Chapter 4

Generating non-uniform random variables

4.1 Inversion

We saw in the last chapter that if the CDF is strictly increasing, then $F(X)$ has a uniform distribution. Conversely, it is easy to show in this case that if U is uniformly distributed on $[0,1]$ then $F^{-1}(U)$ has the distribution $F(x)$. For this we do not need that the CDF is strictly increasing. In this case the usual inverse function need not be defined. We define

$$F^{-1}(u) = \inf\{x : F(x) \geq u\} \tag{4.1}$$

for $0 < u < 1$. (really should not denote this by F^{-1} since it is not the inverse of F.)

Theorem 1 *Let $F(x)$ be a CDF. Define F^{-1} as above. Let U be uniformly distributed on $[0,1]$. Then the CDF of $F^{-1}(U)$ is $F(x)$.*

Proof: Consider the set $\{x : F(x) \geq U\}$. Since F is non-decreasing it must be of the form (c, ∞) or $[c, \infty)$. The right continuity of F implies it must be $[c, \infty)$. We need to compute $P(F^{-1}(U) \leq t)$. We claim that $F^{-1}(U) \leq t$ if and only if $U \leq F(t)$. (This is not quite as trivial as the notation makes it look since F^{-1} is not really the inverse function for F.) The claim will complete the proof since the probability that $U \leq F(t)$ is just $F(t)$.

First suppose $F^{-1}(U) \leq t$. Then t must belong to the set $\{x : F(x) \geq U\}$. So $F(t) \geq U$. Now suppose that $U \leq F(t)$. Then t belongs to $\{x : F(x) \geq U\}$. So t is greater than or equal to the inf of this set, i.e., $t \geq F^{-1}(U)$.

QED

In principle this allows us to generate samples of any distribution. The catch is that it may be difficult to compute F^{-1}. Note that this works for any CDF, continuous or discrete.

Example (exponential) The CDF of the exponential is

$$F(x) = 1 - \exp(-\lambda x) \tag{4.2}$$

and so

$$F^{-1}(u) = -\frac{\ln(1-u)}{\lambda} \tag{4.3}$$

So if U is uniform on $[0,1]$, then $-\ln(1-U)/\lambda$ has the exponential distribution with parameter λ. Note that if U is uniform on $[0,1]$ then so is $1-U$. So we can also use $-\ln(U)/\lambda$ to generate the exponential distribution.

Example - discrete distributions Let $x_1 < x_2 < \cdots < x_n$ be the values and p_i their probabilities. The CDF is piecewise constant and F^{-1} only takes on the values $x_1, x_2, \cdots, x_n$. The value of $F^{-1}(U)$ is x_k where k is the integer such that

$$\sum_{i=1}^{k-1} p_i < U \leq \sum_{i=1}^{k} p_i \tag{4.4}$$

This corresponds to the obvious way to generate a discrete distribution.

To implement this we need to find the k that satisfies the above. We can do this by searching starting at $k = 1$ and computing the partial sums as we go. This takes a time $O(n)$.

If we are going to generate a random variable with this same set of p_i many times we can do better. To implement this we first set up a table with the partial sums above. This takes a time $O(n)$, but we do this only once. Then when we want to generate a sample from the distribution we have to find the k that satisfies the above. If we use a bisection search we can reduce the time to $O(\ln(n))$. **Explain bisection** So there is a one time cost that is $O(n)$ and then the cost per sample is $O(\ln(n))$.

If we have a discrete distribution with an infinite number of values we can still do inversion. Doing this by starting at $k = 1$ and searching for the k that satisfies the above is straightforward. How long does this take on average? Is there a version of bisection that does better?

Example - Normal distribution At first glance it looks like we cannot use the inversion method for the normal distribution since we cannot explicitly compute the CDF and so cannot compute its inverse. Nonetheless, there are very fast methods for computing F^{-1} to high accuracy. See Owen for more details.

Scaling and shifting For many families of random variables there is a parameter that just corresponds to scaling the random variable, i.e., multiplying by a constant, or a parameter that corresponds to shifting the random variable, i.e., adding a constant.

For example consider the exponential random variable which has density

$$f(x) = \lambda \exp(-\lambda x), x \geq 0 \tag{4.5}$$

If Y is exponential with parameter 1 and we let $X = \lambda Y$ then X is exponential with the above density.

The normal density with mean μ and variance σ^2 is

$$f(x) = c \exp(-\frac{1}{2}(x-\mu)^2/\sigma^2) \tag{4.6}$$

If Z is a standard normal then $X = \sigma Z + \mu$ will have the above density.

The above can be thought of as applying an affine transformation to our random variable. If $g(x)$ is an increasing function and we let $Y = g(X)$ then the density of Y is related to that of X by

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{g'(g^{-1}(y))} \tag{4.7}$$

4.2 Acceptance-rejection

The method of acceptance-rejection sampling is also called rejection sampling or accept-reject. We discuss the case of continuous random variables. There is a similar method for discrete random varibles. Suppose we want to simulate a random variable with density $f(x)$. We assume there is another density $g(x)$ and a constant c such that

$$f(x) \leq cg(x) \tag{4.8}$$

Note that by integrating this equation we see that c must be greater than 1. (It can equal 1 only if $f(x) = g(x)$.) We also assume it is relatively easy to generate samples with density g. The algorithm is as follows.

repeat

$Y \sim g$

$U \sim U(0,1)$

until $U \le f(Y)/(cg(Y))$

$X = Y$

output X

The notation $Y \sim g$ means we generate a random variable with the density g. $U(0,1)$ denotes the uniform distribution on $[0,1]$, and $U \sim U(0,1)$ means we generate a random varible U with this distribution. Note that the test $U \le f(Y)/(cg(Y))$ means that we accept Y with probability $f(Y)/(cg(Y))$. Otherwise we reject Y and try again.

Theorem 2 *With the reject-accept algorithm above, the probability density of X is given by $f(x)$.*

Review the partition theorem for continuous RV's

The acceptance-rejection algorithm can take multiple attempts to get a value of X that we accept. So we should worry about how long this takes. The number of tries is random; in fact the number of tries has a geometric distribution. Let p be the parameter for this geometric distribution, i.e., the probability we accept on a particular attempt. Given $Y = y$, the probability we accept is $f(y)/(cg(y))$. So

$$p = \int P(accept|Y = y)g(y)\,dy = \int \frac{f(y)}{cg(y)}g(y)\,dy = \frac{1}{c} \tag{4.9}$$

where "accept" is the event that we accept the first try. Note that the closer c is to 1, the closer the acceptance probability is to 100%. The mean of a geometric is $1/p$, so the mean number of tries is c.

Proof: We will compute the CDF of X and show that we get $P(X \le t) = \int_{-\infty}^{t} f(y)dy$.

$$P(X \le t) = P(\{X \le t\} \cap A_1) + P(\{X \le t\} \cap A_1^c) \tag{4.10}$$

where A_1 is the event that we accept the first try. For the second term we use

$$P(\{X \le t\} \cap A_1^c) = P(X \le t|A_1^c)P(A_1^c) \tag{4.11}$$

If we reject on the first try, then we just repeat the process. So $P(X \le t|A_1^c) = P(X \le t)$.

For the first term we use the continuous form of the partition theorem. Let Y_1 be the Y random variable on the first attempt. We condition on the value of Y_1:

$$P(\{X \le t\} \cap A_1) = \int P(\{X \le t\} \cap A_1|Y_1 = y)g(y)dy \tag{4.12}$$

Note that $P(\{X \leq t\} \cap A_1 | Y_1 = y) = P(\{Y \leq t\} \cap A_1 | Y_1 = y)$. This is zero if $y > t$ and $P(A_1|Y_1 = y)$ if $y \leq t$. Since $P(A_1|Y_1 = y) = f(y)/(cg(y))$, we get

$$P(\{X \leq t\} \cap A_1) = \int_{-\infty}^{t} \frac{f(y)}{cg(y)} g(y)dy = \frac{1}{c}\int_{-\infty}^{t} f(y)dy \tag{4.13}$$

So we have shown

$$P(X \leq t) = \frac{1}{c}\int_{-\infty}^{t} f(y)dy + (1 - \frac{1}{c})P(X \leq t) \tag{4.14}$$

Solving for $P(X \leq t)$ we find it equals $\int_{-\infty}^{t} f(y)dy$. **QED**

Example: Suppose we want to generate samples from the standard normal distribution. So

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \tag{4.15}$$

We use

$$g(x) = \frac{1}{\pi}\frac{1}{1+x^2} \tag{4.16}$$

the Cauchy distribution. Let

$$c = \sup_x \frac{f(x)}{g(x)} = \sup_x \sqrt{\frac{\pi}{2}} \exp(-x^2/2)(1+x^2) \tag{4.17}$$

It is obvious that c is finite. A little calculus shows the sup occurs at $x = \pm 1$. So we get $c = \sqrt{\frac{\pi}{2}} e^{-1/2} 2 \approx 1.52$. Note that it is easy to sample from the Cauchy distribution. This method will accept about 66 % of the time.

Exercise Consider the normal distribution again, but now we let $g(x) = \frac{1}{2}\exp(-|x|/2)$. Show that this works for the acceptance-rejection method and compute the acceptance probability.

There is a nice geometric interpretation of this method. First a fact from probability.

Theorem 3 *Let $f(x)$ be a density function for a continuous random variable. Let A be the region in the plane given by*

$$A = \{(x, y) : 0 \leq y \leq f(x)\} \tag{4.18}$$

So A is the area under the graph of $f(x)$. Let (X, Y) have the uniform distribution on A. Then the marginal distribution of X is $f(x)$.

Proof: Note that A has area 1. So the joint density function is just the indicator function of A. To get the marginal density of X at x, call it $f_X(x)$, we integrate out y.

$$f_X(x) = \int 1_A(x,y)dy = f(x) \tag{4.19}$$

QED

The theorem goes the other way: if we generate X using the density $f(x)$ and then pick Y uniformly from $[0, f(X)]$, then (X, Y) will be uniformly distributed on A. To see this note that we are making

$$f_{Y|X}(y|x) = \frac{1}{f(x)} 1_{0 \le y \le f(x)} \tag{4.20}$$

So the joint density will be

$$f_{X,Y}(x,y) = f_{Y|X}(y|x)f(x) = 1_{0 \le y \le f(x)} = 1_A \tag{4.21}$$

Now consider the acceptance-rejection method. Let A be area under the graph of $f(x)$ and let B be the area under the graph of $cg(x)$. By our assumption, A is contained in B. Note that B is the area under the graph of $g(x)$ stretched in the vertical direction by a factor of c. So if we sample Y and sample U from the uniform distribution on $[0, 1]$, then $(Y, Ucg(x))$ will be uniformly distributed on B. (Notation is confusing.) The process of accepting it only if $U \le f(Y)/(cg(Y))$ means that we accept the point in B only if it is in A

We have considered continuous random variables but the acceptance-rejection works for discrete random variables as well. If we want to simulate a discrete RV X with pdf $f_X(x)$, i.e., $f_X(x) = P(X = x)$ then we look for another discrete RV Y with pdf $f_Y(x)$ that we know how to simulate. If there is a constant c such that $f(x) \le cg(x)$, then we are in business. The algorithm is essentially identical to the continuous case, and the proof that it works is the same as the proof in the continuous case with integrals replaced by sums.

4.3 Alias method for discrete distribution

This is a method for generating a sample from a discrete distribution with a finite number of values. Let n be the number of values. The method takes a time $O(n)$ to set up, but once it is set up the time to generate a sample is $O(1)$. It is no loss of generality to take the values to be $1, 2, \cdots, n$. Let p_i be the probability of i.

Proposition 1 *Given $p_i > 0$ with $\sum_{i=1}^n p_i = 1$, we can find $q_i \in [0,1]$ for $i = 1, \cdots, n$ and functions $u(i), l(i) : \{1, 2, \cdots, n\} \to \{1, 2, \cdots, n\}$ such that for each value i, we have*

$$\frac{1}{n} \sum_{k:l(k)=i} q_k + \frac{1}{n} \sum_{k:u(k)=i} (1 - q_k) = p_i \tag{4.22}$$

DRAW A PICTURE.

The algorithm is then as follows. Pick k uniformly from $\{1, 2, \cdots\}$ and pick U uniformly from $[0, 1]$. If $U \leq q_k$ we return $l(k)$. If $U > q_k$ we return $u(k)$. The equation in the proposition insures that the probability we will return i is p_i.

Proof: It is convenient to let $P_i = np_i$. So the sum of the P_i is n and we want

$$\sum_{k:l(k)=i} q_k + \sum_{k:u(k)=i} (1 - q_k) = P_i \tag{4.23}$$

If the P_i all equal 1 the solution is trivial. Otherwise we can find k, m such that $P_k < 1$ and $P_m > 1$. Define $q_1 = P_k$, $l(1) = k$, $u(1) = m$. This takes care of all of the "probability" for value k. For value m it takes care of only some of the "probability", $1 - q_1$ to be precise. Now let $P'_1, \cdots, P'_{n-1}$ be $P_1, \cdots, P_n$ with P_k deleted and P_m replaced by
$P_m - (1 - q_1) = P_m + P_k - 1$. Note that the sum of the P'_i for $i = 1$ to $n - 1$ is $n - 1$. So we can apply induction to define $q_2, \cdots, q_n$ and define $l()$ and $u()$ on $2, 3, \cdots, n$. Note that $l(i)$ and $u(i)$ will not take on the value k for $i \geq 2$.

QED

Another method for fast simulation of a discrete RV is the method of guide tables. They are discussed in Owen's book.

Stop - Mon, 2/1

4.4 Tricks for specific distributions

For specific distributions there are sometimes clever tricks for simulating them. We give a few of them for distributions that are frequently used.

4.4.1 Box-Muller

Suppose we want to generate samples from the standard normal distribution. We can use inversion, but then we have the nontrivial problem of inverting the CDF. We can also use acceptance-rejection as we saw. Another method is called the Box-Muller method. It it based on the following observation. Let X, Y be independent, both with the standard normal distribution. So the joint density is

$$\frac{1}{2\pi}\exp(-(x^2+y^2)/2) \tag{4.24}$$

We change to polar coordinates, i.e., we define two new random variables Θ and R by

$$X = R\cos(\Theta) \tag{4.25}$$
$$Y = R\sin(\Theta) \tag{4.26}$$

Then the joint density of R, Θ is $\exp(-r^2/2)r/(2\pi)$. This shows that R and Θ are independent. Θ is uniform on $[0, 2\pi]$ and R has density $r\exp(-r^2/2)$.

We can go the other way; this is the Box Mueller algorithm. Generate Θ and R with these densities. It is trivial to generate Θ. For R we can use inversion since the density function is explicitly integrable. Define X and Y as above and they will be independent, each with a standard normal distribution.

4.4.2 Geometric

We can use inversion to simulate the geometric distribution. If the mean is large this can be slow. Here is another method. Recall that N has the geometric distribution with parameter $p \in (0,1)$ if $P(N=k) = (1-p)^{k-1}p$. Now let X have an exponential distribution with parameter λ. So the density is

$$f(x) = \exp(-\lambda x)\lambda \tag{4.27}$$

So

$$P(k-1 \le X < k) = \int_{k-1}^{k} \exp(-\lambda x)\lambda\, dx = \exp(-(k-1)\lambda) - \exp(-k\lambda) \tag{4.28}$$
$$= \exp(-(k-1)\lambda)[1-\exp(-\lambda)] \tag{4.29}$$

So if we take $p = 1-\exp(-\lambda)$, then X round down to the nearest integer has the geometric distribution. In other words $N = \lfloor X \rfloor$ has a geometric distribution.

4.4.3 Random permutations

We are interested in the uniform distribution on the set of permutations on $\{1, 2, \cdots, n\}$. If n is fairly small we treat this as a discrete random variable with a finite set of values and use techniques we have discussed. But the number of factorials is $n!$ for this will be impractical for even modest values of n. Here is an algorithm that generates a sample in time $O(n)$. The idea is simple. First randomly pick $\pi(i)$. It is just uniform on $\{1, 2, \cdots, n\}$. Now pick $\pi(2)$. It will be uniform on $\{1, 2, \cdots, n\}$ with $\pi(1)$ removed. Then $\pi(3)$ will be uniform on $\{1, 2, \cdots, n\}$ with $\pi(1)$ and $\pi(2)$ removed. The slightly tricky part is keeping track of the integers which are not yet in the range of π. In the following $(k_1, \cdots, k_n)$ will be the integers which are not yet in the range.

Initialize $(k_1, \cdots, k_n) = (1, 2, \cdots n)$
For $i = 1$ to n
    Generate I uniformly from $\{1, 2, \cdots, n - i + 1\}$
    Set $X_i = k_I$
    Set $k_I = k_{n-i+1}$
Return $(X_1, \cdots, X_n)$

4.4.4 RV's that are sums of simpler RV's

Suppose we want to simulate a binomial RV X with n trials. If X_i takes on the values $0, 1$ with probabilities $1 - p, p$ and $X_1, \cdots, X_n$ are independent, then $\sum_{i=1}^{n} X_i$ has a binomial distribution. The X_i are trivial to simulate. Of course if n is large this will not be a fast method.

Similarly the χ^2 distribution with k degrees of freedom is given by

$$\sum_{i=1}^{k} Z_i^2 \tag{4.30}$$

where the Z_i are independent standard normals.

4.4.5 Mixtures

A mixture means that we have random variables $X_1, X_2, \cdots, X_n$ and probabilities $p_1, p_2, \cdots, p_n$. Let I be a random variable with $P(I = i) = p_i$, independent of the X_i. Then we let $X = X_I$. The CDF of X is

$$F_X(x) = \sum_{i=1}^{n} p_i F_{X_i}(x) \tag{4.31}$$

If the X_i are all absolutely continuous with densities $f_{X_i}(x)$ then X is absolutely continuous with densities

$$f_X(x) = \sum_{i=1}^{n} p_i f_{X_i}(x) \tag{4.32}$$

If we can sample the X_i then it should be obvious how we sample X.

4.5 Generating multidimensional random variables

We now consider the case of generating samples of a random vector $(X_1, \cdots, X_d)$. If the random vector $(X_1, \cdots, X_d)$ is discrete with a finite set of values, then sampling from it is no different than sampling from a random variable X that only takes on a finite set of values. So we can use the techniques of the previous sections. Hence we will focus on continuous RV's in this section.

If the components X_i are independent, then this reduces the problem we have considered in the previous sections. This section considers what to do when the components are dependent.

4.5.1 Composition method or conditioning

For simplicity of notation we consider the case of $d = 2$. We want to generate jointly continuous random variables with density $f(x, y)$. Compute $f_X(x)$ by integrating out y. Define $f_{Y|X}(y|x)$ in the usual way. Then we generate X according to $f_X(x)$ and then generate Y according to $f_{Y|X}(y|x)$. Then (X, Y) will have the desired joint distribution. Several things can do wrong. Computing $f_X(x)$ requires doing an integral which may not be explicitly computable. And we need to be able to generate samples from $f_X(x)$ and $f_{Y|X}(y|x)$.

4.5.2 Acceptance-rejection

The acceptance-rejection method generalizes immediately to random vectors. We consider the case of a jointly continuous random vector. Suppose we want to sample from the density $f(x_1, \cdots, x_d)$. If $g(x_1, \cdots, x_d)$ is a density that we can sample from and there is a constant c such that

$$f(x_1, \cdots, x_d) \leq c g(x_1, \cdots, x_d) \tag{4.33}$$

then we generate a random vector $\vec{Y} = (Y_1, \cdots, Y_n)$ according to g and we generate a uniform RV U on $[0, 1]$. We accept $\vec{Y}$ if $U \leq f(Y_1, \cdots, Y_d)/(cg(Y_1, \cdots, Y_d))$. If we reject we repeat.

This is useful only if we can sample from g. If g is the density of independent random variables this may be doable. Note that g corresponds to independent components if (and only if) $g(x_1, \cdots, x_d)$ factors into a product of functions $g_i(x_i)$. So it is natural to seek an upper bound on $f(x_1, \cdots, x_d)$ of this form.

If we want to generate the uniform distribution on some subset S of R^d and we can enclose S in a hypercube, then we can do this with acceptance rejection. The acceptance ratio will be the volume of S divided by the volume of the hypercube. If this is unacceptably small we can try to do better by enclosing S in geometry that "fits" better but for which we can still uniformly sample from the geometry.

4.5.3 The multivariate normal

The random vector $(X_1, \cdots, X_d)$ has a multivariate normal distribution with means $(\mu_1, \cdots, \mu_d)$ and covariance matrix Σ if its density is

$$f(x_1, \cdots, x_d) = c \exp(-\frac{1}{2}((\vec{x} - \vec{\mu}), \Sigma^{-1}(\vec{x} - \vec{\mu})) \tag{4.34}$$

where c is a normalizing constant. This distribution is easily simulated using the following.

Proposition 2 *Let $Z_1, \cdots, Z_d$ be independent standard normal RV's. Let*

$$\vec{X} = \Sigma^{1/2}\vec{Z} + \vec{\mu} \tag{4.35}$$

Then $\vec{X}$ has the multivariate normal distribution above and $\Sigma^{1/2}$ is the matrix square root of Σ.

Idea of proof: Diagonalize Σ and do a change of variables.

4.5.4 Affine transformations

Let $\vec{X} = (X_1, \cdots, X_d)$ be an absolutely continuous random vector with density $f_{\vec{X}}(x_1, \cdots, x_d)$. Let $\vec{Z} = A\vec{X} + \vec{b}$ where A is a d by d matrix. Then $\vec{Z} = (Z_1, \cdots, Z_d)$ is an absolutely continuous random vector with density

$$f_{\vec{Z}}(z_1, \cdots, z_d) = \frac{1}{|det(A)|} f_{\vec{X}}(A^{-1}(\vec{z} - \vec{b})) \tag{4.36}$$

4.5.5 Uniform point in a sphere, on a sphere

Suppose we want to generate a point in R^d that is uniformly distributed in the ball of radius 1. We can do this by acceptance rejection. We generate a point uniformly on the hypercube $[-1,1]^d$ and accept it if it is inside the ball. The acceptance ratio depends on the dimension d and goes to zero as d goes to infinity. So for large d we would like a better method. We would also like to be able to generate points uniformly on the sphere, i.e., the boundary of the ball. (By ball I mean $\{(x_1, \cdots, x_d) : \sum_i x_i^2 \leq 1\}$ and by sphere I mean $\{(x_1, \cdots, x_d) : \sum_i x_i^2 = 1\}$.

Note that generating a point uniformly inside the ball and generating a point uniformly on the sphere are closely related. We can see the relation by thinking of hyper-spherical coordinates. If $\vec{X}$ is uniformly distributed inside the ball, then $\vec{X}/||\vec{X}||$ will be uniformly distributed on the sphere. Conversely, if $\vec{X}$ is uniformly distributed on the sphere and R is an independent, continuous RV with density r^{n-1}/n on $[0,1]$, then $R\vec{X}$ will be uniformly distributed inside the ball.

So we only need a method to generate the uniform distribution on the sphere. Let $Z_1, Z_2, \cdots, Z_d$ be independent, standard normal random variables. Then their joint density is $c \exp(-\frac{1}{2}\sum_{i=1}^d z_i^2)$. Note that it is rotationally invariant. So $\vec{Z}/||\vec{Z}||$ will be uniformly distributed on the sphere.