Lecture 32.

Simulation of Random Variables by inverse CDF.

- Problem: given a uniform r.v. \( U \sim \text{Uniform}(0,1) \) r.v., i.e. with density

\[
 f_U(u) = \begin{cases} 
 1 & 0 < u < 1 \\
 0 & \text{otherwise}
\end{cases}
\]

and an arbitrary CDF \( F(x) \) (either continuous or piecewise constant), can we find a function

\[ x = g(U) \]

such that the new random variable \( X = g(U) \) has exactly the CDF \( F(x) \), i.e. such that

\[
 P(g(U) \leq x) = F(x) \]

The answer is \( \text{YES} \).

This is pretty amazing, because it means that if I can simulate, on a computer, a uniform r.v. \( U \), then I can simulate a random variable \( X \) with arbitrary distribution \( F(x) \):

\[ U \sim \text{Uniform}(0,1) \Rightarrow g(U) \text{ is a random variable with CDF } F(x) \]

the function \( g \) is to be found! It will depend on the chosen \( F(x) \).
Example (from last week): \( P(X = k) = \begin{cases} 
\frac{1}{4} & k = 1 \\
\frac{1}{2} & k = 2 \\
\frac{1}{4} & k = 3 
\end{cases} \) (\( * \))

These are exactly the probabilities (\( * \)) that we want to generate from a \( U \sim \text{Uniform}(0,1) \) continuous distribution. Now, \( U \in (0,1) \) with prob 1.

\[
U \rightarrow \begin{cases} 
1 & \text{if } U \rightarrow \left( \frac{1}{4}, \frac{3}{4} \right) \\
\frac{1}{2} & \text{if } U \rightarrow \left( \frac{3}{4}, 1 \right) \\
0 & \text{if } U \rightarrow \left( 0, \frac{3}{4} \right) 
\end{cases}
\]
then we define \( g(U) = 3 \). (This happens with probability \( \frac{1}{4} \)).

More precisely, define:

\[
g(u) = \begin{cases} 
3 & \text{if } \frac{3}{4} < u \leq 1 \\
2 & \text{if } \frac{1}{4} < u \leq \frac{3}{4} \\
1 & \text{if } 0 \leq u \leq \frac{1}{4} 
\end{cases}
\]

and the new random variable \( g(U) \).
We have:

\[ P(g(U) = 1) = P(0 \leq U \leq \frac{1}{4}) = \frac{1}{4}, \]
\[ P(g(U) = 2) = P\left(\frac{1}{4} < U \leq \frac{3}{4}\right) = \frac{1}{2}, \]
\[ P(g(U) = 3) = P\left(\frac{3}{4} < U \leq 1\right) = \frac{1}{4}. \]

So \( g(U) \) and \( X \) have the same distribution.

Remark: they're not the same random variable! They are just identically distributed. Our goal was to start from a given distribution \((*)\), or equivalently from the corresponding CDF, and find a function \( g \) such that \( g(U) \), with \( U \sim \text{Uniform}(0,1) \), has exactly that distribution.

This can be used for the simulation of random variables. On a computer, we can generate a random number \( U \sim \text{Uniform}(0,1) \), and then apply \( g \) to it: \( g(U) \) will be a r.v. with distribution \((*)\).

How do we see it? I generate a sequence of iid r.v.'s \( U_1, U_2, \ldots, U_n \); I compute \( g(U_1), g(U_2), \ldots, g(U_n) \) and verify that, for large \( n \), I get:

\[
\frac{\# \{ i : g(U_i) = k \}}{n} \sim P(X = k), \quad (***)
\]

as it should by by the frequentist interpretation of probability. (Show simulation).
Matlab code for the simulation of a discrete random variable $Y$
with $P(Y = 1) = \frac{1}{4}$, $P(Y = 2) = \frac{1}{2}$, $P(Y = 3) = \frac{1}{4}$, from Uniform(0,1) samples

```matlab
clear all; close all; format long;

%% prescribed probability distribution
p1 = 1/4; p2 = 1/2; p3 = 1/4;

%% number of samples
n = 10000;

%% Generation of a sequence of n iid RV's $U_i$, Uniform(0,1)
U = rand(n,1);

%% visualization of the uniform random variables
figure; stem(U,'r','fill'); axis([1 n 0 1]); grid on;

%% Computation of $Y_i = g(U_i)$
Y = zeros(n,1);
for i=1:n;
    if U(i) <= p1; Y(i) = 1;
    elseif U(i) <= (p1+p2); Y(i) = 2;
    else; Y(i) = 3;
    end;
end;

%% Computation of the empirical histogram of $Y$
C = zeros(3,1);
for i=1:n;
    if Y(i)==1; C(1,1)=C(1,1)+1;
    elseif Y(i)==2; C(2,1)=C(2,1)+1;
    else C(3,1)=C(3,1)+1;
    end;
end;
P = C/n;  %% empirical histogram

figure; bar(1:3,P(:,1),'b'); grid on;
```

Uniform(0,1) samples $U_i$ (first 200 out of 10000 samples)

Samples $Y_i = g(U_i)$ (first 200 out of 10000 samples)

Empirical distribution of the $Y_i$ samples
(can skip this)

Remark: more precisely, (*** ) happens by the law of large numbers. In fact, consider the event:

\[ A_i = \{ g(U_i) = k \} \]

(for \( k \) fixed). The indicators \( I_{A_1}, I_{A_2}, \ldots, I_{A_n} \)
are independent r.v.'s,

\[ I_{A_i} = \begin{cases} 1 & \text{if } g(U_i) = k \text{ (with prob. } P(X=k)) \\ 0 & \text{otherwise} \end{cases} \]

and identically distributed. By the LLN:

\[ \frac{I_{A_1} + I_{A_2} + \ldots + I_{A_n}}{n} \xrightarrow{n \to \infty} E(I_{A_i}) = P(A_i) = P(g(U_i) = k) = P(X=k), \]

with very large probability.

This is a general procedure that can be used to simulate any discrete distribution, even if defined on a countable set, like the Poisson distribution — in this case you break \((0,1)\) in a countable number of intervals to define \( g \).
What if the CDF $F(x)$ that we start with is continuous?

How do we find a function $g$ such that

$$P(g(U) \leq x) = F(x).$$

Reminder: a function $h: \mathbb{R} \to \mathbb{R}$ is invertible if, for a fixed $y$, there is only one $x$ such that $h(x) = y$. In this case, we write $x = h^{-1}(y)$.

- For example, $h(x) = x^3$ is invertible:
  
  - $y = x^3$ has only one solution,
    
    $$x = \sqrt[3]{y},$$
    
    so $h^{-1}(y) = \sqrt[3]{y}$.

- But $h(x) = x^2$ is not invertible:
  
  for $y$ fixed, $y = x^2$ has two solutions:
  
  $$x_1 = \sqrt{y} \text{ and } x_2 = -\sqrt{y},$$
  
  so $h^{-1}(y)$ is not defined.

- In general, if a function $h$ is monotone increasing ($a < b \Rightarrow h(a) < h(b)$) or decreasing ($a < b \Rightarrow h(a) > h(b)$) then $h$ is invertible.
Distributions of continuous random variables can be:

- **Strictly monotone increasing**: \( a < b \Rightarrow F(a) < F(b) \)
  
  ![Strictly monotone increasing graph](image)

- **Not strictly monotone increasing**: \( a < b \Rightarrow F(a) \leq F(b) \)
  
  ![Not strictly monotone increasing graph](image)

Why all of this? The function \( g \) that we are looking for is **EXACTLY** \( F^{-1} \).

**Theorem.** Given a continuous CDF \( F(x) \), let \( g = F^{-1} \). If \( U \sim \text{Uniform}(0,1) \), then \( g(U) \) has CDF \( F \).

**Proof:**

\[
P( g(U) \leq x ) = P( F^{-1}(U) \leq x ) = P( F(F^{-1}(U)) \leq F(x) ) = P( U \leq F(x) ) = F(x)
\]

\( U \sim \text{Uniform}(0,1) \), so \( P(U \leq a) = a \) for \( 0 \leq a \leq 1 \).
Example: the CDF of an Exponential$(\lambda)$ random variable is:

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Set $F(x) = u$, and solve for $x$: 

$$1 - e^{-\lambda x} = u$$

$$\Rightarrow e^{-\lambda x} = 1 - u, \quad -\lambda x = \ln(1 - u), \quad x = -\frac{1}{\lambda} \ln(1 - u)$$

$$\Rightarrow g(u) = F^{-1}(u) = -\frac{1}{\lambda} \ln(1 - u), \quad u \in (0, 1)$$

Let's verify $g(U) \sim$ Exponential $\lambda$. 

Define: $X = g(U)$ with $U \sim$ Uniform

- **Range of $X$:** $U \in (0, 1) \Rightarrow X \in (0, \infty)$
- **Inversion:** $x = -\frac{1}{\lambda} \ln(1 - u) \iff u = 1 - e^{-\lambda x}$
- **Derivative:** $g'(u) = -\frac{1}{\lambda} \frac{1}{1 - u} (-1) = \frac{1}{\lambda} \frac{1}{1 - u}$
- **Apply formula:** 
  $$f_X(x) = \frac{f_U(u)}{|g'(u)|} = \lambda (1 - u), \quad u \in (0, 1),$$

substitute $u = 1 - e^{-\lambda x}$,

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

Show simulation: $U_i \sim$ Uniform $(0, 1)$, iid, $X_i = g(U_i)$, compute histogram: break $x$-axis into bins

$$\# \{i : n\Delta x < g(U_i) \leq (n+1)\Delta x \}.$$
Matlab code for the simulation of a r.v. $Y \sim \text{Exponential}(\lambda)$ from Uniform(0,1) samples

```matlab
clear all; close all; format long;
%% number of samples
n = 100000;
%% Generation of a sequence of $n$ iid RV’s $U_i$, Uniform(0,1)
U = rand(n,1);
%% visualization of the uniform random variables
figure;
stem(U,'r','fill'); axis([1 n 0 1]); grid on;
%% Computation of $Y_i = -\frac{1}{\lambda} \ln(1 - U_i)$
lambda = 2;
Y = -1/lambda*log(1-U);
figure;
stem(Y,'b','fill'); grid on;
%% Computation of histogram
binsize = .04;
X= binsize/2 : binsize : 40;
H=hist(Y,X)/(n*binsize);
figure;
bar(X,H,'b'); grid on;
axis([0 4/lambda 0 lambda*1.5]);
```

Uniform(0,1) samples $U_i$ (first 200 out of 100000 samples)

Exponential($\lambda$) samples $Y_i = -\frac{1}{\lambda} \ln(1 - U_i)$ (first 200 out of 100000 samples)

Empirical distribution of the $Y_i$ samples, with bin size = 0.04