由密度函数生成随机数的方法

一、 CDF 逆变换法

1.1 理论

假设随机变量 X 的密度函数为 $f(x)$，分布函数为 $F(x)$，分布函数的反函数为 $F^{-1}(x)$，假设 $u \sim U(0,1)$，那么 $x = F^{-1}(u)$ 为服从 $f(x)$ 的随机数.

这是因为 $P(F^{-1}(u) \leq x) = P(u \leq F(x)) = F_u(F(x)) = F(x)$.

1.2 示例

以生成 Gamma 分布的随机数为例，其密度函数为

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}$$

最终结果如图 1所示.

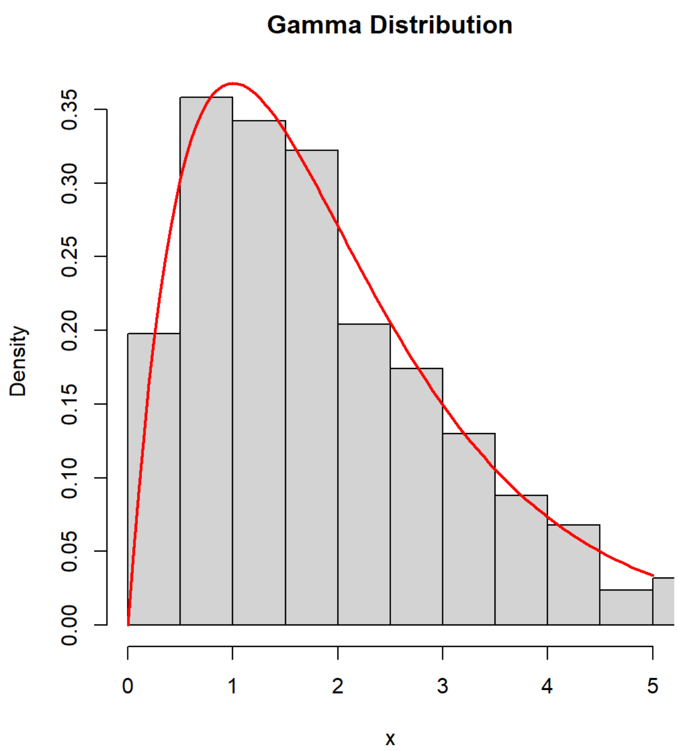

图 1　Gamma 分布随机数

```
# 定义伽玛分布的概率密度函数和累积分布函数
shape <- 2
```

```r
rate <- 1
f <- function(x) ifelse(x > 0, (rate^shape / gamma(shape)) * x^(shape - 1) * exp(-rate * x), 0)  # 伽玛分布的概率密度函数
F <- function(x) ifelse(x <= 0, 0, pgamma(x, shape, rate))  # 伽玛分布的累积分布函数

# 定义累积分布函数的反函数
inv_F <- function(u) qgamma(u, shape, rate)

# 生成均匀分布的随机数
u <- runif(1000)

# 映射到伽玛分布
x <- inv_F(u)

# 绘制直方图和真实的概率密度函数进行比较
hist(x, breaks = 30, freq = FALSE, xlim = c(0, 5), main = "Gamma Distribution", xlab = "x", ylab = "Density")
curve(f, from = 0, to = 5, col = "red", add = TRUE, lwd = 2)
```

二、 接受拒绝法（Acceptance-Rejection Method）

2.1 理论

Acceptance-Rejection Method 的具体过程如下：

- Step1：假设随机变量 X 的密度函数为 $f(x)$, 建议分布 G 的概率密度为 $g(x)$，找到一个常熟 c, 使得对所有的 x, 有 $c \cdot g(x) \geq f(x)$.
- Step2：从建议分布 G 抽样，得到样本 Y.
- Step3：从辅助分布 $U(0,1)$ 抽样，得到样本 U.
- Step4：如果 $U \leq \frac{f(Y)}{c \cdot g(Y)}$，则令 $X = Y$, 即接受 Y 作为样本，否则返回步骤 2 继续抽样.
- Step5：重复步骤 2 至 4 直至抽到足够样本.

Acceptance-Rejection Method 有效性的证明：

易知 $\frac{f(Y)}{c \cdot g(Y)}$ 和 U 是独立的，且 $0 < \frac{f(Y)}{c \cdot g(Y)} \leq 1$, 从建议分布和均匀分布中成功一次获得 X 的抽样（迭代）次数 N 也是个随机变量，服从概率是 p 的几何分布，其中

$$p = P(U \leq \frac{f(Y)}{c \cdot g(Y)}) = \int_{-\infty}^{+\infty} \frac{f(y)}{c \cdot g(y)} \times g(y) dy = \frac{1}{c}$$

我们需要证明 $P(Y \leq y | U \leq \frac{f(Y)}{c \cdot g(Y)}) = F(y)$, 利用贝叶斯公式显然有

$$\begin{aligned} P\left(Y \leqslant y \middle| U \leqslant \frac{f(Y)}{c * g(Y)}\right) &= P\left(U \leqslant \frac{f(Y)}{c * g(Y)} \middle| Y \leqslant y\right) \times \frac{G(y)}{1/c} \\ &= \frac{F(y)}{cG(y)} \times \frac{G(y)}{1/c} \\ &= F(y) \end{aligned}$$

其中

$$\begin{aligned} P\left(U \leqslant \frac{f(Y)}{c * g(Y)} \middle| Y \leqslant y\right) &= \frac{P\left(U \leqslant \frac{f(Y)}{c*g(Y)}, Y \leqslant y\right)}{G(y)} \\ &= \int_{-\infty}^{y} \frac{P\left(U \leqslant \frac{f(Y)}{c*g(Y)} \middle| Y = \omega \leqslant y\right)}{G(y)} g(\omega) d\omega \\ &= \frac{1}{G(y)} \int_{-\infty}^{y} \frac{f(\omega)}{cg(\omega)} g(\omega) d\omega \\ &= \frac{1}{cG(y)} \int_{-\infty}^{y} f(\omega) d\omega \\ &= \frac{F(y)}{cG(y)} \end{aligned}$$

2.2 示例

假设我们的目标概率密度函数为

$$f(x) = 2x^3 e^{-x^2}$$

对此分布生成样本。最终生成的随机数结果如图 2所示.

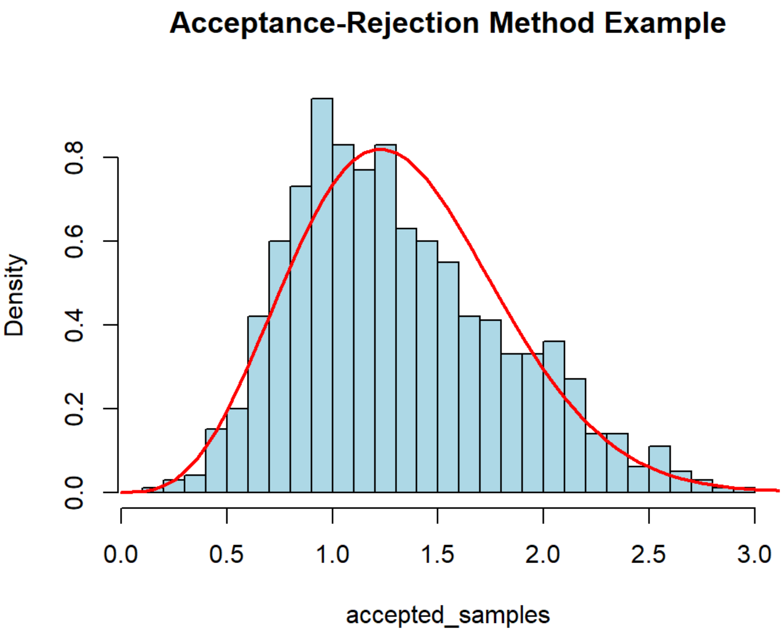

图 2　Acceptance-Rejection Method 示例

```
# 定义目标概率密度函数
f <- function(x) 2 * x^3 * exp(-x^2)

# 定义辅助概率密度函数g，这里选择一个简单的指数分布作为辅助函数
g <- function(x) dexp(x, rate = 1)

# 定义常数k，使得f(x) <= k * g(x) 对所有x都成立
k <- 2

# 生成随机数的数量
n <- 1000

# 初始化接受的样本
accepted_samples <- numeric(0)

# 使用接受-拒绝法生成样本
while(length(accepted_samples) < n) {
  # 从辅助分布中生成一个随机数
```

```r
  x <- rexp(1, rate = 1)
  # 从均匀分布中生成一个随机数
  u <- runif(1)
  # 接受条件
  if(u <= f(x) / (k * g(x))) {
    accepted_samples <- c(accepted_samples, x)
  }
}

# 绘制生成的样本与目标分布进行比较
hist(accepted_samples, breaks = 30, freq = FALSE, col = "lightblue", main = "Acceptance-Rejection Method Example")

# 绘制目标概率密度函数
curve(f, from = 0, to = 5, add = TRUE, col = "red", lwd = 2)
```

三、 内置函数

3.1 理论

在 R 语言中，可以利用内置函数可以直接生成一些常见分布的随机数. 部分随机数生成函数与分布对应表如表 1 所示.

表 1　随机数生成函数与分布对应表

函数	分布
runif	均匀分布
rnorm	正态分布
rexp	指数分布
rgamma	Gamma 分布
rgeom	几何分布
rhyper	超几何分布
rlogis	Logistic 分布
rmultinom	多项式分布
rpois	泊松分布
rt	t 分布

3.2 示例

以 rnorm 函数生成正态分布随机数为例.

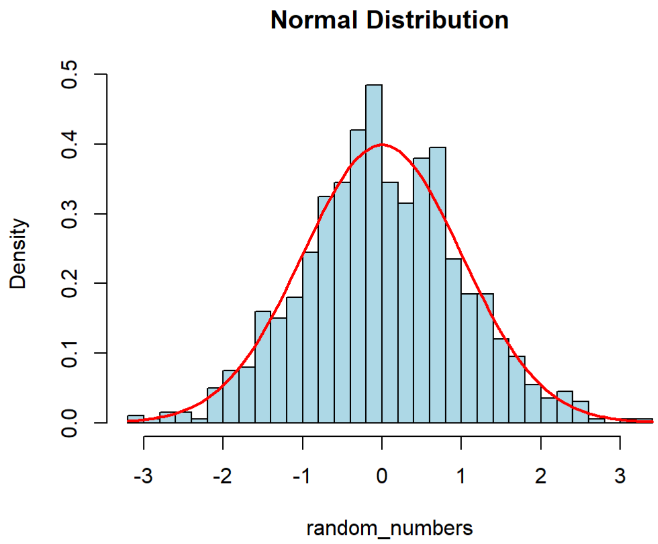

图 3　正态分布

正态分布随机数

```
# 设置参数
N <- 1000
mean <- 0
sd <- 1

# 生成服从正态分布的随机数
random_numbers <- rnorm(N, mean, sd)

# 绘制直方图
hist(random_numbers, breaks = 30, freq = FALSE, col = "lightblue", main = "Normal Distribution")

# 绘制密度函数
curve(dnorm(x, mean, sd), add = TRUE, col = "red", lwd = 2)
```

四、 重要性采样（Importance Sampling）

4.1 理论

重要性采样（Importance Sampling）的基本思想是利用一个已知的易于抽样的分布（提议分布）来近似计算另一个较难抽样的目标分布.

假设需要获得服从概率密度函数为 $p(x)$ 的随机数，但 $p(x)$ 的形式可能很复杂，无法直接抽样. 因此可以选择一个提议分布 $q(x)$，它的形式应该与目标分布相似，但更容易抽样.

对于不容易计算的

$$E(f)=\int_X f(x)p(x)dx$$

可以根据提议分布 Q 进行大量随机采样，样本量为 N，样本点为 $\{x_1,x_2,\cdots,x_N\}$，当 N 足够大时，有

$$E(f)=\int_X f(x)p(x)dx=\int_X f(x)\frac{p(x)}{q(x)}q(x)dx\approx\frac{1}{N}\sum_{i=1}^{N}\frac{p(x_i)}{q(x_i)}f(x_i)$$

其中 $\frac{p(x_i)}{q(x_i)}$ 为 $f(x_i)$ 的重要性，即重要性采样修正因子.

因此，通过重要性采样生成给定概率密度函数的随机数的具体步骤如下：

- Step1：选择提议分布 $q(x)$，常情况下，选择一个与目标分布 $p(x)$ 形状相似的分布是比较合适的.
- Step2：抽样：从提议分布 $q(x)$ 中抽取样本 $\{x_1,x_2,\cdots,x_N\}$.
- Step3：计算重要性：对于每个样本 x_i，计算相应的重要性 $\omega_i=\frac{p(x_i)}{q(x_i)}$.
- Step4：归一化权重：为了使权重和为 1，对所有权重进行归一化处理.
- Step5：生成随机数：从样本集合中按照归一化权重进行随机抽取，得到生成的随机数.

4.2 示例

然而，通过重要性采样生成随机数具有局限性。为了保证估计的准确性，提议分布 $q(x)$ 应该尽可能接近目标分布 $p(x)$，否则估计的误差可能会很大. 因此为了方便起见，本文选择两个正态分布作为示例.

目标分布服从 $N(0,1)$，提议分布服从 $N(0,4)$，最终重要性采样结果如图 4所示.

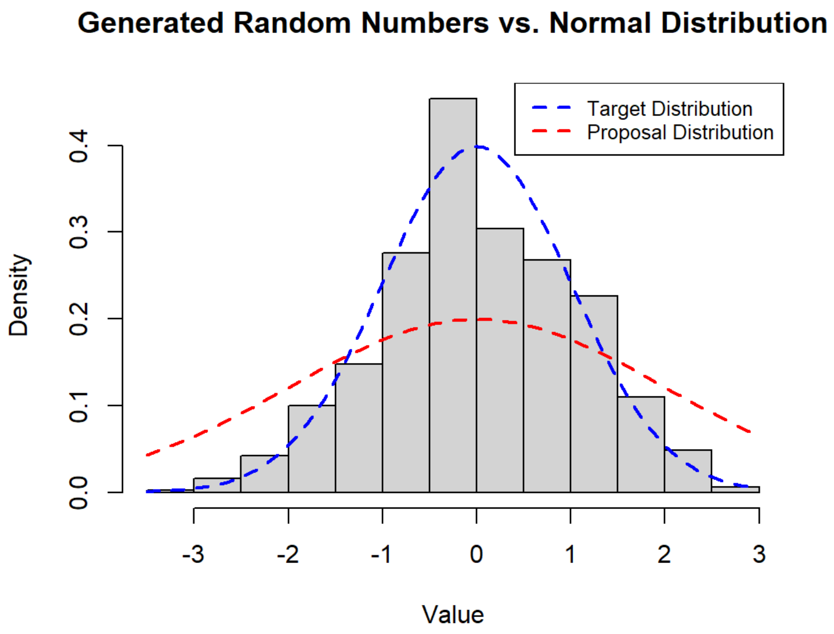

图 4　重要性采样生成随机数示例

```
# 目标分布：正态分布
target_pdf <- function(x) {
  dnorm(x, mean = 0, sd = 1)
}
# 提议分布：另一个正态分布
proposal_pdf <- function(x) {
  dnorm(x, mean = 0, sd = 2)  # 使用均值为0，标准差为2的正态分
   布作为提议分布
}
# 生成样本数量
N <- 1000
# 从提议分布中抽样
samples <- rnorm(N, mean = 0, sd = 2)
# 计算权重
weights <- target_pdf(samples) / proposal_pdf(samples)
# 归一化权重
normalized_weights <- weights / sum(weights)
# 生成随机数
random_indices <- sample.int(N, size = N, replace = TRUE, prob
```

```
    = normalized_weights)
random_samples <- samples[random_indices]
# 绘制生成的随机数的直方图与正态分布的密度曲线进行比较
hist(random_samples, freq = FALSE, main = "Generated Random
   Numbers vs. Normal Distribution",
     xlab = "Value", ylab = "Density")
curve(dnorm(x, mean = 0, sd = 1), add = TRUE, col = "blue",
   lwd = 2, lty = 2)  # 目标分布的密度曲线
curve(dnorm(x, mean = 0, sd = 2), add = TRUE, col = "red", lwd
    = 2, lty = 2)   # 提议分布的密度曲线
legend("topright", legend = c("Target Distribution", "Proposal
    Distribution"),
       col = c("blue", "red"), lty = 2, lwd = 2, cex = 0.8)
```

五、 混合变换法

5.1 理论

对于一个复杂的概率密度函数，根据全概率公式，可以将其转化为若干个已知的简单分布的组合，以简化问题.

对于离散型，可以写成

$$F_X(x) = \sum \theta_i F_{X_i}(x)$$

其中，$\sum \theta_i = 1$.

对于连续性，可以写成

$$F_X(x) = \int F_{X|Y=y} f_Y(y) dy$$

5.2 示例

三角形分布的概率密度函数为

$$f(x) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} if\ a \leq x \leq c \\ \frac{2(b-x)}{(b-a)(b-c)} if\ c \leq x \leq b \\ 0\ otherwise \end{cases}$$

这个分布可以表示为两个独立的均匀分布的加权和，即 $f(x) = \frac{1}{2}U(a, c) + \frac{1}{2}U(c, b)$，其中 $U(a, b)$ 表示取值范围在 $[a, b]$ 上的均匀分布。可以先生成一个均匀分布的随机数 U，然后根据 U 的取值来确定是从区间 $[a, c]$ 还是 $[c, b]$ 中生成随机数。

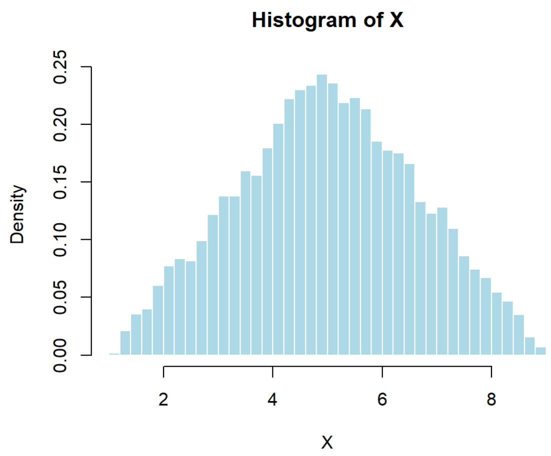

图 5　混合变换法生成随机数示例

```r
# 定义生成三角形分布随机数的函数
triangle_random <- function(a, b, c, size) {
  U <- runif(size)
  X <- ifelse(U < (c - a) / (b - a), a + sqrt(U * (b - a) * (c
    - a)), b - sqrt((1 - U) * (b - a) * (b - c)))
  return(X)
}
# 设置参数
a <- 1
b <- 9
c <- 5
size <- 10000
# 生成三角形分布的随机数
X <- triangle_random(a, b, c, size)
# 绘制直方图
hist(X, breaks = 50, prob = TRUE, col = "lightblue", border =
   "white")
```