五、特殊的抽样方法

针对具体的问题，有时采用近似抽样方法是十分必要的。如果实验测量或理论计算得到的近似分布密度函数在抽样范围内是有界的，我们总是可以采用舍选法。但是这种方法对分布密度函数在抽样范围内起伏比较大时，其抽样效率很低。

部份近似抽样方法。

1．对由直方图给出的分布的抽样

一维直方图给出的分布反映了某一随机变量出现的频数。实际上是以图形形式给出随机变量在各道上的分布密度函数 $f(x)$ 和分布函数 $F(x)$ 的值。

如果随机变量在第 j 道内的频数为 n_j，则到该道的累积分布数为 $\sum_{i=1}^{j} n_i$，再假定抽样范围是从 1 道到 N 道，则在第 j 道上的分布函数值为

$$F(x_j) = \sum_{i=1}^{j} n_i \Big/ \sum_{i=1}^{N} n_i \quad .$$

它的抽样可以采用阶梯近似法，即抽取均匀分布随机数 ξ，找出满足不等式

$$F(x_{i-1}) \le \xi < F(x_i) \ .$$

的 i 值，把对应的 x_i 值作为抽样值，即取 $\eta = x_i$。

这种做法实际上就是用若干个前后相接的阶梯性函数值来近似 $F(x)$。

进一步作细致的考虑时，我们可以用线性插值法求出抽样值。从判断不等式决定出的 i，和 x_i 的值，求出

$$x_i' = x_{i-1} + \frac{\xi - F(x_{i-1})}{F(x_i) - F(x_{i-1})}(x_i - x_{i-1})$$

取 $\eta = x_i'$ 作为抽样值。

上述方法由于需要逐道地计算累计分布数 $F(x_i)$，来判断与随机数 ξ 值对应的满足不等式的 x_i 值，因而效率很低。

折半查找法是以计算最靠近 ξ 的 $F(x_{i-1})$ 和 $F(x_i)$ 的值，并求出线性插值来作为抽样值。这种方法可以提高抽样效率。

2．对由经验公式给出分布的抽样

当随机变量样本的一维分布密度函数是由平滑的经验公式 $f(x)$ 给出时，常用的技巧是采用如下方法：

（1） 将抽样区间划分为若干等份的子区间；然后在各个子区间内对分布密度函数积分；

（2） 计算出对应于各个区间的分布函数值，即 $F(j) = \sum_{i=1}^{j} \int_{x_{i-1}}^{x_i} f(x)dx$ ；

（3） 再采用与由直方图分布抽样中使用的相同办法来求出抽样值。

这种方法在求对应于各子区间的一组分布函数值时比较耗时，但依据这些数产生随机数时却相当快。

3． 反函数近似

设随机变量η以分布函数$F(x)$分布。采用直接抽样法，取$\eta=F^{-1}(\xi)$，则可以从均匀分布的随机变量抽样值ξ得到随机变量η的抽样值。

在实际抽样中，往往反函数$F^{-1}(y)$的解析形式求不出来，因而就用近似计算方法求得$F^{-1}(y)\approx Q(y)$。以$Q(y)$作为η的抽样近似值。这就是反函数近似。

假如$F^{-1}(y)$具有如下性质：$y\in[0,1]$，$\lim\limits_{y\to 0}F^{-1}(y)=-\infty$和$\lim\limits_{y\to 1}F^{-1}(y)\approx+\infty$，此时，可以利用最小二乘法拟合曲线$F^{-1}(y)$的函数。例如我们取

$$F^{-1}(y)\approx Q(y)=a+by+cy^2+\alpha(1-y)^2\ln y+\beta y^2\ln(1-y)\text{.}$$

这样的近似取法对相当广泛的分布函数抽样是可行的。其中系数α,β,a,b,c是待定参数。当然$Q(y)$也可以取其它数学表示形式，如帕迪(Pade)近似。

以标准正态分布$N(0,1)$为例，利用分布函数定义的公式，近似有

$$y=\int_{-\infty}^{F^{-1}(y)}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx\approx\int_{-\infty}^{Q(y)}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx\quad.$$

取点$y_k=\dfrac{k}{200}$，$(k=1,2,\ldots,199)$，即将[0，1]区间分成 200 等份，取区间内有 199 个点，得到

$$y_k=\int_{-\infty}^{Q(y)}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx,\qquad(k=1,2,\ldots,199)$$

$$Q(y_k) \approx \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

利用逐步回归法计算出各个系数为：

$a = -0.8268$ ， $b = 1.6736$， $c = 0$ ，

$\alpha = 0.3315$ ， $\beta = -0.3315$ 。

4． 近似修正抽样

对于任意已知的分布密度函数 $f(x)$，若 $f_1(x)$ 是 $f(x)$ 的一个近似分布密度函数，并且以 $f_1(x)$ 分布的抽样简单，运算量也小，则可以令

$$m = \min_{f_1(x) \neq 0} \frac{f(x)}{f_1(x)}$$

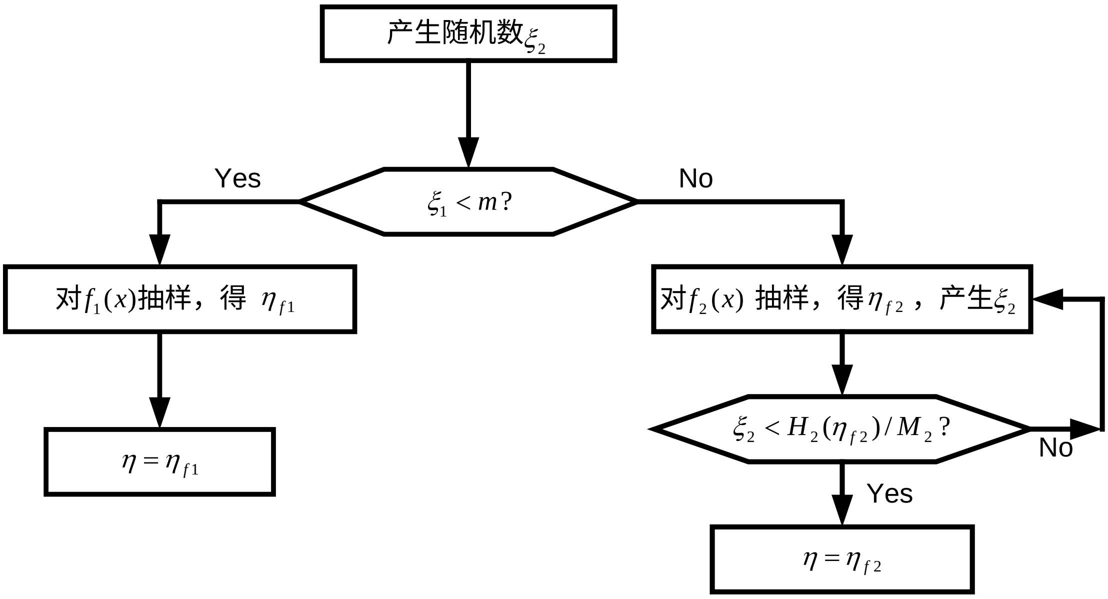
使分布密度函数可以表示成乘加分布抽样的分布形式

$$f(x) = mf_1(x) + H_2(x) f_2(x) .$$

其中 $H_2(x) f_2(x)$ 是对近似 $f(x) \approx mf_1(x)$ 的一个修正，即

$$H_2(x) f_2(x) = f(x) - mf_1(x) .$$

令 $M_2 = \max H_2(x)$，与乘加分布的公式比较，可以看到这里有 $H_1(x) = m$。这样我们就可以采用如下的抽样框图。



5．极限近似法

中心极限定理可以用来产生具有正态分布的随机变量抽样。

它利用任意分布的随机数的和来产生正态分布的抽样值。假如$\xi_1,\xi_2,\ldots,\xi_n$是在[0，1]区间上 n 个均匀分布的独立随机变量的抽样样本。它的平均值为$1/2$，方差为$1/12$。事实上，我们有

$$E\{\xi\} = \int_{-\infty}^{+\infty} x \cdot f(x)dx = \int_0^1 x \cdot 1dx = \frac{1}{2} \ .$$

$$V\{\xi\} = E\{\xi^2\} - [E\{\xi\}]^2 = \int_0^1 x^2 \cdot f(x)dx - \left(\frac{1}{2}\right)^2 = \frac{1}{12} \ .$$

设$R_n = \xi_1 + \xi_2 + \ldots + \xi_n$，则

$$E\{R_n\} = \int_{-\infty}^{+\infty} nx \cdot f(x)dx = \int_0^1 nx \cdot 1dx = \frac{n}{2} \ .$$

$$V\{R_n\} = E\{R_n^{\ 2}\} - [E\{R_n\}]^2 = \frac{n^2}{12} \ .$$

根据中心极限定理，引入新的随机变量δ_n，

$$\delta_n=\frac{R_n-\frac{n}{2}}{\sqrt{\frac{n}{12}}}\quad ,$$

则

$$\lim_{n\to\infty}p(\delta_n\le x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-t^2/2}dt=N(0,1)\ .$$

通常取$n=12$，此时随机变量δ_{12}为

$$\delta_{12}=R_{12}-6\quad .$$

这种抽样的方法称为极限近似法。但是要注意：如果取$n=12$，采用这种方法抽样时，则$|x|>6$的情况已经完全忽略。若要考虑$|x|<6$处的情况，必须取$n>12$或改用其它的抽样办法。

六、多维随机向量的抽样方法

1．舍选法

设随机向量变量$\vec{\eta}$的各分量为$\eta_1,\eta_2,\cdots,\eta_n$ ，

$$\vec{\eta}=(\eta_1\eta_2\eta_3\cdots\eta_n)^T\ .$$

它的联合分布密度函数为$f(x_1,x_2,\cdots,x_n)$，抽样范围在平行多面体

$$\{a_1\le x_1\le b_1,a_2\le x_2\le b_2,\cdots,a_n\le x_n\le b_n\}$$

内。令在该范围内，

$$L=\sup f(x_1,x_2,\cdots,x_n)<+\infty$$

我们将一维舍选法推广到这里，得到 n 维舍选法的做法如下：

（1）产生 n+1 个[0，1]上的均匀分布随机数$\xi_1,\xi_2,\cdots,\xi_{n+1}$，

（2）然后判断如下不等式

$$\xi_{n+1} < \frac{1}{L} F\left[(b_1 - a_1)\xi_1 + a_1, (b_2 - a_2)\xi_2 + a_2, \cdots, (b_n - a_n)\xi_n + a_n\right]$$ ．

是否成立。

（3）若不等式成立，则得到 $\vec{\eta}$ 的一个抽样值，该向量的各个分量值为 $\eta_i = (b_i - a_i)\xi_i + a_i$ ， $(i = 1, 2, \cdots, n)$

（4） 若不等式不成立，再重新产生 n+1 个随机数 ξ_i ，

（5） 重复上面的步骤，直至该不等式成立。这种方法的效率为 $E = \dfrac{1}{L\prod_{i=1}^{n}(b_i - a_i)}$ ．

显然这个抽样效率较低，而且 L 的计算也很困难。这就在很多情况下限制了它的使用。

2．条件密度法

设 $\vec{\eta} = (\eta_1, \eta_2)^T$ 的联合密度函数为 $f(x_1, x_2)$ ，若在某一特定的点 x_1 处，

$$\int_{-\infty}^{+\infty} f(x_1, x_2) dx_2 > 0$$ ．

则定义

$$f(x_2 \mid x_1) = f(x_1, x_2) / \int_{-\infty}^{+\infty} f(x_1, x_2) dx_2$$ ．

其中 $f(x_2 \mid x_1)$ 称为在 $\eta_1 = x_1$ 条件下，η_2 的条件分布密度函数。这时可以将 $f(x_1, x_2)$ 表示成

$$f(x_1, x_2) = f(x_2 \mid x_1) \cdot \int_{-\infty}^{+\infty} f(x_1, x_2) dx_2$$ ．

用类似的方法可以将三维随机向量的联合分布密度函数写为

$$f(x_1, x_2, x_3) = f_1(x_1) \cdot f_2(x_2 \mid x_1) \cdot f_3(x_3 \mid x_1, x_2)$$ ．

上面公式中，

$$\left.\begin{array}{r} f_1(x_1)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x_1,x_2,x_3)dx_2dx_1 \\ f_2(x_2\mid x_1)=\int_{-\infty}^{+\infty} f(x_1,x_2,x_3)dx_3 / f_1(x_1) \\ f_3(x_3\mid x_1,x_2)= f(x_1,x_2,x_3)/[f_1(x_1)\cdot f_2(x_2\mid x_1)] \end{array}\right\}$$

进一步推广到 n 维随机向量也是容易的。

$f(x_1,x_2,\cdots,x_n)= f_1(x_1)\cdot f_2(x_2\mid x_1)\cdot f_3(x_3\mid x_1,x_2)\cdots f_n(x_n\mid x_1,x_2,\cdots x_{n-1})$.

抽样步骤：

（1）由 $f_1(x_1)$ 为分布密度函数产生 η_1 的抽样值 $\eta_1=x_1$。

（2）在 $\eta_1=x_1$ 的条件下，由分布密度函数 $f_2(x_2\mid x_1)$ 抽取 $\eta_2=x_2$。

（3）在 $\eta_1=x_1$， $\eta_2=x_2$ 的条件下，由分布密度函数 $f_3(x_3\mid x_1,x_2)$ 抽取 $\eta_3=x_3$。

⋮

⋮

（n）在 $\eta_1=x_1$， $\eta_2=x_2$， $\cdots,\eta_{n-1}=x_{n-1}$ 的条件下，由分布密度函数 $f_n(x_n\mid x_1,x_2,\cdots,x_{n-1})$ 抽取 $\eta_n=x_n$。

最后就得到了 $\vec{\eta}=(\eta_1,\eta_2,\eta_3,\cdots,\eta_n)^T$ 的抽样值 $(x_1,x_2,x_3,\cdots,x_n)^T$。

例　中子入射角 (φ,θ) 服从联合分布密度函数

$$\begin{cases} f(\varphi,\theta)=\dfrac{1}{\alpha}(1+\sqrt{3}\sin\varphi\sin\theta)\sin\varphi\sin^2\theta, \\ \pi/2>\varphi\geq 0,0\leq\theta\leq\pi/2,\alpha=(3+2\sqrt{3})\pi/12. \end{cases}$$

α 为归一化常数，现要求对其余弦 $\eta=\cos\varphi,\delta=\cos\theta$ 做抽样。

解　容易证明对上式进行变量代换后，η 和 δ 联合分布密度函数为

$$f(x,y)=\frac{12}{(3+2\sqrt{3})\pi}\sqrt{1-y^2}\left[1+\sqrt{3}\sqrt{(1-x^2)(1-y^2)}\right]$$ ．

取

$$f_1(x)=\frac{12}{(3+2\sqrt{3})\pi}\left(\frac{\pi}{4}+\frac{2\sqrt{3}}{3}\sqrt{1-x^2}\right)$$ ．

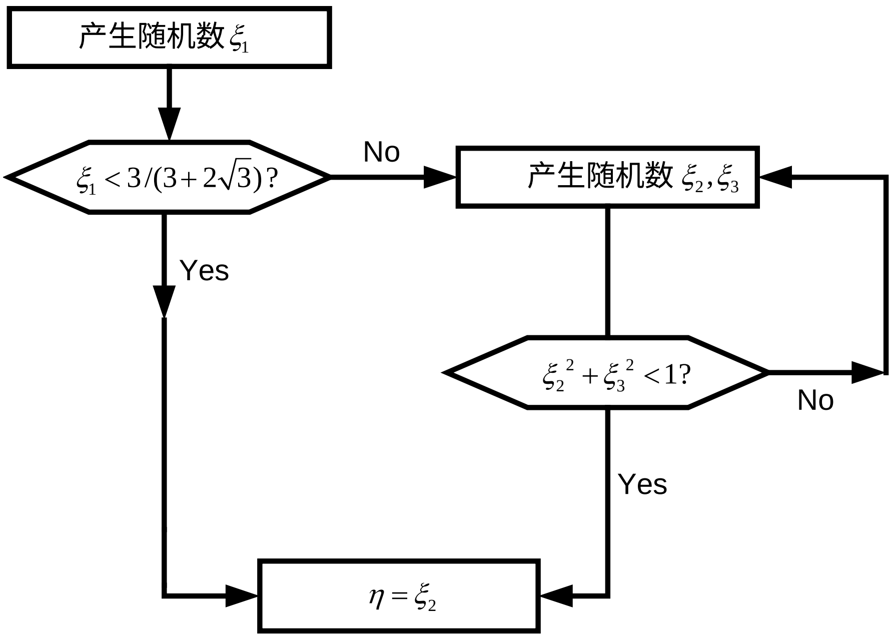
$$f_2(y|x)=\frac{1}{\frac{\pi}{4}+\frac{2\sqrt{3}}{3}\sqrt{1-x^2}}\sqrt{1-y^2}\left(1+\sqrt{3}\sqrt{(1-x^2)(1-y^2)}\right)$$ ．

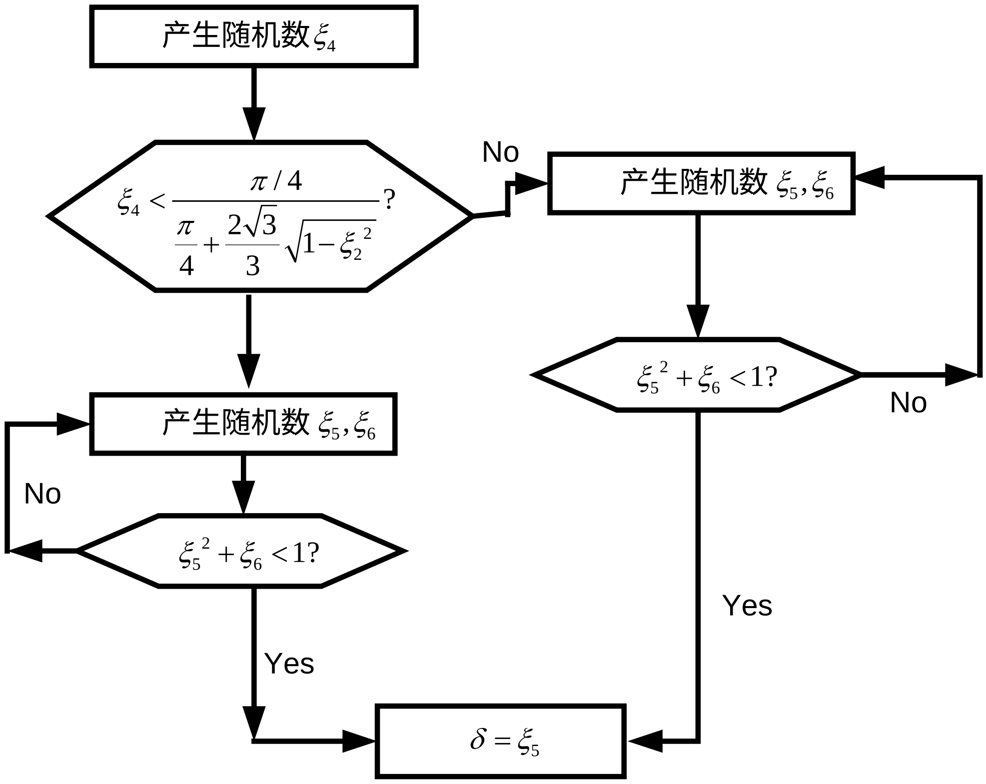
则 $f(x,y)=f_1(x)f_2(y|x)$。我们先对 $f_1(x)$ 抽样，将其化为

$$\begin{cases} f_1(x)=p_1+p_2\cdot\frac{4}{\pi}\sqrt{1-x^2} \\ p_1=\frac{3}{3+2\sqrt{3}},\quad p_2=\frac{2\sqrt{3}}{3+2\sqrt{3}} \end{cases}$$

用前面介绍过的加分布抽样，可以得到它的抽样框图。抽出 $\eta=\xi_2$ 后，再对 $f_2(y|\xi_2)$ 抽取 δ 值。这时同样可以使用加分布抽样法。

抽样框图





3．n 维正态分布随机向量的抽样

当 n 维随机向量$\vec{\eta}=(\eta_1,\eta_2,\eta_3,\cdots,\eta_n)^T$服从如下标准正态分布时，

$$f(x_1,x_2,\cdots,x_n)=\frac{1}{\sqrt{2\pi}}\exp\left\{-x_1^2/2\right\}\cdot\frac{1}{\sqrt{2\pi}}\exp\left\{-x_2^2/2\right\}\cdots\frac{1}{\sqrt{2\pi}}\exp\left\{-x_n^2/2\right\}$$

各分量是互相独立的。我们可以用一维变量正态分布的抽样法，对各分量分别抽取η_{x_i}，构成总体抽样值$\vec{\eta}_x=\left(\eta_{x_1},\eta_{x_2},\eta_{x_3},\cdots,\eta_{x_n}\right)^T$。

对 n 维正态分布的抽样可以在对 n 维标准正态分布抽样的基础上进行.如果 n 维随机向量$\vec{\eta}$服从的联合分布密度函数可以表示为如下的正态分布形式:

$$f(x_1,x_2,\cdots x_n)=(2\pi)^{-\frac{n}{2}}\cdot|M|^{-\frac{1}{2}}\cdot\exp\left\{-\frac{1}{2}(\vec{\chi}-\vec{\mu})^T M^{-1}(\vec{\chi}-\vec{\mu})\right\}\ .$$

其中

$$\vec{\chi}=\left(x_1,x_2,\cdots,x_n\right)^T,$$

$$\vec{\mu}=\left(\mu_1,\mu_2,\cdots,\mu_n\right)^T.$$

$\vec{\mu}=E\{\vec{\eta}\}$，即$\vec{\mu}$为$\vec{\eta}$的期望值。M 称为$\vec{\eta}$的协方差矩阵，它是正定对称的 n 阶方阵，其矩阵元σ_{ij}为

$$\sigma_{ij}=E\{(\eta_i-\mu_i)(\eta_j-\mu_j)\}=\sigma_{ji}.$$

因为 M 是正定对称的，所以总可以找到一个非奇异的下三角矩阵 A 。将 M 分解为

$$M=AA^T.$$

可以证明，一般 n 维正态分布的抽样值η_x，可以通过 n 维标准正态分布抽样值η_y，经过变换

$$\vec{\eta}_x=\vec{\mu}+A\vec{\eta}_y.$$

来得到。

例　二维正态分布的抽样

解　设$\vec{\eta}=(\eta_1,\eta_2)^T$服从二维正态分布，对其协方差矩阵

$$M=\begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}.$$

进行分解，以得到$M=AA^T$的形式。我们可以得到下三角矩阵 A 为

$$A=\begin{pmatrix}\sqrt{\sigma_{11}} & 0\\ \dfrac{\sigma_{12}}{\sqrt{\sigma_{11}}} & \sqrt{\dfrac{\sigma_{11}\sigma_{22}-\sigma_{12}^2}{\sigma_{11}}}\end{pmatrix}.$$

设$\vec{\eta}$的期望值为

$$\vec{\mu}=\left(\mu_1,\mu_2\right)^T.$$

若相应的二维标准正态分布已抽得$\vec{\eta}=\left(\eta_{y_1},\eta_{y_2}\right)^T$，则得到最后的

抽样结果$\vec{\eta}_x=\left(\eta_{x_1},\eta_{x_2}\right)^T$为

$$\begin{cases} \eta_{x_1}=\mu_1+\sqrt{\sigma_{11}}\eta_{y_1} \\ \eta_{x_2}=\mu_2+\dfrac{1}{\sqrt{\sigma_{11}}}\left[\sigma_{12}\eta_{y_1}+(\sigma_{11}\sigma_{22}-\sigma_{12}^2)^{1/2}\eta_{y_2}\right] \end{cases}.$$