第1-2章 随机样本

《统计推断》 第5章

感谢清华大学自动化系江瑞教授提供PPT

内容提要

- 收敛性定义
  - 依概率收敛
  - 几乎处处收敛
  - 依分布收敛
- Delta方法
- 样本生成

Convergence in Probability

A sequence of random variables $X_1, \ldots, X_n$, **converges in probability** to a random variable X if, for every $\varepsilon > 0$,

$$\lim_{n \to \infty} P(|X_n - X| \geq \varepsilon) = 0,$$

or equivalently,

$$\lim_{n \to \infty} P(|X_n - X| < \varepsilon) = 1.$$

记为： $X_n \xrightarrow{p} X.$

Convergence in Probability

对于任意 $\epsilon > 0, \delta > 0$，存在 n_0，只要 $n > n_0, P(|X_n - X| \geq \epsilon) < \delta$
或者等价地，只要 $n > n_0, P(|X_n - X| < \epsilon) > 1 - \delta$

也就是说，n充分大时X_n以很大的概率充分靠近X.

Suppose that $X_1, X_2, \ldots$ converges in probability to a random variable X and that h is a continuous function. Then $h(X_1), h(X_2), \ldots$ converges in probability to $h(X)$.

Example I

Let the sample space $\mathcal{S}$ be the closed interval [0,1] with sample points uniformly distributed.

Define random variable

$$X(s) = s, \ \ s \in S.$$

Define random variables

$$X_n(s) = s + s^n, \ \ s \in S.$$

Then

$$X_n(s) - X(s) = s^n, s \in S.$$

Now, for every $\varepsilon > 0$,

$$\begin{aligned}
&\lim_{n\to\infty} P(|\, X_n(s) - X(s) \,| \ \ < \varepsilon) \\
&= \lim_{n\to\infty} P(s^n < \varepsilon) \\
&= P(s \in [0,1)) \\
&= 1.
\end{aligned}$$

Therefore, X_n converges in probability to X.

Example II

Let the sample space $\mathcal{S}$ be the closed interval [0,1] with sample points uniformly distributed.

Define random variable

$$X(s) = s,\ \ s \in S.$$

Define random variables X_n as follows

$$\begin{aligned}
&X_1(s) = s + I_{[0,1]}(s),\\
&X_2(s) = s + I_{[0,1/2]}(s), X_3(s) = s + I_{[1/2,1]}(s),\\
&X_4(s) = s + I_{[0,1/3]}(s), X_5(s) = s + I_{[1/3,2/3]}(s), X_6(s) = s + I_{[2/3,1]}(s),\\
&\ldots
\end{aligned}$$

Then, for every $\varepsilon > 0$,

$$\begin{aligned}
&\lim_{n\to\infty} P(|\ X_n(s) - X(s)\ | \ \ < \varepsilon)\\
&= \lim_{k\to\infty} P(\text{length}(I[0,1/k]) < \varepsilon)\\
&= 1.
\end{aligned}$$

Therefore, X_n converges in probability to X.

Almost Sure Convergence
(Convergence with Probability 1)

A sequence of random variables $X_1, \ldots, X_n$, **converges almost surely** to a random variable X if, for every $\varepsilon > 0$,

$$P\left(\lim_{n\to\infty} |X_n - X| < \varepsilon\right) = 1.$$

简记为 $X_n \to X, a.s.$

几乎处处收敛的理解

- 令

$$\Omega_0 = \{\omega | \lim_{n \to +\infty} X_n(\omega) = X(\omega)\}$$

- 几乎处处收敛 $X_n \to X, a.s.$ 的等价定义是：

$$P(\Omega_0) = 1$$

Example I

Let the sample space $\mathcal{S}$ be the closed interval $[0,1]$ with sample points uniformly distributed.

Define random variable

$$X(s) = s,\ \ s \in S.$$

Define random variables

$$X_n(s) = s + s^n,\ \ s \in S.$$

For every $s \in [0,1)$, as $n \to \infty$,

$$X_n(s) \to s = X(s).$$

However, for $s = 1$,

$$X_n(s) = 1 + 1^n = 2 \neq X(s).$$

Since

$$P(s = 1) = 0 \text{ and } P(s \in [0,1)) = 1,$$

X_n converges almost surely to X.

Example II

Let the sample space $\mathcal{S}$ be the closed interval [0,1] with sample points uniformly distributed.

Define random variable

$$X(s) = s,\ \ s \in S.$$

Define random variables X_n as follows

$$X_1(s) = s + I_{[0,1]}(s),$$
$$X_2(s) = s + I_{[0,1/2]}(s), X_3(s) = s + I_{[1/2,1]}(s),$$
$$X_4(s) = s + I_{[0,1/3]}(s), X_5(s) = s + I_{[1/3,2/3]}(s), X_6(s) = s + I_{[2/3,1]}(s),$$
$$\ldots$$

Then, for every $s \in S,$ the value $X_n(s)$ alternates between the value of s and $s+1$ infinitely often. Therefore, there is no value of $s \in S$ for which $X_n(s) \to s = X(s)$. In other words,

altuhough X_n converges in probability to X,

X_n does **NOT** converge almost surely to X.

几乎处处收敛蕴含依概率收敛

- 定理：如果 $X_n \to X, a.s.$, 则 $X_n \xrightarrow{p} X.$
- 证明：对于任何 $\epsilon > 0$, 事件 $\cap_{n=1}^{\infty} \cup_{k=n}^{\infty} \{|X_k - X| \geq \epsilon\}$

表示有无穷多个 $\{|X_k - X| \geq \epsilon\}$ 发生，因此由 $X_n \to X, a.s.$ 得

$$P(\cap_{n=1}^{\infty} \cup_{k=n}^{\infty} \{|X_k - X| \geq \epsilon\}) = 0$$

而事件 $A_n = \cup_{k=n}^{\infty}\{|X_k - X| \geq \epsilon\}$ 是n的单调递减序列，所以利用

$$\{|X_n - X| \geq \epsilon\} \subset \cup_{k=n}^{\infty}\{|X_k - X| \geq \epsilon\}$$

和概率的连续性得到

$$\begin{aligned}\lim_{n \to +\infty} P(|X_n - X| \geq \epsilon) &\leq \lim_{n \to +\infty} P(\cup_{k=n}^{\infty}\{|X_k - X| \geq \epsilon\}) \\ &= P(\cap_{n=1}^{\infty} \cup_{k=n}^{\infty} \{|X_k - X| \geq \epsilon\}) = 0\end{aligned}$$

Convergence in Distribution

A sequence of random variables $X_1, \ldots, X_n$, **converges in distribution** to a random variable X if,
for every $\varepsilon > 0$,

$$\lim_{n \to \infty} F_{X_n} = F_X(x)$$

at all points x where $F_X(x)$ is continuous.

If the sequence of random variables, $X_1, X_2, \ldots$ converges in probability to a random variable X, the sequence also converges in distribution to X.

Example

Let $X_1, X_2, \cdots$ be iid *uniform*$(0,1)$ random variables. Let $X_{(n)} = \max_{1\le i\le n} X_i$.
As $n \to \infty$, $X_{(n)}$ gets close to 1, but must necessarily be less than 1. Therefore

$$
\begin{aligned}
P(|X_{(n)} - 1| \ \ge \varepsilon) &= \underbrace{P(X_{(n)} \ge 1+\varepsilon)}_{=0} + P(X_{(n)} \le 1-\varepsilon) \\
&= P(X_{(n)} \le 1-\varepsilon).
\end{aligned}
$$

However,

$$
\begin{aligned}
P(X_{(n)} \le 1-\varepsilon) &= P(\max_{1\le i\le n} X_i \le 1-\varepsilon) \\
&= P(X_i \le 1-\varepsilon, i = 1,\ldots,n) \\
&= (1-\varepsilon)^n \\
&\to 0, \text{ as } n \to \infty.
\end{aligned}
$$

Therefore, $X_{(n)}$ converges to 1 in probability.

Example (continued)

Furthermore, let $\varepsilon = t / n,$ then

$$P(X_{(n)} \leq 1 - t / n) = (1 - t / n)^n \to e^{-t},$$

that is

$$P(n(1 - X_{(n)}) \leq t) \to 1 - e^{-t}.$$

Recall the *exponential*(1) distribution.

$$f(x) = e^{-x},$$

$$F(x) = \int_0^x e^{-t} dt = -e^{-t}\Big|_0^x = 1 - e^{-x}.$$

Hence,

$n(1 - X_{(n)})$ converges in distribution to an exponential random variable.

依概率收敛蕴含依分布收敛

- 定理：如果 $X_n \xrightarrow{p} X$, 则 $X_n \xrightarrow{d} X$.

证明：对于F的连续点x, 取 $\delta > 0, x_0 = x - \delta, x_1 = x + \delta$

$$
\begin{aligned}
F_n(x) - F(x) &= Pr(X_n \le x) - F(x) \\
&= Pr(X_n \le x, X > x_1) + Pr(X_n \le x, X \le x_1) - F(x) \\
&\le Pr(|X_n - X| > \delta) + F(x_1) - F(x)
\end{aligned}
$$

$$
\begin{aligned}
F(x) - F_n(x) &= [1 - Pr(X > x) - [1 - Pr(X_n > x)] \\
&= Pr(X_n \le x) - Pr(X > x) \\
&= Pr(X_n \ge x, X \le x_0) + Pr(X_n \ge x, X > x_0) - Pr(X > x) \\
&\le Pr(|X_n - X| > \delta) + F(X > x_0) - Pr(X > x) \\
&= Pr(|X_n - X| > \delta) + F(x) - F(x_0)
\end{aligned}
$$

依概率收敛蕴含依分布收敛

综上有

$$|F_n(x) - F(x)| \leq 2Pr(|X_n - X| \geq \delta) + F(x_1) - F(x_0)$$

当 $n \to +\infty$ 时得到

$$\overline{\lim_{n\to+\infty}} |F_n(x) - F(x)| \leq F(x_1) - F(x_0)$$

令 $\delta \to 0$，由F的连续性有

$$\overline{\lim_{n\to+\infty}} |F_n(x) - F(x)| = 0$$

Markov Inequality

Let X be a random variable and let $g(x)$ be a nonnegative function. Then for any $r > 0$

$$P(g(X) \geq r) \leq \frac{\mathrm{E}g(X)}{r}.$$

$$\begin{aligned}
\mathrm{E}g(X) &= \int_{-\infty}^{\infty} g(x)f(x)dx \\
&\geq \int_{\{x:g(x)\geq r\}} g(x)f(x)dx \\
&\geq r\int_{\{x:g(x)\geq r\}} f(x)dx \\
&= rP(g(X) \geq r)
\end{aligned}$$

Chebychev's Inequality

- 如果取

$$g(X) = (X - E(X))^2$$

- 就有Chebychev不等式

$$Pr(|X - E(X)|^2 \geq \epsilon^2) \leq \frac{Var(X)}{\epsilon^2}$$

Weak Law of Large Numbers (WLLN)

Let $X_1,\dots,X_n$ be iid random variables with $\mathrm{E}X_i = \mu$ and $\mathrm{Var}X_i = \sigma^2 < \infty$. Define $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$. Then,

$$P(|\bar{X}_n - \mu| \geq \varepsilon) = \underbrace{P(|\bar{X}_n - \mu|^2 \geq \varepsilon^2) \leq \frac{\mathrm{E}(\bar{X}_n - \mu)^2}{\varepsilon^2}}_{\text{Chebychev's inequality}} = \frac{\mathrm{Var}\bar{X}_n}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

Hence,

$$P(|\bar{X}_n - \mu| < \varepsilon) = 1 - P(|\bar{X}_n - \mu| \geq \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2} \to 1, \text{ as } n \to \infty.$$

In other words

$$\lim_{x \to \infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1.$$

Weak law of Large Numbers

Let $X_1, \ldots, X_n$ be iid random variables with $\mathrm{E}X_i = \mu$ and $\mathrm{Var}X_i = \sigma^2 < \infty$. Define $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$. Then, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1;$$

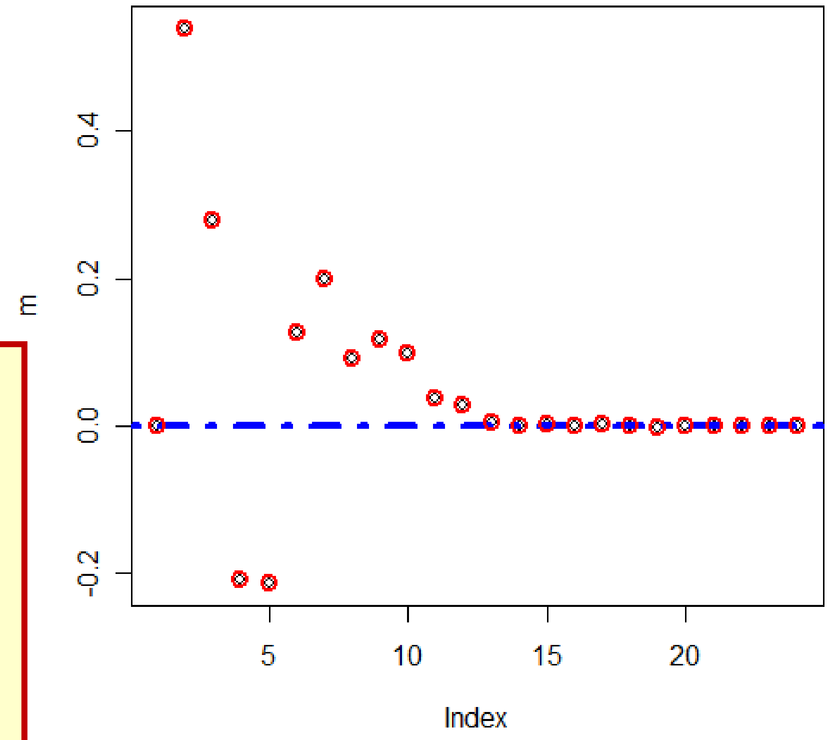
that is, $\bar{X}_n$ converges in probability to μ.

Sample mean becomes population mean when the sample size tends to infinity.

A simulation Study of WLLN

```
t <- 10000000;
x <- rnorm(t, 0, 1);
m <- 0;

n <- 2;
while(n < t){
        m <- c(m, mean(x[1:n]));
        n = n * 2;
}
```

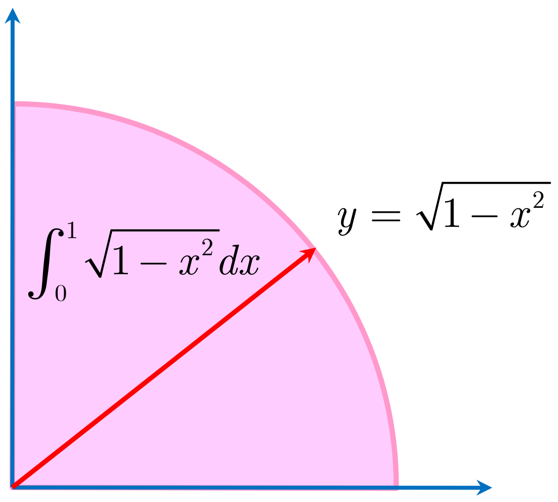


Monte Carlo Integration

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathrm{E}_{p(x)} f(X)\right| < \varepsilon\right) = 1 \Rightarrow \mathrm{E}_{p(x)} f(X) \approx \frac{1}{n}\sum_{i=1}^{n} f(X_i),\ \text{as } n \to \infty$$

$$\mathrm{E}_{p(x)}\left[f(X)\right] = \int_{-\infty}^{\infty} f(x)p(x)dx \Rightarrow \underbrace{\int_{-\infty}^{\infty} h(x)dx = \int_{-\infty}^{\infty} f(x)p(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)}_{\text{Monte Carlo integration}}$$

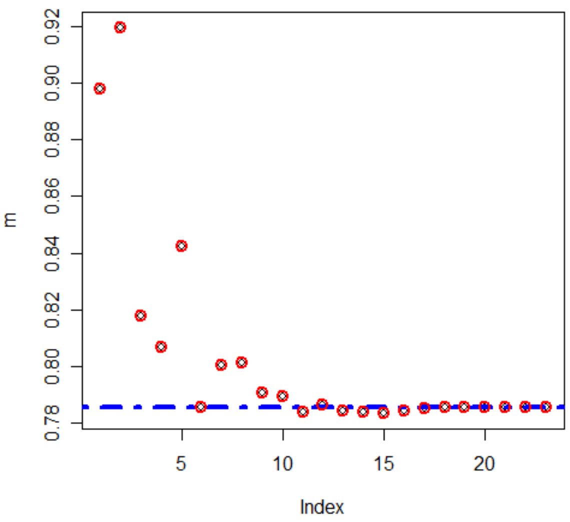
$$\int_0^1 \sqrt{1-x^2}\,dx = \frac{1}{2}\left(x\sqrt{1-x^2} + \arcsin x\right)\Big|_0^1 = \frac{\pi}{4}$$



$$\int_0^1 \sqrt{1-x^2}\,dx = \int_0^1 \underbrace{\sqrt{1-x^2}}_{f(x)} \cdot \underbrace{1}_{p(x)}\,dx$$

$$\approx \frac{1}{n}\sum_{i=1}^{n}\sqrt{1-x_i^2},$$

where every x_i is sampled from a uniform(0,1) distribuiton.



Convergence of Sample Variance

Let $X_1, \ldots, X_n$ be iid random variables with $\mathrm{E}X_i = \mu$ and $\mathrm{Var}X_i = \sigma^2 < \infty$. Define

$$\bar{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Then, for every $\varepsilon > 0$,

$$P(|\bar{S}_n^2 - \sigma^2| \geq \varepsilon) \leq \frac{\mathrm{E}(\bar{S}_n^2 - \sigma^2)^2}{\varepsilon^2} = \frac{\mathrm{Var}\bar{S}_n^2}{\varepsilon^2}.$$

So, if $\mathrm{Var}\bar{S}_n^2 \to 0$, then $\bar{S}_n^2$ converges to σ^2 in probability.

Strong law of Large Numbers (SLLN)

Let $X_1,\ldots,X_n$ be iid random variables with $\mathrm{E}X_i = \mu$ and $\mathrm{Var}X_i = \sigma^2 < \infty$. Define $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$.
Then, for every $\varepsilon > 0$,

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

that is, $\bar{X}_n$ converges almost surely to μ.

The Central Limit Theorem (CLT)

The central limit theorem

Let $X_1,\ldots,X_n$ be iid random variables with $\mathrm{E}X_i = \mu$ and $\mathrm{Var}X_i = \sigma^2 < \infty$. Define $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$. Then, for any x, $-\infty < x < \infty$,

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

$\sqrt{n}(\bar{X}_n - \mu)/\sigma$ has a limiting standard normal distribution.

The distribution of normalized sample mean becomes standard normal distribution when sample size tends to infinity.

Normal Approximation of Binomial

- Bernoulli trial
$$X_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

$$Y_i = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, p)$$

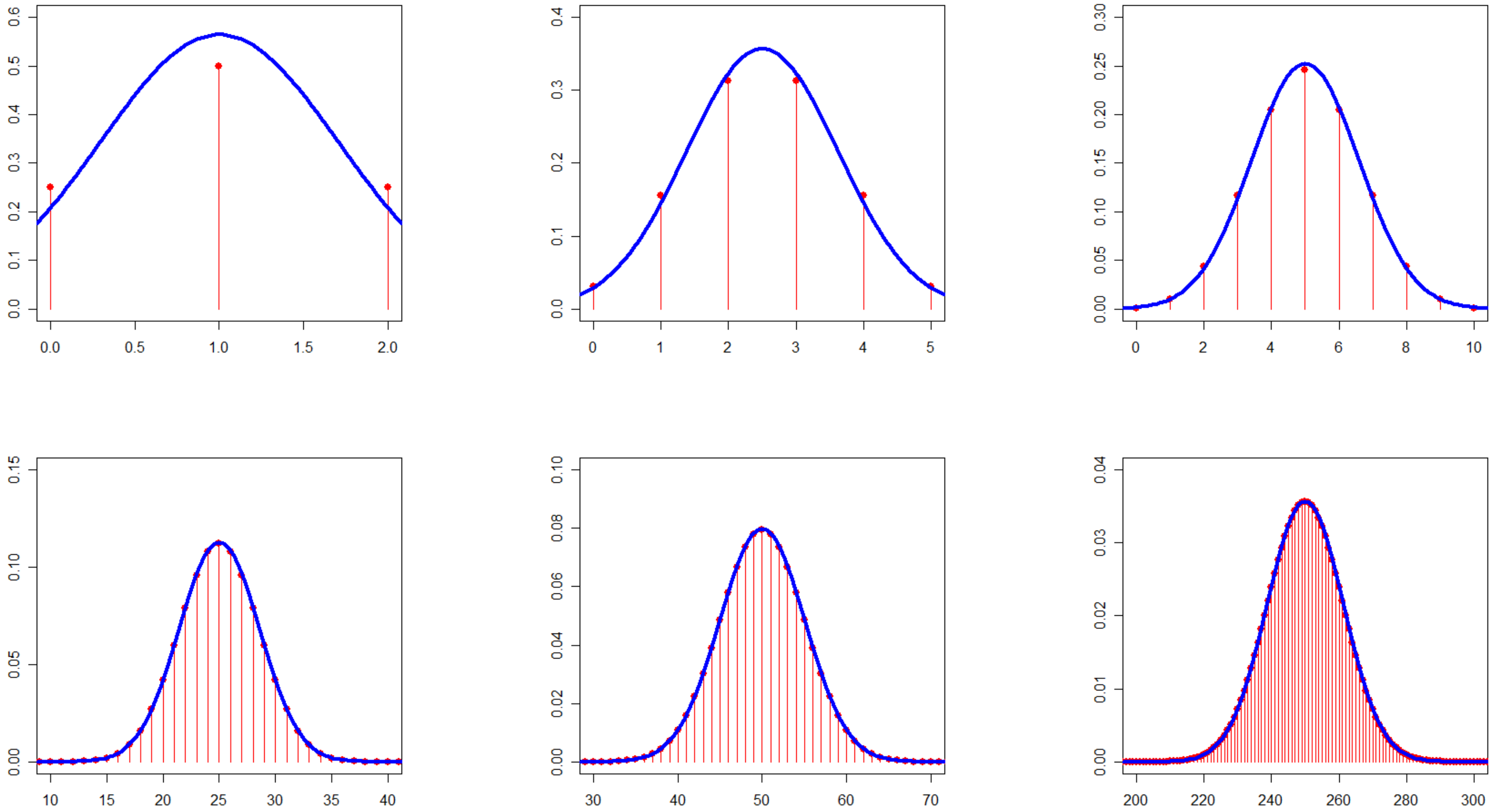
- Multiple Bernoulli trials
  - From concept

$$Z_n = \frac{Y_i - np}{\sqrt{np(1-p)}} = \frac{Y_i \,/\, n - p}{\sqrt{p(1-p) \,/\, n}} \sim N(0,1), \text{ as } n \to \infty$$

  - From the central limit theorem

$$Y_i \sim N\left(np, np(1-p)\right), \text{ as } n \to \infty$$

Normal Approximation of Binomial



Proof (矩母函数存在)

Define $Y_i = (X_i - \mu)/\sigma$, and let $M_Y(t)$ denote the common mgf of the Y_is.

Since

$$\frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} = \frac{(1/n)\Sigma_{i=1}^n X_i - \mu}{\sigma / \sqrt{n}} = \frac{\Sigma_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mu}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i,$$

We have

$$M_{\sqrt{n}(\bar{X}_n - \mu)/\sigma}(t) = M_{(1/\sqrt{n})\Sigma_{i=1}^n Y_i}(t) = M_{\Sigma_{i=1}^n Y_i}(t/\sqrt{n}) = [M_Y(t/\sqrt{n})]^n.$$

Because if $Y = aX + b$, then $M_Y(t) = e^{tb} M_X(at)$;

if $Y = X_1 + \cdots + X_n$, then $M_Y(t) = [M_X(t)]^n$.

Now, define

$$M_Y^{(k)}(0) = \frac{d^k}{dt^k} M_Y(t)\Big|_{t=0},$$

We can expand $M_Y(t/\sqrt{n})$ in a Taylor series around 0, as

$$M_Y\left(\frac{t}{\sqrt{n}}\right) = \sum_{k=0}^{\infty} M_Y^{(k)}(0) \frac{(t/\sqrt{n})^k}{k!} = 1 + \frac{(t/\sqrt{n})^2}{2!} + R_Y\left(\frac{t}{\sqrt{n}}\right)$$

because $M_Y^{(0)}(0) = 1, M_Y^{(1)}(0) = \mu_Y = 0, M_Y^{(2)}(0) = \sigma_Y^2 + \mu_Y^2 = 1.$

Proof

Now, there is a Taylor's theorem says that

if $g^{(r)}(a) = \frac{d^r}{dx^r} g(x)\Big|_{x=a}$ exists, then $\lim_{x \to a} \frac{g(x) - T_r(x)}{(x-a)^r} = 0.$

Apply the theorem to

$$R_Y\left(\frac{t}{\sqrt{n}}\right) = \sum_{k=3}^{\infty} M_Y^{(k)}(0) \frac{(t/\sqrt{n})^k}{k!} = g(x) - T_2(x),$$

yielding

$$\lim_{n \to \infty} \frac{R_Y(t/\sqrt{n})}{(t/\sqrt{n})^2} = 0, \text{ for any fixed } t \neq 0.$$

Since t is fixed, we have further

$$\lim_{n \to \infty} \frac{R_Y(t/\sqrt{n})}{(1/\sqrt{n})^2} = \lim_{n \to \infty} n R_Y\left(\frac{t}{\sqrt{n}}\right) = 0,$$

which is also true for $t = 0$ since $R_Y(0) = 0$.

Proof

Therefore,

$$\begin{aligned}
\lim_{n\to\infty}\left[M_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n &= \lim_{n\to\infty}\left[\sum_{k=0}^{\infty} M_Y^{(k)}(0)\frac{(t/\sqrt{n})^k}{k!}\right]^n \\
&= \lim_{n\to\infty}\left[1+\frac{(t/\sqrt{n})^2}{2!}+R_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n \\
&= \lim_{n\to\infty}\left[1+\frac{1}{n}\left(\frac{t^2}{2}+nR_Y\left(\frac{t}{\sqrt{n}}\right)\right)\right]^n \\
&= e^{t^2/2},
\end{aligned}$$

which is the standard normal mgf.

The last equality comes from the following theorem:

let $a_1, a_2, \ldots,$ be a sequence of numbers converging to a, that is $\lim_{n\to\infty} a_n = a$. Then

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

中心极限定理

- 定理5.5.15：设 $X_1, X_2, \cdots$ ，是一列独立同分布随机变量，且$E(X_i) = \mu, 0 < \mathrm{Var}(X_i) = \sigma^2 < +\infty,$令$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,\ G_n(x)$ 为 $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ 的累积分别函数，则对任意x 都有

$$\lim_{n \to +\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$$

即$\sqrt{n}(\overline{X}_n - \mu)/\sigma$ 依分布收敛到标准正态分布随机变量

证明思路

- 证明：基本上同前面利用矩母函数的证明方式，只不过将矩母函数替换成特征函数。矩母函数不一定存在，但任意分布的特征函数总是存在。

- 特征函数将随机变量的和转化为特征函数乘积，容易计算。

Slutskey's Theorem

The Slutskey's theorem

If $X_n \to X$ in distribution and $Y_n \to a$, a constant, in probability, then

$$X_n Y_n \to aX \text{ in distribution, and}$$

$$X_n + Y_n \to X + a \text{ in distribution.}$$

Since $Z_n = \dfrac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \to N(0,1)$ in distribution, if $\dfrac{\sigma}{S_n} \to 1$ in probability (need to proof), we have

$$\frac{\bar{X}_n - \mu}{S_n / \sqrt{n}} = \frac{\sigma}{S_n} \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \to N(0,1) \text{ in distribution.}$$

Limits of the *Student's t* Distribution

Student's t pdf

$$f(x \mid p) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)} \frac{1}{\sqrt{p\pi}} \frac{1}{(1+x^2 / p)^{(p+1)/2}}, -\infty < x < \infty, p = 1, \ldots$$

Standard normal pdf

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, -\infty < x < \infty$$

What is the relation? Can you prove that

$$\lim_{p \to \infty} \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)} \frac{1}{\sqrt{p\pi}} \frac{1}{(1+x^2 / p)^{(p+1)/2}} = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

§5.5 Delta方法

- 本节希望给出随机变量函数的抽样分布

- 方法：利用泰勒展开进行近似

多变量实值函数

- 胜算(Odds ratio): 设 $X_1, \cdots, X_n \sim Ber(p)$, 人们往往关心胜算率

$$g(X) = \frac{\hat{p}}{1-\hat{p}} = \frac{\overline{X}}{1-\overline{X}}$$

- 例子：设两个随机变量X与Y的均值非0，函数为他们的均值之比值

$$g(X,Y) = \frac{X}{Y}$$

Taylor展开

- 函数g(x)在点a处的r阶Taylor多项式

$$T_r(x) = \sum_{k=0}^{r} \frac{g^{(k)}(a)}{k!}(x-a)^k$$

- Taylor近似：如果函数g(x)在x=a处存在r阶导数，则

$$\lim_{x \to a} \frac{g(x) - T_r(x)}{(x-a)^r} = 0$$

- 即g(x)在x=a附近可以由r阶Taylor多项式近似.

实值函数Taylor近似

- 设随机变量 $X_1, \cdots, X_k$ 的期望分别是$\theta_1, \cdots, \theta_k$,
将函数在变量的期望处做一阶近似

$$g(x_1, \cdots, x_k) \approx g(\theta_1, \cdots, \theta_k) + \sum_{i=1}^{k} \frac{\partial g}{\partial x_i}(\theta)(x_i - \theta_i)$$

- 于是

$$Eg(X_1, \cdots, X_k) = g(\theta)$$

$$Varg(X_1, \cdots, X_k) = \sum_{i=1}^{k} [\frac{\partial g}{\partial x_i}(\theta)]^2 Var(X_i) + 2\sum_{i<j} \frac{\partial g}{\partial x_i}(\theta)\frac{\partial g}{\partial x_j}(\theta) Cov(X_i, X_j)$$

Delta方法

- 定理5.5.24：设随机变量序列Y_n满足：$\sqrt{n}(Y_n-\theta)$依分布收敛到正态分布$N(0,\sigma^2)$, 函数g在指定点θ满足 $g'(\theta)\neq 0$, 则

$$\sqrt{n}[g(Y_n)-g(\theta)]\to N(0,\sigma^2[g'(\theta)]^2).$$

- 证明思路：g(x)在θ处展开到一阶近似

$$g(Y_n)-g(\theta)=g'(\theta)(Y_n-\theta)+o(Y_n-\theta)$$

二阶Delta方法

- 定理5.5.26：设随机变量序列Y_n满足：$\sqrt{n}(Y_n-\theta)$ 依分布收敛到正态分布 $N(0,\sigma^2)$, 函数g在指定点 θ 满足 $g'(\theta)=0, g''(\theta)\neq 0.$ 则

$$n[g(Y_n)-g(\theta)] \to \sigma^2 \frac{g''(\theta)}{2}\chi_1^2.$$

- 证明思路：g(x)在 θ 处展开到二阶近似

$$g(Y_n)-g(\theta)=\frac{g''(\theta)}{2}(Y_n-\theta)^2+o(Y_n-\theta)^2$$

多元Delta方法

- 定理5.5.28：设高维随机向量序列 $\vec{X}_1, \cdots, \vec{X}_n$ 满足：$E(X_{ij}) = \mu_i, cov(X_{ik}, X_{jk}) = \sigma_{ij}$. 函数g连续一阶偏导数，且在指定点 $\mu = (\mu_1, \cdots, \mu_p)$ 处满足

$$\tau^2 = \sum\sum \sigma_{ij} \frac{\partial g(\mu)}{\partial \mu_i} \frac{\partial g(\mu)}{\partial \mu_j} > 0$$

则

$$\sqrt{n} g(\overline{X}_1, \overline{X}_2, \cdots, \overline{X}_p) - g(\mu_1, \cdots, \mu_p)] \to N(0, \tau^2)$$

随机变量生成及其应用

- 伪随机数生成
- 直接方法
- Monte Carlo方法
- Monte Carlo近似计算

Monte Carlo方法

- Von Neumann
- S.Ulam (1946)
- N.Metropolis (1953)
- Hasting (1975)

- Monte Carlo (Monaco), 著名赌城

Monte Carlo方法特点

- 概率模型：所求问题的解是模型参数或特征量；
- 抽样：根据概率模型进行随机抽样或模拟随机过程；
- 估计：多次抽样给出参数的估计以及参数的统计特性，给出解的近似值。

线性同余法
Linear congruence generator (LCG)

- One of the earliest and fastest algorithm:

$$X_{n+1} = (aX_n + c) \mod M$$

where $0 \leq X_n < M$ M is the modulus, a is multiplier, c is increment. All of them are integers. Choice of a, c, M must be done with special care.

http://random.mat.sbg.ac.at/links/rando.html

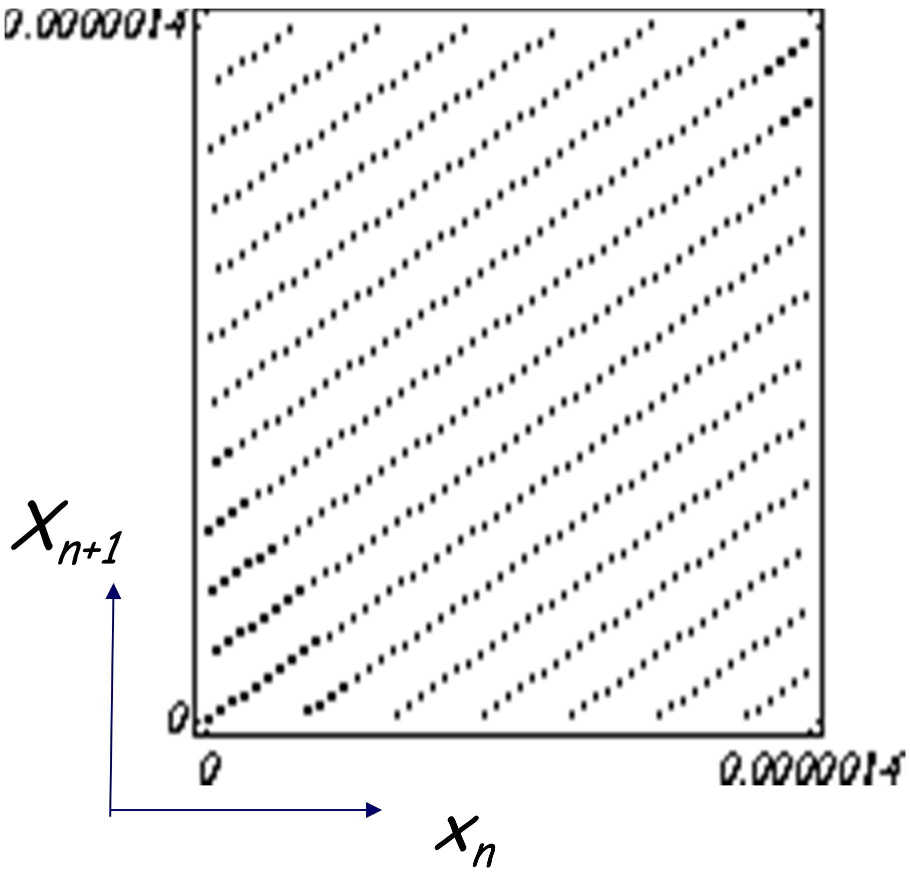
Choice of Parameters

Source	m	a	c	output bits of seed in *rand()* / *Random(L)*
Numerical Recipes	2^{32}	1664525	1013904223	
Borland C/C++	2^{32}	22695477	1	bits 30..16 in *rand()*, 30..0 in *lrand()*
glibc (used by GCC)[4]	2^{32}	1103515245	12345	bits 30..0
ANSI C: Watcom, Digital Mars, CodeWarrior, IBM VisualAge C/C++	2^{32}	1103515245	12345	bits 30..16
Borland Delphi, Virtual Pascal	2^{32}	134775813	1	bits 63..32 of *(seed * L)*
Microsoft Visual/Quick C/C++	2^{32}	214013	2531011	bits 30..16
RtlUniform from Native API [5]	$2^{31}-1$	2147483629	2147483587	
Apple CarbonLib	$2^{31}-1$	16807	0	see MINSTD
MMIX by Donald Knuth	2^{64}	6364136223846793005	1442695040888963407	
VAX's **MTH$RANDOM**,[6] old versions of glibc	2^{32}	69069	1	
Random class in Java API	2^{48}	25214903917	11	bits 47...16
LC53[7] in Forth (programming language)	$2^{32}-5$	$2^{32}-333333333$	0	

$$x_{n+1} = (ax_n + c) \mod M$$

http://en.wikipedia.org/wiki/Linear_congruential_generator

LCG方法的缺点



When (x_n, x_{n+1}) pairs are plotted for all *n*, a lattice structure is shown.

现代随机数产生器

- Mersenne Twister (MT19937)

  Extremely long period (2^{19937}-1), fast

- 逆同余随机数产生器

$$X_{n+1} = (ax_n^{-1} + c) \mod M$$

  nonlinear, no lattice structure

http://en.wikipedia.org/wiki/Mersenne_twister
http://en.wikipedia.org/wiki/Inversive_congruential_generator

直接方法-多点分布

- 设有多点分布

$$X \sim \{p_1, p_2, \cdots, p_n\}$$

- 只要将 [0 , 1] 区间分割为 n 份, 使得各份的长度分别为 $p_1, \cdots, p_n$.
- 产生随机变量 $u \sim U[0, 1]$.
- 若 u 落在上述区间的第 k 份， 取 X=k.

概率积分变换

- Let F(*x*) be the cumulative distribution function of a random variable *X*, then *x* can be generated from

$$x = F^{-1}(\xi)$$

- where $\xi \sim U(0,1)$, and $F^{-1}(x)$ is the inverse function of F(*x*).

例1：指数分布

- Density function,

$$p(x) = \exp(-x), x \geq 0.$$

- Distribution function,

$$F(x) = \int_0^x \exp(-y)dy = 1 - \exp(-x)$$

- So we generate *x* by

$$x = -\log(\xi), \quad \xi \sim U(0,1)$$

例2：二维标准正态分布

- Take 2D gaussian distribution

$$p(x,y)dxdy = \frac{1}{2\pi}\exp(-\frac{x^2+y^2}{2})dxdy$$

- Work in polar coordinates,

$$p(r,\theta)rdrd\theta = \frac{1}{2\pi}\exp(-\frac{r^2}{2})rdrd\theta$$

Box-Muller Method

- The formula implies that the variable ϑ is distributed uniformly between 0 and 2π, $\frac{1}{2}r^2$ is exponentially distribution, we have

$$\begin{cases} x = \sqrt{-2\log(\xi_1)}\cos(2\pi\xi_2) \\ y = \sqrt{-2\log(\xi_1)}\sin(2\pi\xi_2) \end{cases}$$

ξ_1 and ξ_2 are two independent, uniformly distributed random numbers.

Von Neuman 取舍原则

- 对于比较复杂的或不常见的分布, 下面的 **Von Neuman 取舍原则** 提供了一个非常有效的方法。
- 若密度为p_0(x)的随机变量容易生成，现在要生成密度为p(x)的随机变量，而且

$$p(x) \le c p_0(x)$$

Von Neuman 取舍原则

- 生成密度为$p_0(x)$的随机变量列

$$\{\xi_1, \xi_2, \cdots, \xi_n\}$$

- 对 $\{\xi_n\}$进行随机筛选, 留下的随机变量列就是密度为p(x)的随机变量列。
- 筛选方法：独立生成$U_1, \ldots, U_n \sim U(0,1)$，如果

$$U_k \geq \frac{p(x)}{cp_0(x)}$$

删除 ξ_k，否则保留 ξ_k.

Von Neuman 取舍原则的论证

- 留下的随机变量仍然是相互独立同分布的，它们中的每一个 (记为 η) 的分布是

$$
\begin{aligned}
& Pr(\eta < x) = Pr(\xi_k < x | \xi_k \text{ kept}) \\
& = Pr(\xi_k < x | U_k \leq \frac{p(\xi_k)}{cp_0(\xi_k)} \\
& = \frac{Pr\left(\xi_k < x, U_k \leq \frac{p(\xi_k)}{cp_0(\xi_k)}\right)}{Pr\left(U_k \leq \frac{p(\xi_k)}{cp_0(\xi_k)}\right)} \\
& = \frac{\int_{-\infty}^{x} \frac{p(y)}{cp_0(y)} p_0(y) dy}{\int_{-\infty}^{+\infty} \frac{p(y)}{cp_0(y)} p_0(y) dy} = \int_{-\infty}^{x} p(y) dy = F(y)
\end{aligned}
$$

马氏链的遍历性与遍历极限

- **马氏链的遍历性定理：**若有限状态Markov链的所有状态都是互通的，f是状态空间上的有界实值函数且满足 $\sum_i |f(i)|\pi_i < +\infty$，则

$$Pr\left(\frac{f(\xi_1)+\cdots+f(\xi_n)}{n} \to \sum_i f(\xi_i)\pi_i\right) = 1$$

其中π是MC的不变分布，而且此极限与初分布无关.

MCMC 算法的思想

- 对于非周期的MC (例如$p_{ij}>0$时)，我们还有

$$p_{ij}(n) \to \pi_j, \qquad n \to +\infty$$

- **Markov Chain Monte Carlo** 算法的思想就是：设计一个马氏链, 使得它的极限分布π与f成比例, 于是当我们模拟马氏链足够多步后, 它的分布就近似于π. 也正是因为有遍历性，我们就将此马氏链上的样本不加区分地当成是π的样本了。

Markov Chain Monte Carlo (MCMC)

- 设计一个 Markov 链，使其不变分布为我们关心的分布, (如高维分布, 或样本空间非常大的离散分布)。用这个 Markov 链的样本, 来对该分布作采样, 并用以作随机模拟。这样的方法, 统称为 **Markov Chain Monte Carlo (MCMC)**方法。
- 由于这种方法的问世, 使随机模拟在很多领域的计算中, 相对于决定性算法，显示出它的巨大的优越性。而有时随机模拟与决定性算法的结合使用, 会显出更多的长处.

Gibbs 采样法

- 生成一元随机变量是并不困难的， 但是生成高维各分量不独立的随机向量就非常困难。

- Gibbs 采样法的思想是**通过条件分布**得到以给定分布π为不变分布的马氏链的转移概率.

Gibbs采样法

- 这里$\pi(x)=\pi(x_1,x_2,\cdots,x_m)$是一个m-维分布密度, 相应的马氏链的状态是m-维向量, 其转移矩阵是(p_{xy}). 具体地, 取

$$p_{xy}=p(x,y)=\prod_{k=1}^{m}\pi(y_k|y_1,\cdots,y_{k-1},x_{k+1},\cdots,x_m)$$

- 意思是依次改变状态的m个分量, 每次只改变状态的一个分量, 使x变为y
- 容易验证：$\sum_x \pi(x)p_{xy}=\pi(y)$,即　$\pi\times(p_{xy})=\pi$

Gibbs 采样法的论证

$$
\begin{aligned}
&\sum_x \pi(x)p(x,y) = \sum_x \pi(x)\prod_{k=1}^{m}\pi(y_k|y_1,\cdots,y_{k-1},x_{k+1},\cdots,x_m) \\
&= \sum_{x_1,\cdots,x_m}\left[\sum_{x_1}\pi(x_1,\cdots,x_m)\right]\frac{\pi(y_1,x_2,\cdots,x_m)}{\sum_{x_1}\pi(x_1,\cdots,x_m)}\prod_{k=2}^{m}\pi(y_k|y_1,\cdots,y_{k-1},x_{k+1},\cdots,x_m) \\
&= \sum_{x_2,\cdots,x_m}\pi(y_1,x_2,\cdots,x_m)\prod_{k=2}^{m}\pi(y_k|y_1,\cdots,y_{k-1},x_{k+1},\cdots,x_m) \\
&= \cdots\cdots\cdots \\
&= \pi(y_1,\cdots,y_m) = \pi(y)
\end{aligned}
$$

构造 Markov 链轨道的方法

- 从已有的在时刻 n 的样本 ξ_n 去求时刻 $n+1$ 的样本 ξ_{n+1}
  1. 先由 $\boldsymbol{\xi_n}=(\xi_n^1, \xi_n^2, \ldots, \xi_n^s)$ 得到服从分布为

$$\pi(y_1|x_2,\cdots,x_m)=\frac{\pi(y_1,x_2,\cdots,x_m}{\pi(+\infty,x_2,\cdots,x_m)}$$

     的一元随机变量 ξ_{n+1}^1，其中 $x_k=\xi_n^k$
  2. 用同样的方法，再得到服从分布

$$\pi(y_2|y_1,x_3,\cdots,x_m)$$

     的样本 ξ_{n+1}^2，其中 $y_1=\xi_{n+1}^1$

构造 Markov 链轨道的方法

3. 依此下去…，得到 ξ_{n+1} 的所有分量.

4. 不断重复 1) － 3)， 我们就可以得到 上述马氏链的一 个样本轨道， 即一列随机向量 $\{\xi_n\}$, 其中 ξ_n 当 n 充分大时, 都可以近似地作为 π 的样本。

Metropolis采样法
(Metropolis sampler)

Metropolis 采样法概述 I

- 与Gibbs采样法一样，Metropolis方法也给出了在计算机上用 马氏链近似模拟遵从一个分布 $\boldsymbol{\pi}$ 的随机变量(向量) 的一个算法。
- Metropolis 提出了这种采样法, 称为**Metropolis**采样法。

Metropolis 采样法概述 II

- 它与 Gibbs 采样法的不同处在于, 对于 Metropolis采样法的转移概率如下：

$$p_{i,j} = \begin{cases} \overline{p}_{ij}\frac{\pi_j}{\pi_i} & \forall j \neq i, \\ 1 - \sum_{k \neq i} \overline{p}_{ik}\frac{\pi_k}{\pi_i} & j = i \end{cases}$$

Metropolis 采样法概述 III

- 其中 $\overline{P} = (\overline{p}_{ij})$ 是一个对称的互通转移矩阵, 称为**预选矩阵**，使用它是为了减少状态间的连接, 以加快 Markov 链的分布向不变分布收敛的速度。
- 由于预选矩阵是一个对称的互通转移矩阵, 所以

$$\sum_i \pi_i \left(\overline{p}_{ij} \frac{\pi_j}{\pi_i} \right) = \left(\sum_i \overline{p}_{ij} \right) \pi_j = \pi_j$$

Metropolis 采样法概述 IV

- **预选矩阵 $\overline{P} = (\overline{p}_{ij})$ 的选取：** 在许多情况下, 由于总的状态数非常多, 我们希望每次转移只能到达很少的几个状态, 但是保持不变分布仍然为 π.
- 我们通常选取非常稀疏的预选矩阵, 只要保证它是互通而且对称的转移概率阵就行了.
- 例如当 $\boldsymbol{\pi}$ 为多维分布, 从一个状态 $\boldsymbol{x}$ 出发,可规定它只能到达与它只有一个分量不同的状态.

Monte Carlo 近似计算

- 例: 求 π 的近似值
- 考虑在单位方块上均匀分布的二元随机变量 (ξ, η),

$$Pr(\xi^2 + \eta^2 < 1) = \frac{\pi}{4}$$

- 因此如果生成一组样本

$$(\xi_1, \eta_1), \cdots, (\xi_n, \eta_n),$$

- 其落在单位圆内的频率v_n就是 $\frac{\pi}{4}$ 的估计值

Monte Carlo 近似计算

- 利用 Chebychev 不等式, 立刻可以得到要保证误差不超过一个 给定值 ε 的概率小于 α , 应把 n 取多大

$$Pr\left(\left|v_n - \frac{\pi}{4}\right| \geq \epsilon\right) \leq \frac{D(v_1)}{n\epsilon^2}$$

Monte Carlo 积分近似计算

- Monte Carlo方法计算积分 $\int_a^b f(x)dx$
- g-样本方法: 设ξ的分布函数g(x)在f(x)非零点恒正，则

$$I=\int_a^b f(x)dx=\int_a^b \frac{f(x)}{g(x)}g(x)dx=E_g\left[\frac{f(x)}{g(x)}\right]$$

- 如果 $\xi_1,\cdots,\xi_n \sim g(x),$

$$I_n=\frac{1}{n}\left[\frac{f(\xi_1)}{g(\xi_1)}+\cdots+\frac{f(\xi_n)}{g(\xi_n)}\right]$$

Monte Carlo 积分近似计算

- 它是无偏估计

$$E_g(I_n) = \int_a^b f(x)dx$$

- 当g(x)=cf(x)时，估计的方差

$$Var(I_n) = \frac{1}{n}\left[\int_a^b \left(\frac{f(x)}{g(x)}\right)^2 g(x)dx - I^2\right]$$

最小(此结论可以由Schwartz不等式证明)

Monte Carlo 近似计算

- 但是上面的分布密度函数g(.)含未知数C, 其实它正是我们所要求的积分。所以上面的最优g-样本方法似乎并不可行。
- 注意到Gibbs 抽样中，只利用条件概率进行抽样，常数C可以被消去。此时可以抽样出g=Cf(x)的样本。