Bayesian notes

1 在做什么

在估计参数, $p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)\, d\theta}$, 我们想知道参数 θ 的后验分布是什么, 但是所知道的东西只有 prior $p(\theta)$ 和 likelihood$p(y|\theta)$. 而现实中往往面临的情况是分母这个积分根本没有办法积出来.

2 怎么做

一共有三种方法估计参数:
(1) Analytic: 直接能求出来 $p(\theta|y)$ 的具体表达式
(2) Sampling: 用随机样本逼近后验分布. 例如:Gibbs, Metropolis–Hastings, HMC
(3) Approximation: 用优化或展开近似分布

2.1 Analytic

什么时候能直接写出后验？当先验与似然属于“共轭”组合时，后验与先验同族，只是参数更新了一下。这样无需数值方法，可直接写出后验分布。常见的共轭对有:
(1) 正态均值
prior: $\mu \sim N(m, s^2)$
likelihood: $y_i|\mu \sim N(\mu, \sigma^2)$ (σ^2 已知)
posterior: $\mu|y \sim N(\hat{m}, \hat{s}^2)$
(2) 正态线性回归

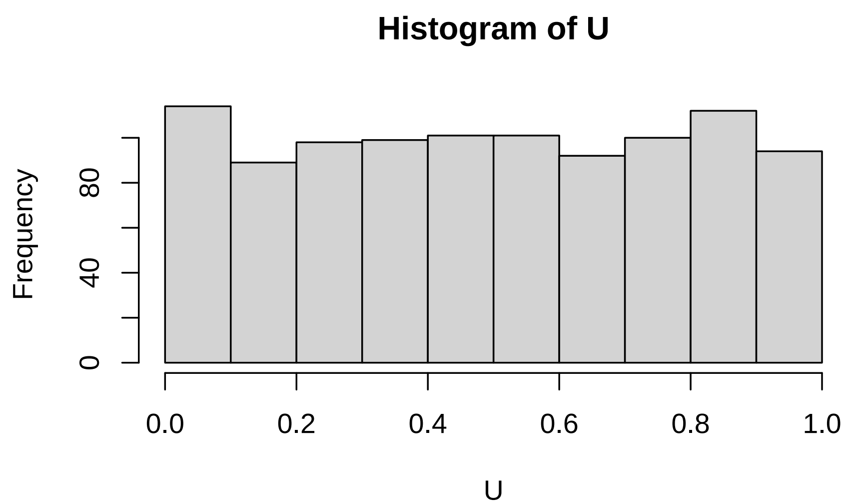
Conjugate priors: $\beta|\sigma^2 \sim N(B, \sigma^2 \sum), \sigma^2 \sim Inv - Gamma(a, b)$
Posterior: $\beta|\sigma^2, y$ 仍为正态,$\sigma^2|y$ 为逆伽马

2.2 Sampling

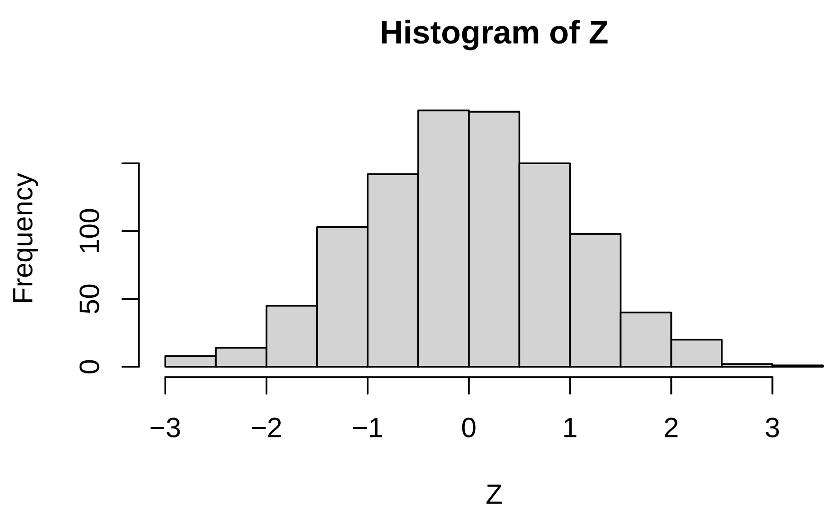
Direct Sampling

首先要回答第一个问题, 什么是采样, 如果我说我从正态分布 $N(0,1)$ 里取 1 个随机数是什么意思. 先说过程, 我先从 $Uniform(0,1)$ 中取出一个数 y (可以闭眼从 1cm 的直尺上随机指出一个数), 接着使用 $\Phi(z) = y$ 计算出 z 的值, 那么这就是我们从 $N(0,1)$ 中取出的一个随机数, 为什么? 接下来给出证明:
(1) 目的是要证明这样计算出的 Z 应该服从正态分布: $p(Z \leq z) = \Phi(z)$
(2) Z 是由 $\Phi(Z) = Y$ 也就是 $\Phi^{-1}(Y)$ 计算出来的, 因此 $p(Z \leq z) = p(\Phi^{-1}(Y) \leq z) = p(Y \leq \Phi(z)) = \Phi(z)$ (中间等号根据分布函数的单调性和反函数的性质)

```
U = runif(1000,0,1) # 生成 1000 个服从 U(0,1) 的数
hist(U)
```




```
Z = qnorm(U) # 做正态分布的逆变换
hist(Z)
```



这个方法叫做直接采样 (Direct Sampling), 但缺点显而易见, 我们必须能知道怎么计算后验分布的反函数, 但很多分布太过于复杂根本没办法求解反函

数, 就需要一些办法来采样.

Rejection Sampling

拒绝采样的原理并不复杂, 这个方法依赖于我们提出一个 proposal density, 要求这个提议分布必须得完全包裹住我们想要采样的目标分布, 也就是条件 $cq(x) \geq p(x)$, 这个常数项 c 只是为了保证 proposal density 一定能包裹住目标分布, 接着想象在 $cq(x)$ 曲线上随机撒点, 我们只留下那些在目标分布 $p(x)$ 曲线下的点, 对于每个从 $q(x)$ 里取样出来的点, 是否接受它们取决于它们自己的权重, 这个权重用 $U \leq \frac{p(X)}{cq(X)}$ 来表示, 为什么? 如果定义事件 A 为接受这些点, 那么 $P(A) = P(U \leq \frac{p(X)}{cq(X)})$, 只有当满足 $U \sim U(0,1)$ 的时候 $P(A) = P(U \leq \frac{p(X)}{cq(X)}) = \frac{p(X)}{cq(X)}$, 意味着此时接受 X 的概率正好等于这个点自己在目标分布 $q(x)$ 曲线下方的概率, 那当把所有这些接受的 X 都收集起来, 就能得到目标分布的分布曲线.
以下是个简单的例子如何来运行拒绝采样. 我们的目标分布是 $f(x) = 2x$, 提议分布是 $U(0,1)$, 这样令 $c = 2$ 正好使得提议分布在目标分布之上, 步骤一共五步:
(1) 选择 c
(2) 从提议分布里采样一个数 x_i
(3) 从 $U(0,1)$ 里采样一个数 u_i
(4) 比较 $u_i \leq \frac{p(x_i)}{cq(x_i)}$ 是否成立, 若成立则留下 x_i
(5) 重复以上步骤多次直到生成完整的样本

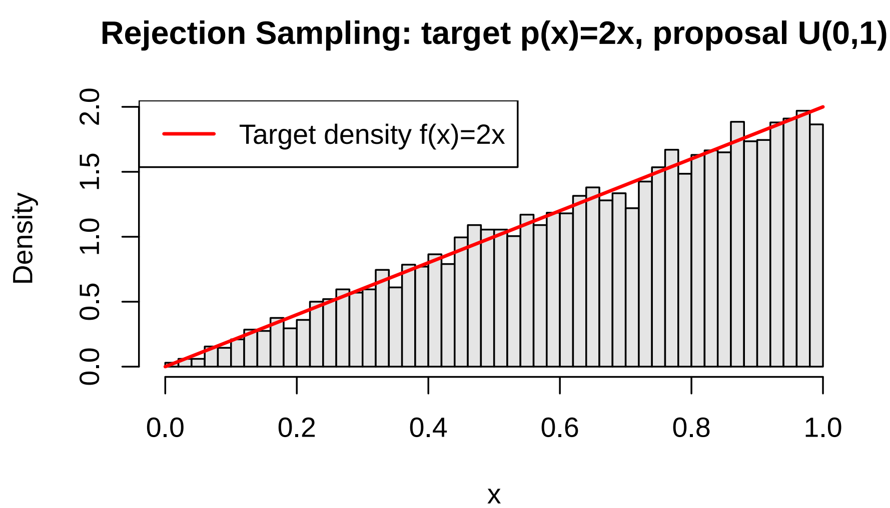
```r
set.seed(123)

n <- 10000                 # 想要的样本数
accepted <- c()            # 存放被接受的样本
cst <- 2                   # 因为 p(x)/q(x)=2x, 最大值是 2, 所以 c = 2 就够

while (length(accepted) < n) {
  x <- runif(1)            # 从 U(0,1) 提样
  u <- runif(1)            # 从 U(0,1) 提样作为接受概率
```

```
  if (u < (2 * x) / cst) {  # 接受条件
    accepted <- c(accepted, x)
  }
}

# 画图对比目标密度
hist(accepted, breaks = 50, freq = FALSE, col = "grey90",
     main = "Rejection Sampling: target p(x)=2x, proposal U(0,1)",
     xlab = "x")
curve(2*x, from = 0, to = 1, add = TRUE, lwd = 2, col = "red")
legend("topleft", legend = c("Target density f(x)=2x"), col = "red", lwd = 2)
```



拒绝性采样的缺陷来自几点：
（1）效率低，一旦 proposal density 和 taget density 差距很大，在判别此处时 $U \leq \frac{p(X)}{cq(X)}$ 就会丢弃非常多的样本
（2）需要知道上界常数 c ，使用该方法必须满足 $cq(x) \geq p(x)$ ，但在分布非常复杂时无法做到这一点

Importance Sampling

重要性采样的原理也并不复杂, 核心思想是：用一个容易采样的分布来近似原本难以采样的目标分布，并通过加权修正来保证估计无偏。
比如说要计算的期望是 $E[f(X)] = \int f(x)p(x)dx$, 很多情况下 $p(x)$ 太过于复杂这个积分没有办法计算, 我们转而采用 Monte Carlo 的办法模拟近似估计:$E[\hat{f(X)}] = \frac{1}{n}\sum f(x_i)$, 但问题在于 x_i 都需要从分布 $p(x)$ 中直接采样得到, 但很难做到这一点, 所以才进行了数学变换 $E_p[f(X)] = \int f(x)p(x)dx = \int \frac{f(x)p(x)}{q(x)}q(x)dx = E_q[\frac{f(X)p(X)}{q(X)}]$, 通过这样的方式, 此时如果我们再用 Monte Carlo 来近似的话, 样本只需要从我们设计的 proposal density$p(x)$ 里采样即可, 此时再用 Monte Carlo 计算估计量表示为:$E[\hat{f(X)}] = \frac{1}{n}\sum \frac{f(x_i)p(x_i)}{q(x_i)}$, 值得注意的是这个估计量虽然是无偏的但是方差却很大, 因此我们实际使用的估计量是:$E[\hat{f(X)}] = \sum f(x_i)\widetilde{w_i}$, 其中 $\widetilde{w_i} = \frac{w_i}{\sum w_j}, w_i = \frac{p(x_i)}{q(x_i)}$, 这个估计量的推导原理也很简单:$E_p[f(X)] = \frac{\int f(x)p(x)dx}{\int p(x)dx} = \frac{\int \frac{f(x)p(x)}{q(x)}q(x)dx}{\int \frac{p(x)}{q(x)}q(x)} = \frac{E_q[\frac{f(X)p(X)}{q(X)}]}{E_q[\frac{p(X)}{q(X)}]}$, 分母上的 $\int p(x)dx = 1$ 因为这是 pdf 的积分, 此外这里也能看得出来一个明显的区别就是如果我们只知道 $p(x) \propto g(x)$ 的话, 那在前一个估计量处进行计算 $E[\hat{f(X)}] = \frac{1}{n}\sum \frac{f(x_i)p(x_i)}{q(x_i)}$ 就是不正确的, 因为使用的 $p(x)$ (其实是 $g(x)$) 是一个未归一化的函数, 所谓的未归一化指的是 $\int p(x)dx = 1$ 但是如果 $p(x) \propto g(x)$ 那么 $1 = \int p(x)dx = \int \theta g(x)dx, \int g(x)dx = \frac{1}{\theta}$

```r
set.seed(1234)

# --- 目标分布：混合高斯 ---
x <- c(rnorm(5000,-1,1), rnorm(5000,6,1))

# --- 提议分布：简单高斯 ---
m <- 100000
g <- rnorm(m, 0, 2)
g.sorted <- sort(g)

# --- 计算权重 w = p/q ---
weights <- (0.7*dnorm(g.sorted, -1, 1) + 0.3*dnorm(g.sorted, 6, 1)) /
            dnorm(g.sorted, 0, 2)
```

```r
# --- 归一化 + 重采样 ---
w_norm <- weights / sum(weights)
resample_idx <- sample(1:length(g.sorted), size=1000, replace=TRUE, prob=w_norm)
x_resampled <- g.sorted[resample_idx]

# --- 可视化 ---
plot(density(x), col="black", lwd=2, main="Importance Sampling + Resampling",
     xlab="x", xlim=c(-5,10),ylim=c(0, 0.6))
lines(density(g.sorted), col="gray", lwd=2, lty=2)
lines(density(x_resampled), col="blue", lwd=2)
legend("topright", legend=c("Target p(x)", "Proposal q(x)", "Resampled Samples"),
       col=c("black", "gray", "blue"), lty=c(1,2,1), lwd=2)
```