ORIE 5530: Modeling Under Uncertainty

Lecture 5 (Simulation and Continous Random Variables)

Professor Mark S. Squillante,
Adapted from Professor Itai Gurvich's original notes

5.1 Simulation of discrete random variables

In class we started from Bernoulli random variables and (quickly) built our way up. I just want to provide here the final algorithm. The premise of simulation is that you have a code command that generates a $U[0,1]$ (that is uniform on $[0,1]$) random variable. We want to generate a sample of size n from the distribution of a discrete random variable X. That is, we want to generate a sequence $X_1,\dots,X_n$ but we can only generate a sequence $U_1,\dots,U_n$ from a uniform distribution.

Here is the general algorithm: Suppose you have a random variable taking values $x_1,\dots,x_n$. Let $x_0=-\infty$. Notice that the distribution is given by $p_k=F(x_k)=\sum_{i=1}^{k}\mathbb{P}(X=x_k)$.

Then, do the following.

Algorithm 1 *For $i=1,\dots,n$:*

1. *Generate a $U[0,1]$ random variable U_i (independent of all previous RVs)*
2. *Find k such that $p_{k-1}<U_i\le p_k$. Set $Y_i=x_k$.*

Why does this do the right thing? In other words, if I generated Y_i this way, is the distibution of Y_i the same as that of X? The answer is yes because:

$$\mathbb{P}(Y_i=x_k)=\mathbb{P}(p_{k-1}<U_i\le p_k)=p_k-p_{k-1}=F(x_k)-F(x_{k-1})=\mathbb{P}(X=x_k).$$

Example 1 *With the binomial distribution, the possible outcomes are $x_1=0,x_2=1,\dots,x_{n+1}=n$. Further, $\mathbb{P}(X=i)=\binom{n}{i}p^i(1-p)^{n-i}$ and $p_k=\sum_{i=0}^{k-1}\mathbb{P}(X=i)$.*

5.2 Continuous random variables

The presentation of continuous random variables is aimed at relating the discrete to the continuous by showing that, at an intuitive level, whatever we did with the probability mass function (pmf) of a discrete random variable, we can translate to a continuous random variable by replacing the pmf with the probability density function (pdf).

Definition 1 *We say that X is a continuous random variable if there exists a function $f_X(x)$ such that for any interval $[a,b]$ of the real line*

$$\mathbb{P}(a<x\le b)=\int_a^b f_X(x)dx.$$

We call $f_X(\cdot)$ the probability density function of X.

It may be useful to go back to thinking about sample spaces and events. The sample space of a normal random variable is $S = \{x : -\infty < x < \infty\}$ (namely, the real line). What are events of interest here? The "important" events are intervals of the real line. Namely, we have events of the form $A = [a, b]$ and their probability is determined by the normal density. That is, $\mathbb{P}(A) = \mathbb{P}(a < X \leq b)$ which can be computed as in the definition.

The easiest way to think about density is via the distribution. We define (as before) the distribution (CDF) of a random variable X as

$$F_X(x) = \mathbb{P}(X \leq x).$$

The density is how the distribution grows with x; i.e., it is the derivative:

$$F_X(x) = \int_{-\infty}^{x} f_X(y)dy \Rightarrow f_X(x) = F_X'(x).$$

In some way then f_X captures the local behavior. Remember your basic Taylor theorem? If you have a differentiable function g then $g(y) \approx g(x) + g'(x)(y - x)$ when y is sufficiently close to x. In this way $F_X(x + \epsilon) = F_X(x) + f_X(x)\epsilon$, or equivalently,

$$F_X(x + \epsilon) - F_X(x) = \mathbb{P}(x < X \leq x + \epsilon) \approx f_X(x)\epsilon.$$

Expectations: In the discrete case we had, for a random variable taking values in $\{x_1, \ldots, x_n\}$, $\mathbb{E}[X] = \sum_{k=1}^{n} x_k \mathbb{P}\{X = x_k\}$. Similarly here (with the pmf replaced by the pdf), we have

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} y f_X(y)dy.$$

In general, for a function g,

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(y) f_X(y)dy.$$

Taking this function to be $g(X) = X^2$ we can compute the variance

$$\mathbb{V}\text{ar}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \int_{-\infty}^{\infty} y^2 f_X(y)dy + \left(\int_{-\infty}^{\infty} y f_X(y)dy\right)^2.$$

Some important random variables

- *Uniform:* $X \sim U[a, b]$
  - Density: $f_X(x) = \frac{1}{b-a}$
  - Distribution: $F_X(x) = \frac{x-a}{b-a}$ for $x \in [a, b]$
  - Expectation: $\mathbb{E}[X] = \frac{b-a}{2}$
  - Variance: $\mathbb{V}\text{ar}(X) = \frac{1}{12}(b - a)^2$
- *Exponential:* $X \sim exp(\lambda)$
  - Density: $f_X(x) = \lambda e^{-\lambda x}$
  - Distribution: $F_X(x) = \int_0^x f_X(y)dy = 1 - e^{-\lambda x}$ for $x \geq 0$
  - Expectation: $\mathbb{E}[X] = \frac{1}{\lambda}$
  - Variance: $\mathbb{V}\text{ar}(X) = \frac{1}{\lambda^2}$
- *Normal:* $X \sim N(\mu, \sigma^2)$
  - Density: $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

  - Distribution: $F_X(x) = \int_0^x f_X(y)dy$ (there is no closed form)
  - Expectation: $\mathbb{E}[X] = \mu$
  - Variance: $\mathbb{V}\text{ar}(X) = \sigma^2$

- *Log-Normal:* $X \sim Log\text{-}Normal(\mu, \sigma^2)$
  - Density: $f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(log(x)-\mu)^2}{2\sigma^2}}$
  - Distribution: $F_X(x) = \int_0^x f_X(y)dy$ (there is no closed form)
  - Expectation: $\mathbb{E}[X] = e^{\mu+\frac{1}{2}\sigma^2}$
  - Variance: $\mathbb{V}\text{ar}(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$

The important thing about log-normal is that if X is $log\text{-}normal(\mu, \sigma^2)$ then $Y = log(X) \sim N(\mu, \sigma^2)$.

The most important case is the *standard normal* $X \sim N(0, 1)$ (that is, mean of 0 and variance of 1). Why is that? Because any other normal can be constructed from it. Specifically, if $X \sim N(\mu, \sigma^2)$ then $Y = aX + b \sim N(a\mu + b, a^2\sigma^2)$. To make sure we get the variance right, notice that by linearity of expectation $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$ and by laws of variance

$$\mathbb{V}\text{ar}(aX + b) = \mathbb{V}\text{ar}(aX) = a^2\mathbb{V}\text{ar}(X).$$

So, if you have $X \sim N(0, 1)$, then $Y = \mu + \sigma X \sim N(\mu, \sigma^2)$. Vice versa, if $Y \sim N(\mu, \sigma^2)$ then $X = (Y - \mu)/\sigma \sim N(0, 1)$.

Further note that, if X_1 and X_2 are two independent normal random variables, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, then their sum $X_1 + X_2$ will also be normally distributed, with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$.

5.3 Joint distributions

Oftentimes you have random variables that are dependent. With dependent events, to be able to compute conditional probabilities, for example, we had

$$\mathbb{P}\{A|B\} = \frac{\mathbb{P}\{A \bigcap B\}}{\mathbb{P}\{B\}}.$$

So, letting D be the foot-traffic to a store and W be the weather (say temperatures tomorrow), if I wanted to compute the likelihood that demand exceeds 100 if the temperature is below 0 (say this is what the forecast says), then I want to compute

$$\mathbb{P}\{D > 100|W \le 0\} = \frac{\mathbb{P}\{D > 100, W \le 0\}}{\mathbb{P}\{W \le 0\}}.$$

This means, I want to be able to compute things of the form

$$\mathbb{P}\{X \le x, Y \le y\}, \text{ for all } x, y.$$

This is the *joint distribution.*

If I have discrete random variables X and Y that take on values in $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, respectively, then I can compute the joint distribution from the joint probabilities

$$\mathbb{P}\{X = x_l, Y = y_j\}, l = 1, \ldots, n; j = 1, \ldots, n, \tag{5.1}$$

because

$$\mathbb{P}\{X \leq x_k, Y \leq y_m\} = \sum_{l=1}^{k} \sum_{j=1}^{m} \mathbb{P}\{X = x_l, Y = y_j\}. \tag{5.2}$$

I can also recover the marginal distribution (i.e., of one RV) from the joint distribution. Returning to our Demand and Weather example

$$\mathbb{P}\{D = 100\} = \mathbb{P}\{D = 100, W < \infty\},$$

or similarly

$$\mathbb{P}\{D \leq 100\} = \mathbb{P}\{D \leq 100, W < \infty\}.$$

What in all of the above changes with continuous distributions? Not much really. Now we have a joint density $f_{X,Y}(x,y)$

Here, are the equivalents.

- The joint probabilities (5.1) are replaced by a joint density $f_{X,Y}(x,y)$ (see example of normal below).
- The joint distribution is then computed from the joint density in the same way that (5.2) is computed from (5.1) with integration instead of summation:

  $$F_{X,Y}(x,y) = \mathbb{P}\{X \leq x, Y \leq y\} = \int_{z=-\infty}^{x} \int_{h=-\infty}^{y} f_{X,Y}(z,h) dz dh.$$

  This also means that when you actually have the joint distribution, you can get the joint density by differentiating twice.

  $$f_{X,Y}(a,b) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) \bigg|_{a,b}.$$

  Note: Do not be confused by this superscript X, Y. I just added it so that we remember that this is a joint distribution of the two variable X, Y.

- The marginal distribution can then be recovered by

  $$F_X(x) = \mathbb{P}\{X \leq x\} = F_{X,Y}(x, \infty).$$

A useful example is the normal distribution. Say (X,Y) is a two-variate normal with mean vector $\boldsymbol{\mu} = (\mathbb{E}[X], \mathbb{E}[Y]) = (\mu_X, \mu_Y)$, variance σ_X^2 and σ_Y^2, and correlation ρ. In this case, we have

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \, e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]}. \tag{5.3}$$

What can we do with these joint distributions? Well, sometimes we need them because reality has dependencies and correlation – we want to see the interaction between rain and demand or between the fact that a customer likes a pink shirt and the likelihood he/she will buy a foldable bike. We also want it because we can then (as in the beginning of these notes) build conditioning. **There will be in the third homework a couple of questions that will help you practice this.**

Independence: Two random variables are independent if $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ for any values of x, y or in "words" if $\mathbb{P}\{X \leq x, Y \leq y\} = \mathbb{P}\{X \leq x\}\mathbb{P}\{Y \leq y\}$ for any x, y.

For the discrete case, this implies and is implied by the condition that $\mathbb{P}\{X = x_k, Y = y_l\} = \mathbb{P}\{X = x_k\}\mathbb{P}\{, Y = y_l\}$ for all possible values.

Similarly, for the continuous case, this implies and is implied by

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

(that is, the joint density is the product of the densities). Notice that this holds in the normal example above if you set the correlation to $\rho = 0$.

Bottom line: If you want to check independence it suffices that

- Discrete: $\mathbb{P}\{X = x_k, Y = y_l\} = \mathbb{P}\{X = x_k\}\mathbb{P}\{, Y = y_l\}$ for all possible values x_k, y_l.
- Continuous: $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all possible x, y.

5.4 Simulation of continuous random variables

One idea of simulating continuous random variables is in fact very simple and has to do with inverting the distribution function. Specifically, say you want to create a family of independent random variables $X_1, \ldots, X_n$ that follow the exponential distribution $F(x) = 1 - e^{-\lambda x}$ and, as before, the only thing your computer knows how to do is generate uniform random variables.

The basic idea is as follows: if you take a family of independent $uniform[0,1]$ random variables $U_1, \ldots, U_n$, then $F^{-1}(U_1), \ldots, F^{-1}(U_n)$ follow the exponential distribution. (You can visualize this in excel: in column A you have a bunch of rand() commands and in column B you put the function $F^{-1}(A1)$ in the first row, $F^{-1}(A2)$ in the second row, etc.).

Why does it do the right thing? Put another way, why is it the case that $F^{-1}(U)$ is a random variable with the distribution $F(x)$?

This is because (recall $\mathbb{P}\{U \leq b\} = b$):

$$\mathbb{P}\{F^{-1}(U) \leq x\} = \mathbb{P}\{U \leq F(x)\} = F(x).$$

The only challenge with this is that, except for some simple cases, inverting also involves computational work. But, sometimes you can invert by hand as in the case of exponential:

$$y = F(x) = 1 - e^{-\lambda x} \Rightarrow x = -\frac{1}{\lambda}\log(1-y).$$

So we simulate a bunch of uniform random variables $U_1, \ldots, U_n$ and then the values we use are

$$-\frac{1}{\lambda}\log(1-U_1), \ldots, -\frac{1}{\lambda}\log(1-U_n).$$

In fact one can replace here $1 - U_i$ with U_i (convince yourself that if U is $Uniform[0,1]$ so is $1 - U$).

Why the hell is this of any use? Well, say we want to simulate a queue where customer service times follow an exponential distribution. This will tell us how to create the service times.

A small remark: When I say that I am generating a draw from an exponential distribution, what does this mean? After all, I am just generating one number. The "meaning" here is the frequentist view just as the one we apply intuitively with a coin. In the case of a coin, "generating a draw" is just flipping the coin. When I say that this flip follows a Bernoulli distribution with $p = 1/2$, what I mean is that if I do this many times, I will get heads half of the time.

It is the same with the exponential distribution. If I simulate many independent exponential random variables, I should have that

$$\% \text{ of draws } \leq x \approx \mathbb{P}\{X \leq x\} = 1 - e^{-\lambda x}.$$

Lastly, let us assume that we have code that generates **independent standard normal random variables**. How do we generate a vector of dependent normal random variables?

Definition 2 *A multivariate (d-dimensional) normal $X = (X_1, \ldots, X_d)$ with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ and $d \times d$ covariance matrix Σ (where the entry Σ_{ij} is the covariance of X_i and X_j, and the entry $\Sigma_{ii} = \sigma_i^2$ is $\mathbb{V}ar(X_i)$) is a continuous RV with the density*

$$f_X(x) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|\Sigma|}} e^{-\frac{1}{2}(x-\boldsymbol{\mu})^t \Sigma^{-1}(x-\boldsymbol{\mu})},$$

where $|\Sigma|$ is the determinant of the matrix Σ. We write $X \sim N(\boldsymbol{\mu}, \Sigma)$.

In the special bi-variate case (i.e., $d = 2$), the determinant of Σ equals $\sigma_1^2\sigma_2^2 - \rho^2\sigma_1^2\sigma_2^2 = (1-\rho^2)\sigma_1^2\sigma_2^2$ and the inverse of Σ is

$$\Sigma^{-1} = \frac{1}{|\Sigma|}\left[\begin{array}{cc} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{array}\right],$$

where ρ is the correlation and hence we obtain (5.3) from the above definition. Notice, in general, that it suffices if we are given the variances (hence standard deviations) and the correlation matrix $\rho = [\rho_{ij}]$ because then we have the covariance matrix $\Sigma_{ij} = \mathbb{C}\mathrm{ov}(X_i, X_j) = \rho_{ij}\sigma_i\sigma_j$.

The basic idea for simulating correlated multivariate $X = (X_1, \ldots, X_d)$ from a vector of independent standard normals $N = (N_1, \ldots, N_d)$ is built on the multivariate version of the basic fact that, taking a standard normal N, $Y = aN + b$ is normal with mean b and variance a^2.

There is a bit of algebra involved below. To make our life easier as we move to the multidimensional case let us first assume $\boldsymbol{\mu} = 0$ (zero mean). Also, let us just look at the two-dimensional case first.

Let us consider a 2×2 matrix L (we will shortly say what this matrix L is) and define

$$X = LN = \left[\begin{array}{cc} L_{11} & L_{12} \\ L_{21} & L_{22} \end{array}\right]\left[\begin{array}{c} N_1 \\ N_2 \end{array}\right] = \begin{array}{c} L_{11}N_1 + L_{12}N_2 \\ L_{21}N_1 + L_{22}N_2 \end{array}.$$

What is the covariance matrix of L? Using the independence of N_1 and N_2, we have

$\Sigma_{11} = \mathbb{V}\mathrm{ar}(X_1) = L_{11}^2 + L_{12}^2$, $\Sigma_{22} = \mathbb{V}\mathrm{ar}(X_2) = L_{21}^2 + L_{22}^2$ and

$$\begin{aligned}
\Sigma_{12} = \Sigma_{21} &= \mathbb{C}\mathrm{ov}(X_1, X_2) = \mathbb{C}\mathrm{ov}(L_{11}N_1 + L_{12}N_2, L_{21}N_1 + L_{22}N_2) \\
&= L_{11}L_{21}\mathbb{C}\mathrm{ov}(N_1, N_1) + L_{11}L_{22}\mathbb{C}\mathrm{ov}(N_1, N_2) + L_{12}L_{21}\mathbb{C}\mathrm{ov}(N_1, N_2) + L_{12}L_{22}\mathbb{C}\mathrm{ov}(N_2, N_2) \\
&= L_{11}L_{21}\mathbb{V}\mathrm{ar}(N_1) + L_{12}L_{22}\mathbb{V}\mathrm{ar}(N_2) \\
&= L_{11}L_{21} + L_{12}L_{22},
\end{aligned}$$

where I use the rules of covariance:

- $\mathbb{C}\mathrm{ov}(X, X) = \mathbb{V}\mathrm{ar}(X)$;
- $\mathbb{C}\mathrm{ov}(X, Y) = \mathbb{C}\mathrm{ov}(Y, X)$;
- $\mathbb{C}\mathrm{ov}(aX, bY) = ab\mathbb{C}\mathrm{ov}(X, Y)$;
- $\mathbb{C}\mathrm{ov}(X + Y, Z) = \mathbb{C}\mathrm{ov}(X, Z) + \mathbb{C}\mathrm{ov}(Y, Z)$;
- If X, Y are independent $\mathbb{C}\mathrm{ov}(X, Y) = 0$.

So, if I am given a Σ, I would want to choose L such that

$$\Sigma = \begin{bmatrix} L_{11}^2 + L_{12}^2 & L_{11}L_{21} + L_{12}L_{22} \\ L_{11}L_{21} + L_{12}L_{22} & L_{21}^2 + L_{22}^2 \end{bmatrix} = \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix}^t = LL^t.$$

So, in the case of $\boldsymbol{\mu} = 0$, we have arrived at the following recipe to simulate a multivariate normal random variable with covariance matrix Σ.

1. Compute the matrix L such that $\Sigma = LL^t$.
2. Generate a standard normal vector $N_1, \ldots, N_d$.
3. Assign $X = LN$.

If you do this, $X = (X_1, \ldots, X_d)$ will follow the desired normal distribution. Having a non-zero mean $\boldsymbol{\mu}$ is now easy. Replace the last step with

3. Assign $X = \boldsymbol{\mu} + LN$.

A final word about implementation. How do we find the matrix L such that $\Sigma = LL^t$? Well, the matrix Σ is a special case of what is called a positive semi-definite matrix. It turns out that, for such matrices, you can always find a "square-root" matrix L as above. The procedure to do so is called Cholesky factorization and there is a command in Python's scipy.linalg library's that does exactly that. In that command you can choose whether you want to have L upper or lower triangular and you can always choose lower.