Computation

Statistics 220

Spring 2005

Simulation of Random Variables

As we've seen so far, in performing Bayesian analyzes, being able to sample from a wide range of distributions is important. However in most computer languages (e.g. C, Fortran, Pascal, etc) only a very few distributions are available. Often only uniform generators are available. These can be supplemented by numerical analysis libraries for common distributions. This is effectively what happens in most Stat packages (`R`, `SAS`, `Minitab`, etc).

How can we expand past what these libraries allow? How do the routines in these libraries work?

- Grid approach (seen so far)
- Inverse CDF
- Relationships with other distributions
- Acceptance - Rejection Sampling

Grid approach

Assume that the random variable X has density $p(x)$.

- Choose an equally spaced grid of values $x_{1:n} = x_1, \ldots, x_n$
- Evaluate the density for each x_i: $p_i = p(x_i)$ and set $p_0 = 0$
- Normalize the values

$$\tilde{p}_i = \frac{p_i}{\sum_{j=0}^{n} p_j}$$

  giving a valid discrete probability distribution function on $x_{1:n}$

- Calculate the CDF for this discrete distribution

$$\tilde{P}_i = \sum_{j=1}^{i} \tilde{p}_j \approx P[X \leq x_i]$$

Then to generate a draw x from $p(x)$

- Sample $u \sim U(0, 1)$
- Set x to be x_i where $\tilde{P}_{i-1} < u \leq \tilde{P}_i$

Note that this scheme is a mechanism for drawing from any discrete distribution.

In `R` this scheme can be implemented using the `sample()` function.

Advantages:

- Quick and easy
- Will work for any density function
- Easily extended to multivariate densities

Disadvantages:

- Draws are from an approximation to the true distribution
- How many grid points n. If n is too small this will be a poor approximation.
- Most values of the random variable can not be generated by this scheme.

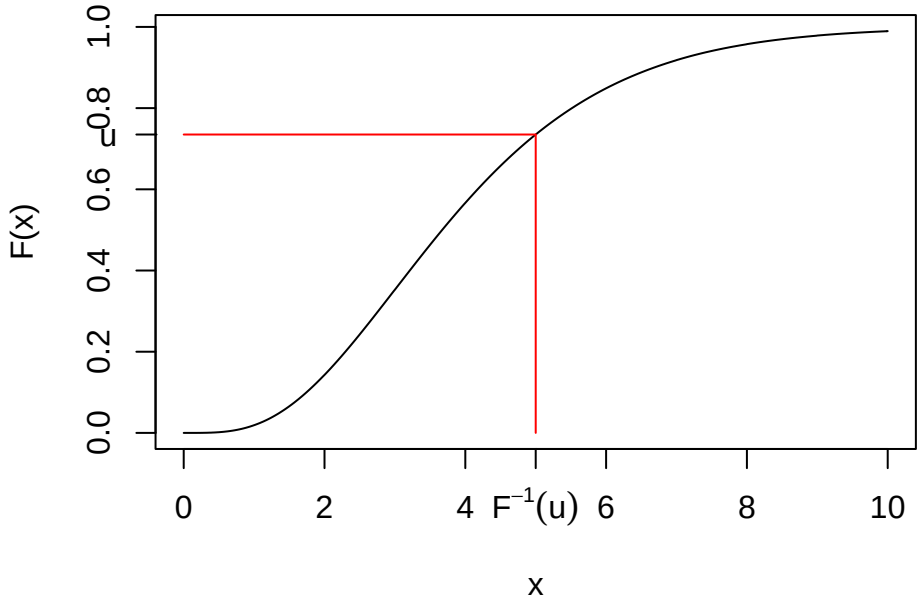
So we need alternate schemes for simulating random variables.

Inverse CDF Method

Let $F(x) = P[X \leq x]$ be the CDF of the random variable X.

Then the inverse CDF (or quantile function) is defined by

$$F^{-1}(u) = \inf\{x : F(x) \leq u\}$$



For continuous RVs

$$P[F(X) \leq u] = P[X \leq F^{-1}(u)] = F(F^{-1}(u)) = u$$

i.e. $F(X) \sim U(0, 1)$

Thus given and iid $U(0,1)$ sample $u_1, \ldots, u_m$, an iid sample $x_1, \ldots, x_m$ from F can be obtained by

$$x_i = F^{-1}(u_i)$$

Examples:

1. $Cauchy(\mu, \sigma)$

$$\begin{aligned} F(x; \mu, \sigma) &= \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{x-\mu}{\sigma}\right) \\ F^{-1}(u; \mu, \sigma) &= \mu + \sigma \tan(\pi(u - 0.5)) \end{aligned}$$

2. $Exp(\mu)$

$$\begin{aligned} F(x; \mu) &= 1 - \exp(-x/\mu) \\ F^{-1}(u; \mu) &= -\mu \log(1-u) \end{aligned}$$

3. Discrete distributions

   The approach discussed earlier for discrete distributions is effectively inverting the CDF.

Note that sometimes its easier to work with the survivor function $S(x) = 1 - F(x)$.

Since U and $1 - U$ both have uniform distributions $S^{-1}(u)$ will also be a draw from F. For example

$$S^{-1}(u; \mu) = -\mu \log u$$

will also give a draw from an exponential distribution with mean μ.

Advantages:

- Will give draws from the correct distribution
- Don't need to worry about things like gridding values.

Disadvantages:

- While the density is not always of a nice form, the CDF and its inverse often aren't (e.g. Normal, Gamma, Beta, etc).
- Though there are often good approximation for the quantile function (e.g. `R` and `Matlab` use a rational function approximation), these are often slow and a poor approximation for simulation purposes (particularly in the tails of the distribution).

  Slightly surprisingly, `R` uses this approach as the default though there are 4 other methods available.

- For a discrete distribution with many classes, they may be alot of comparisons made to determine x_k. For example, `R` doesn't use this approach for Binomial draws if $np > 30$

Relationships with Other Distributions

Examples:

- $X \sim N(\mu, \sigma^2)$ then $Y = e^X \sim LogN(\mu, \sigma^2)$
- $X \sim N(0, 1)$ then $Y = X^2 \sim \chi^2_1$
- $X_\alpha \sim Gamma(1, \alpha), X_\beta \sim Gamma(1, \beta)$ then

$$Y = \frac{X_\alpha}{X_\alpha + X_\beta} \sim Beta(\alpha, \beta)$$

- $X \sim U(0, 1)$ then $Y = -\log X \sim Exp(1)$

The inverse CDF method can be thought of as a special case of this.

Advantages:

- Will give draws from the correct distribution

Disadvantages:

- Many distributions don't have useful relationships
- Can be inefficient as functions like $\log, \sin, \cos$ can be somewhat expensive to calculate

Acceptance-Rejection

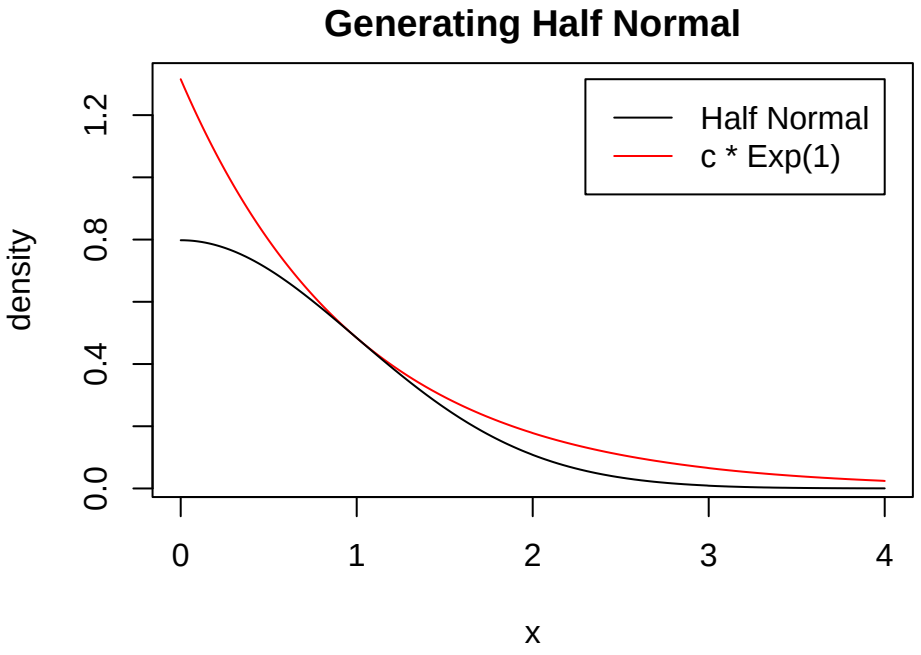
Due to von Neumann (1951)

Want to simulate from a distribution with density $f(x)$.

Need to find a "dominating" or majorizing distribution $g(x)$ where g is easy to sample from and

$$f(x) \leq cg(x) = h(x)$$

for all x and some constant $c > 1$.



Sampling scheme

1. Sample x from $g(x)$ and compute the acceptance ratio

$$r(x) = \frac{f(x)}{cg(x)} = \frac{f(x)}{h(x)} < 1$$

2. Sample $u \sim U(0, 1)$

   If $u \leq r(x)$ accept and return x

   If $u > r(x)$ reject and go back to 1)

   Note that this step is equivalent to flipping a biased coin with success probability $r(x)$

Then the resultant sample is a draw from the density $f(x)$.

Proof. Let I be the indicator of whether a sample x is accepted. Then

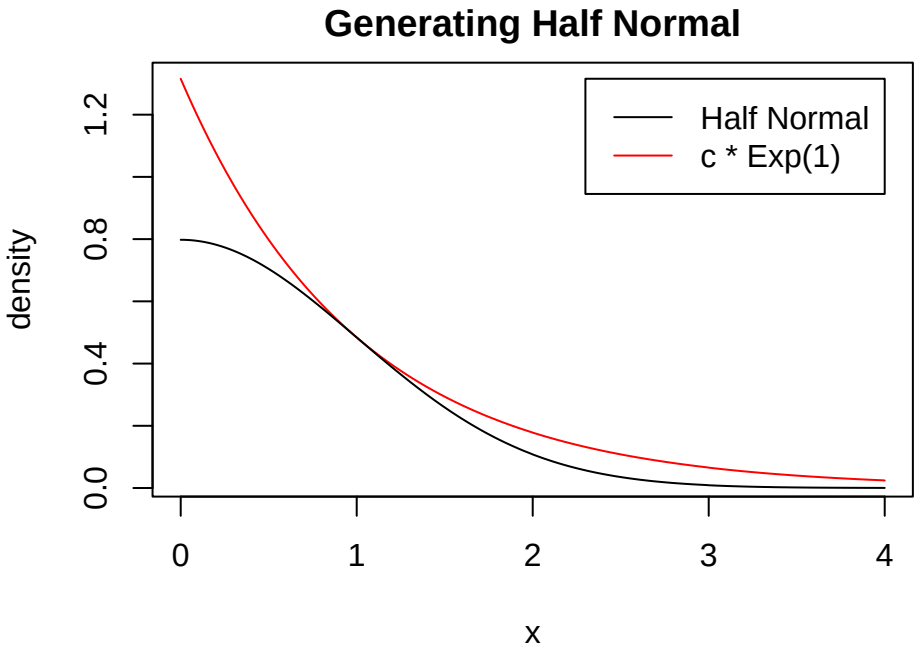
$$\begin{aligned} P[I=1] &= \int P[I=1|X=x]g(x)dx \\ &= \int r(x)g(x)dx \\ &= \int \frac{f(x)}{cg(x)}g(x)dx = \frac{1}{c} \end{aligned}$$

Next

$$\begin{aligned} p(x|I=1) &= \frac{f(x)}{cg(x)}g(x) \Big/ P[I=1] \\ &= \frac{f(x)}{c}c = f(x) \end{aligned}$$

□

For a more geometrical proof see Flury (1990) (on the web site). Its based on the idea of drawing uniform points (x, y) under the curve $h(x)$ and only accepting the points that also lie under the curve $f(x)$.



The number of draws needed until an acceptance occurs is $Geometric(\frac{1}{c})$ and thus the expected number of draws until a sample is accepted is c.

The acceptance probability satisfies

$$\frac{1}{c} = \frac{\int f(x)dx}{\int cg(x)dx} = \frac{\text{Area under } f(x)}{\text{Area under } h(x)}$$

One consequence of this is that c should be made as small as possible to minimize the number of rejections.

The optimal c is given by

$$c = \sup \frac{f(x)}{g(x)}$$

Note that the best c need not be determined, just one that satisfies

$$f(x) \leq cg(x) = h(x)$$

for all x.

Example: Generating from the half normal distribution

$$\begin{aligned} f(x) &= 2\phi(x)I(x \geq 0) \\ &= \sqrt{\frac{2}{\pi}}\exp(-0.5x^2)I(x \geq 0) \end{aligned}$$

Lets use an $Exp(1)$ as the dominating density

$$g(x) = e^{-x}I(x \geq 0)$$

The optimal c for this example is

$$c = \sqrt{\frac{2}{\pi}}\exp(0.5) \approx 1.315$$

so the acceptance rate is approximately 76%

This the acceptance-rejection scheme is

1. Draw $x \sim Exp(1)$

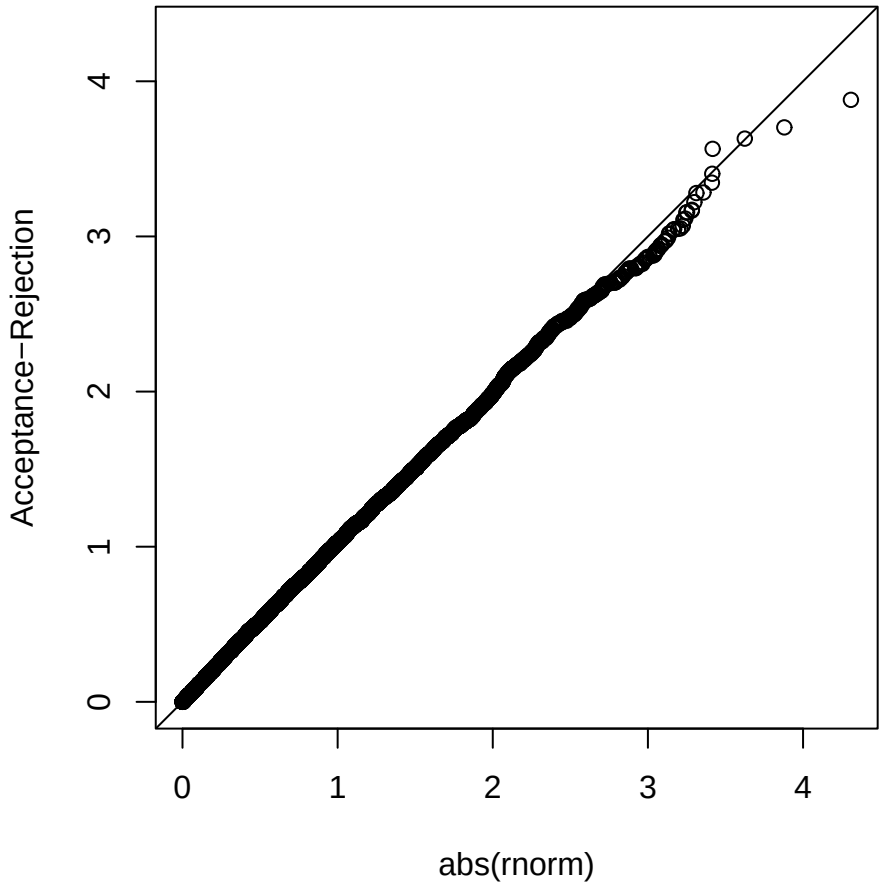
$$r(x) = \exp(-0.5(x-1)^2)$$

2. Draw $u \sim U(0,1)$

   If $u \leq r(x)$ accept and return x

   If $u > r(x)$ reject and go back to 1)

Note that this scheme isn't needed as the half normal distribution is the distribution of the absolute value of a $N(0,1)$



In the above, it was assumed that $f(x)$ was a density function. In fact $f(x)$ only needs to be known up to a multiplicative constant

$$l(x) = bf(x)$$

where b may be unknown.

This is common in our situation as the posterior density is usually only known up to

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

and the normalizing constant is difficult to calculate exactly.

However the acceptance-reject approach does not require knowing this constant. The procedure can be modified to

Find a c such that

$$l(x) \leq cg(x) = h(x)$$

for all x and $c > 1$.

1. Sample x from $g(x)$ and compute the ratio

$$r(x) = \frac{l(x)}{cg(x)} = \frac{l(x)}{h(x)} \leq 1$$

2. Sample $u \sim U(0, 1)$

   If $u \leq r(x)$ accept and return x

   If $u > r(x)$ reject and go back to 1)

Everything is the same except the unnormalized density $l(x)$ is used instead of the normalized density $f(x)$.

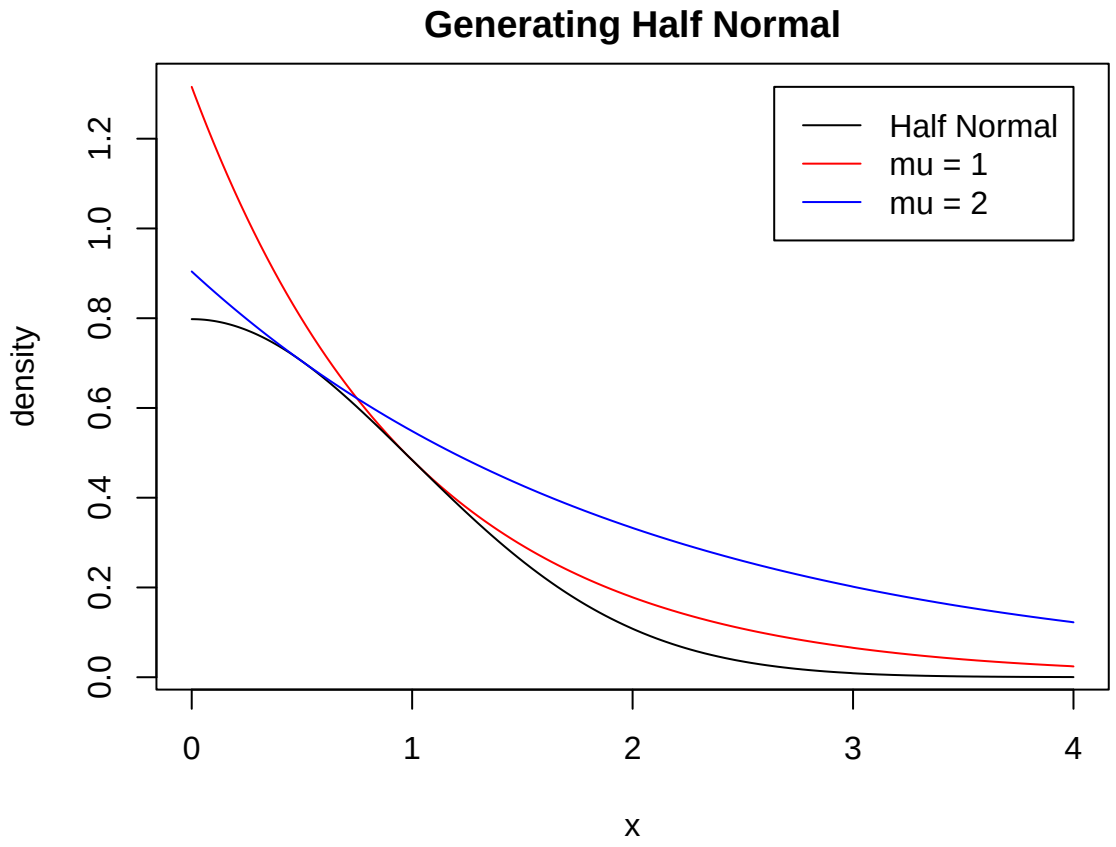
The acceptance probability for this scheme is $\frac{b}{c}$.

In addition to the constant c chosen, the distribution $g(x)$ will also affect the acceptance rate. (c is chosen conditional on $g(x)$).

A good choice $g(x)$ will normally be "close to" $f(x)$. You want to minimize the separation between the two densities.

Often a parametric family will be chosen and the member of the parametric family with the smallest c will then be used.

For example, for the half normal distribution, which $Exp(\mu)$ will minimize $c(\mu)$



In fact $\mu = 1$ will minimize $c(\mu)$ for this problem.

Note that so far its appeared that this has focused on continuous random variables.

In fact acceptance-rejection works fine with discrete random variables and with variables with more than one dimension.

The proof presented earlier goes through in this more general setting by replacing integration over a density to integration over a more general measure.

For discrete problems, you get a sum over the probability mass function.

With higher dimensional problems, the majorization constants (i.e. the cs) tend to be higher, implying the procedure is less efficient.

Advantages:

- Will give draws from the correct distribution.
- Extremely flexible.
- Approach will work for a wide range of problems.
- For many problems there are good choices for the majorizing distribution (i.e. log concave densities).

Disadvantages:

- Maybe inefficient (large c).
- How to pick majorizing distribution not always clear.