其中 $X_1, X_2, \cdots, X_n$ 相互独立，且都服从参数为 p 的 0-1 分布，因而 $\mathrm{E}(X_k) = p$。由辛钦大数定律有

$$1 = \lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_i X_i - p\right| < \epsilon\right) = \lim_{n\to\infty} P\left(\left|\frac{n_A}{n} - p\right| < \epsilon\right)$$

事件 A 发生的频率 $\frac{n_A}{n}$ 趋于其一次试验发生的概率

频率 $\frac{n_A}{n}$ 与 p 有较大偏差 $\left|\frac{n_A}{n} - p\right| \geq \epsilon$ 是小概率事件。因而在 n 足够大时，可以用频率近似代替 p。

引子

“中心极限定理”的由来

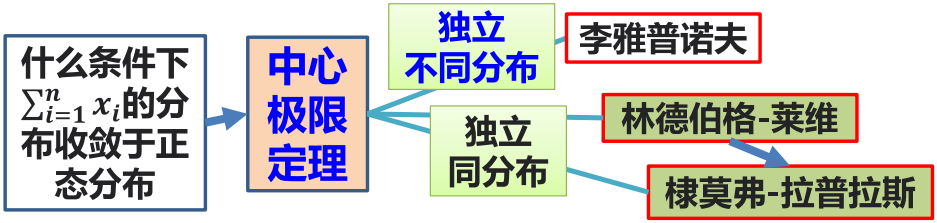
Polya 1920 年

It was generally known that the appearance of the Gaussian probability density e^{-x^2} in a great many situations can be explained by one and the same limit theorem, which plays a central role in probability theory.

“中心极限定理”的由来

Polya 1920 年

It was generally known that the appearance of the Gaussian probability density e^{-x^2} in a great many situations can be explained by one and the same limit theorem, which plays a central role in probability theory.



中心极限定理

林德伯格 Lindberg 列维 Levi 版本

设随机变量序列 $X_1, X_2, \cdots, X_n$ 独立同分布，且数学期望和方差存在：

$$\mathrm{E}(X_k) = \mu, \mathrm{Var}(X_k) = \sigma^2 > 0, k = 1, 2, \cdots$$

则随机变量之和 $X \equiv \sum_{k=1}^{n} X_k$ 的标准化变量

$$Y_n = \frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}} = \frac{X - n\mu}{\sqrt{n}\sigma}$$

的分布函数 $F_n(x)$ 对于任意实数 x 满足

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} P\left(\frac{X - n\mu}{\sqrt{n}\sigma} \le x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \Phi(x)$$

林德伯格 Lindberg 列维 Levi 版本

设随机变量序列 $X_1, X_2, \cdots, X_n$ 独立同分布，且数学期望和方差存在：

$$\mathrm{E}(X_k) = \mu, \mathrm{Var}(X_k) = \sigma^2 > 0, k = 1, 2, \cdots$$

则随机变量之和 $X \equiv \sum_{k=1}^{n} X_k$ 的标准化变量

$$Y_n = \frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}} = \frac{X - n\mu}{\sqrt{n}\sigma}$$

的分布函数 $F_n(x)$ 对于任意实数 x 满足

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} P\left(\frac{X - n\mu}{\sqrt{n}\sigma} \le x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \Phi(x)$$

解读

n 足够大时，Y_n 的分布函数近似于标准正态的。
$X = \sum_{k=1}^{n} X_k = \sqrt{n}\sigma Y_n + n\mu$ 近似服从 $N(n\mu, n\sigma^2)$。

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

$$\Longrightarrow \overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad n \to \infty$$

统计推断的基础

在 n 充分大时，均值为 μ，方差为 σ^2 的独立同分布随机变量 $X_1, X_2, \cdots, X_n$ 的算术平均值 $\overline{X}$，近似服从 $N\left(\mu, \frac{\sigma^2}{n}\right)$.

为何许多随机现象服从正态分布?

- 彼此没有什么相依关系、对随机现象谁也不能起突出影响，而“均匀”地起到微小作用的随机因素共同作用叠加，结果呈现正态分布。
- 若描述此随机现象的随机变量为 X，则它可被看成为许多相互独立的起微小作用的因素 X_k 的总和 $\sum_k X_k$，而这个总和近似服从正态分布。

例：棣莫弗 DeMoivre 拉普拉斯 Laplace 中心极限定理

是 Lindberg-Levi 中心极限定理的二项分布特例。
设 $Y_n \sim b(n,p)$，$0<p<1$，$n=1,2,\cdots$ 则对任一实数 x，有

$$\lim_{n\to\infty} P\left(\frac{Y_n - np}{\sqrt{np(1-p)}} \le x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

$$Y_n \sim N[np, np(1-p)], n \to \infty$$

例：良种数再次近似

例

设一大批种子中良种占 $\frac{1}{6}$. 试用切比雪夫不等式估计在任选的 6000 粒种子中, 良种比例与 $\frac{1}{6}$ 比较上下小于 1% 的概率范围.

设 X 表示 6000 粒种子中的良种数, $X \sim b\left(6000, \frac{1}{6}\right)$, $\mathrm{E}(X) = 1000$, $\mathrm{Var}(X) = \frac{5000}{6}$ 近似有

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

$$P(|X-1000| < 60) \approx \Phi\left(\frac{59}{\sqrt{5000/6}}\right) - \Phi\left(\frac{-59}{\sqrt{5000/6}}\right)$$

$$= 2\Phi\left(\frac{59}{\sqrt{5000/6}}\right) - 1 = 0.9590287$$

例：良种数再次近似

例

设一大批种子中良种占 $\frac{1}{6}$. 试用切比雪夫不等式估计在任选的 6000 粒种子中, 良种比例与 $\frac{1}{6}$ 比较上下小于 1% 的概率范围.

设 X 表示 6000 粒种子中的良种数, $X \sim b\left(6000, \frac{1}{6}\right)$，$\mathrm{E}(X) = 1000$，$\mathrm{Var}(X) = \frac{5000}{6}$ 近似有

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

$$P(|X-1000| < 60) \approx \Phi\left(\frac{59}{\sqrt{5000/6}}\right) - \Phi\left(\frac{-59}{\sqrt{5000/6}}\right)$$

$$= 2\Phi\left(\frac{59}{\sqrt{5000/6}}\right) - 1 = 0.9590287$$

例：精准的高斯

$P(|X-1000|<60)=P(|X-1000|\leq 59$ 对离散型随机变量 X 成立，但正态分布随机变量是连接型的。尝试

$$P(|X-1000|<60)\approx 2\Phi\left(\frac{59.5}{\sqrt{5000/6}}\right)-1=0.9607$$

比较

二项分布 0.9607
泊松分布 0.9401
切比雪夫不等式 0.7685

例：精准的高斯

$P(|X-1000|<60)=P(|X-1000|\leq 59$ 对离散型随机变量 X 成立，但正态分布随机变量是连接型的。尝试

$$P(|X-1000|<60)\approx 2\Phi\left(\frac{59.5}{\sqrt{5000/6}}\right)-1=0.9607$$

比较

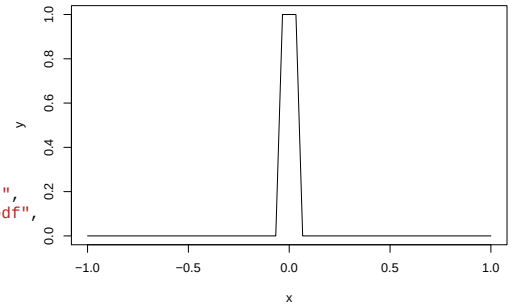
二项分布 0.9607

泊松分布 0.9401

切比雪夫不等式 0.7685

卷积的天花板

```
library(animation)
juan <- function(n=30, ker=c(1,2,1)/4) {
    x <- seq(-n,n)/n
    yt <- rep(0,n-1)
    y <- c(yt, c(1,1,1), yt)
    for (i in seq_len(ani.options("nmax"))) {
        dev.hold()
        plot(x, y, type='l', ylim=c(0,1))
        y <- c(0, convolve(y, ker, type="filter"), 0)
        ani.pause()
    }
}
saveLatex(juan(), nmax=100, img.name="plot/convolution",
          interval = 1, ani.dev = "pdf", ani.type = "pdf",
          ani.width = 7, ani.height = 5,
          latex.filename = NULL, pdfLatex=NULL)
```



证明

中心极限定理 by Levy

$$Y_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \xrightarrow[n\to\infty]{P} Z \sim N(0,1)$$

- 右边的特征函数 $\varphi_Z(t) = e^{-\frac{t^2}{2}}$
- 左边

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

$$\Longrightarrow \varphi_{Y_n}(t) = \left[\varphi_{X_i-\mu}\left(\frac{t}{\sqrt{n}\sigma}\right)\right]^n$$

$$= \left[1 + \frac{1}{2}\varphi''_{X_i-\mu}(0)\left(\frac{t}{\sqrt{n}\sigma}\right)^2 + o\left(\frac{1}{n}\right)\right]^n$$

$$= \left[1 + t^2\frac{1}{2}\mathrm{Var}(X_i)\frac{t^2}{n\sigma^2} + o\left(\frac{1}{n}\right)\right]^n \overset{n\to\infty}{=} e^{-\frac{t^2}{2}}$$

中心极限定理 by Levy

$$Y_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \xrightarrow[n\to\infty]{P} Z \sim N(0,1)$$

- 右边的特征函数 $\varphi_Z(t) = e^{-\frac{t^2}{2}}$
- 左边

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

$$\Longrightarrow \varphi_{Yn}(t) = \left[\varphi_{X_i-\mu}\left(\frac{t}{\sqrt{n}\sigma}\right)\right]^n$$

$$= \left[1 + \frac{1}{2}\varphi''_{X_i-\mu}(0)\left(\frac{t}{\sqrt{n}\sigma}\right)^2 + o\left(\frac{1}{n}\right)\right]^n$$

$$= \left[1 + t^2\frac{1}{2}\operatorname{Var}(X_i)\frac{t^2}{n\sigma^2} + o\left(\frac{1}{n}\right)\right]^n \overset{n\to\infty}{=} e^{-\frac{t^2}{2}}$$

中心极限定理 by Levy

$$Y_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \xrightarrow[n\to\infty]{P} Z \sim N(0,1)$$

- 右边的特征函数 $\varphi_Z(t) = e^{-\frac{t^2}{2}}$
- 左边

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

$$\Longrightarrow \varphi_{Yn}(t) = \left[\varphi_{X_i-\mu}\left(\frac{t}{\sqrt{n}\sigma}\right)\right]^n$$

$$= \left[1 + \frac{1}{2}\varphi''_{X_i-\mu}(0)\left(\frac{t}{\sqrt{n}\sigma}\right)^2 + o\left(\frac{1}{n}\right)\right]^n$$

$$= \left[1 + i^2\frac{1}{2}\mathrm{Var}(X_i)\frac{t^2}{n\sigma^2} + o\left(\frac{1}{n}\right)\right]^n \overset{n\to\infty}{=} e^{-\frac{t^2}{2}}$$

中心极限定理 by Levy

$$Y_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \xrightarrow[n\to\infty]{P} Z \sim N(0,1)$$

- 右边的特征函数 $\varphi_Z(t) = e^{-\frac{t^2}{2}}$
- 左边

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

$$\Longrightarrow \varphi_{Yn}(t) = \left[\varphi_{X_i-\mu}\left(\frac{t}{\sqrt{n}\sigma}\right)\right]^n$$

$$= \left[1 + \frac{1}{2}\varphi''_{X_i-\mu}(0)\left(\frac{t}{\sqrt{n}\sigma}\right)^2 + o\left(\frac{1}{n}\right)\right]^n$$

$$= \left[1 + i^2\frac{1}{2}\operatorname{Var}(X_i)\frac{t^2}{n\sigma^2} + o\left(\frac{1}{n}\right)\right]^n \overset{n\to\infty}{=} e^{-\frac{t^2}{2}}$$

中心极限定理 by Levy

$$Y_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \xrightarrow[n\to\infty]{P} Z \sim N(0,1)$$

- 右边的特征函数 $\varphi_Z(t) = e^{-\frac{t^2}{2}}$
- 左边

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

$$\begin{aligned}
\Longrightarrow \quad \varphi_{Yn}(t) &= \left[\varphi_{X_i-\mu}\left(\frac{t}{\sqrt{n}\sigma}\right)\right]^n \\
&= \left[1 + \frac{1}{2}\varphi''_{X_i-\mu}(0)\left(\frac{t}{\sqrt{n}\sigma}\right)^2 + o\left(\frac{1}{n}\right)\right]^n \\
&= \left[1 + i^2\frac{1}{2}\operatorname{Var}(X_i)\frac{t^2}{n\sigma^2} + o\left(\frac{1}{n}\right)\right]^n \overset{n\to\infty}{=} e^{-\frac{t^2}{2}}
\end{aligned}$$

释义

- 中心极限定理阐明了正态分布的来源；
- 与二项分布、指数分布等由物理世界的性质决定不同，正态分布从极限起源；
- 中心极限定理诠释了正分布的物理意义。

李雅普诺夫

李雅普诺夫 Lyapunov 中心极限定理

设 $X_1, X_2, \cdots, X_n, \cdots$ 是独立的随机变量序列，且具有数学期望和方差 $\mathrm{E}(X_k) = \mu_k$ ， $\mathrm{Var}(X_k) = \sigma_k^2 > 0 (k = 1, 2, \cdots)$ 。记 $B_n^2 = \sum_{k=1}^{n} \sigma_k^2$。若存在 $\delta > 0$，使得 *Lyapunov* 条件

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

成立，则随机变量之和 $\sum_{k=1}^{n} X_k$ 的标准化变量 $Y_n = \frac{\sum_{k=1}^{n} X_k - \sum_{k=1}^{n} \mu_k}{B_n}$ 的分布函数 $F_n(x)$ 对于任意的 x 满足：

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} P\left(\frac{\sum_{k=1}^{n} X_k - \sum_{k=1}^{n} \mu_k}{B_n} \le x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}} \mathrm{d}t$$

李雅普诺夫 Lyapunov 中心极限定理

设 $X_1, X_2, \cdots, X_n, \cdots$ 是独立的随机变量序列，且具有数学期望和方差 $\mathrm{E}(X_k) = \mu_k$ ， $\mathrm{Var}(X_k) = \sigma_k^2 > 0 (k = 1, 2, \cdots)$ 。记 $B_n^2 = \sum_{k=1}^{n} \sigma_k^2$。 若存在 $\delta > 0$，使得 *Lyapunov* 条件

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

成立，则随机变量之和 $\sum_{k=1}^{n} X_k$ 的标准化变量 $Y_n = \frac{\sum_{k=1}^{n} X_k - \sum_{k=1}^{n} \mu_k}{B_n}$ 的分布函数 $F_n(x)$ 对于任意的 x 满足：

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} P\left(\frac{\sum_{k=1}^{n} X_k - \sum_{k=1}^{n} \mu_k}{B_n} \le x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}} \mathrm{d}t$$

李雅普诺夫 Lyapunov 中心极限定理

设 $X_1, X_2, \cdots, X_n, \cdots$ 是独立的随机变量序列，且具有数学期望和方差 $\mathrm{E}(X_k) = \mu_k$ ， $\mathrm{Var}(X_k) = \sigma_k^2 > 0 (k = 1, 2, \cdots)$ 。记 $B_n^2 = \sum\limits_{k=1}^{n} \sigma_k^2$。 若存在 $\delta > 0$，使得 *Lyapunov* 条件

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

成立，则随机变量之和 $\sum\limits_{k=1}^{n} X_k$ 的标准化变量 $Y_n = \dfrac{\sum\limits_{k=1}^{n} X_k - \sum\limits_{k=1}^{n} \mu_k}{B_n}$ 的分布函数 $F_n(x)$ 对于任意的 x 满足：

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} P\left\{\frac{\sum\limits_{k=1}^{n} X_k - \sum\limits_{k=1}^{n} \mu_k}{B_n} \le x\right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}} \mathrm{d}t$$

Lyapunov CLT 要项

- 当 n 很大时， $Z_n = \frac{\sum_{k=1}^{n} X_k - \sum_{k=1}^{n} \mu_k}{B_n}$ 近似服从于标准正态分布 $N(0,1)$。
- 当 n 很大时， $\sum_{k=1}^{n} X_k$ 近似服从于正态分布 $N(\sum_{k=1}^{n} \mu_k, B_n^2)$ 。
- 即无论随机变量 $X_n (n = 1, 2, \cdots)$ 服从什么分布，只要满足定理的条件，它们的和 $\sum_{k=1}^{n} X_k$ 在 n 很大时，近似服从正态分布。

例 (自然现象往往近似的服从正态分布)

- 某一时刻一个城市的用电量是大量用户耗电的总和
- 一个物理实验的测量误差是由观察不到的，可加的微小误差合成的

Lyapunov CLT 要项

- 当 n 很大时，$Z_n = \frac{\sum_{k=1}^{n} X_k - \sum_{k=1}^{n} \mu_k}{B_n}$ 近似服从于标准正态分布 $N(0,1)$。
- 当 n 很大时，$\sum_{k=1}^{n} X_k$ 近似服从于正态分布 $N(\sum_{k=1}^{n} \mu_k, B_n^2)$ 。
- 即无论随机变量 $X_n (n = 1, 2, \cdots)$ 服从什么分布，只要满足定理的条件，它们的和 $\sum_{k=1}^{n} X_k$ 在 n 很大时，近似服从正态分布。

例 (自然现象往往近似的服从正态分布)

- 某一时刻一个城市的用电量是大量用户耗电的总和
- 一个物理实验的测量误差是由观察不到的，可加的微小误差合成的

Lyapunov CLT 要项

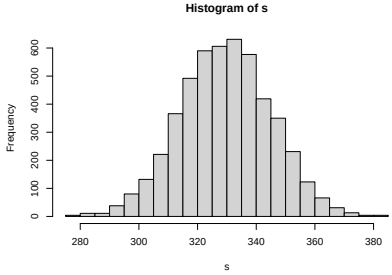
- 当 n 很大时， $Z_n = \frac{\sum_{k=1}^{n} X_k - \sum_{k=1}^{n} \mu_k}{B_n}$ 近似服从于标准正态分布 $N(0,1)$。
- 当 n 很大时， $\sum_{k=1}^{n} X_k$ 近似服从于正态分布 $N(\sum_{k=1}^{n} \mu_k, B_n^2)$ 。
- 即无论随机变量 $X_n (n = 1, 2, \cdots)$ 服从什么分布，只要满足定理的条件，它们的和 $\sum_{k=1}^{n} X_k$ 在 n 很大时，近似服从正态分布。

例 (自然现象往往近似的服从正态分布)

- 某一时刻一个城市的用电量是大量用户耗电的总和
- 一个物理实验的测量误差是由观察不到的，可加的微小误差合成的

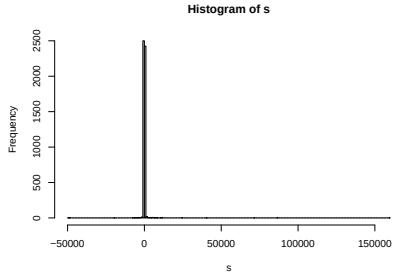
例：数值试验

```
N <- 5000
s <- rep(0, N)
for (i in 1:5) {
    s <- s + rpois(N, i)
    s <- s + rbinom(N, 68 * i, 0.3)
    s <- s + rgamma(N, 0.5 * i, 2)
    s <- s + rbeta(N, 10 - i , 0.2 * i)
}
hist(s, breaks=20)
```



反例柯西：方差不存在

```
N <- 5000
s <- rep(0, N)
for (i in 1:20) {
    s <- s + rcauchy(N)
}
hist(s, breaks=200)
```



马尔可夫

引子

- 切比雪夫的学生，继承了老师的圣彼得堡学派，研究中心极限定理
- Pavel Nekrasov 莫斯科学派领袖，认为大数定律的必要条件是被加的随机变量相互独立
- 自然现象中的被加项有关联
  - 理想气体、植物生长等都有时间演化历史，时间上的因果意识着被加随机变量之间有关联

相互不独立的被加随机变量

- 设 $X_1, X_2, \cdots, X_n, \cdots$ 是不相互独立的随机变量序列，且有

$$P(X_j|X_{j-1}, X_{j-2}, \cdots) = P(X_j|X_{j-1})$$

即 马尔可夫性

- 加上可逆性和可达性条件，使它形成一个 马尔可夫链 （随机过程章讨论）
- 那么

$$\mu = \mathrm{E}(X_1)$$

$$\sigma^2 = \mathrm{Var}(X_1) + 2\sum_{k=1}^{\infty} \mathrm{Cov}(X_1, X_{1+k}) < +\infty$$

$$\mu_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$

$$\Longrightarrow \mu_n \xrightarrow[n\to\infty]{P} Z \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$