Simulation - Lecture 2 - Inversion and transformation methods

Lecture version: Monday 27^{th} January, 2020, 23:25

Robert Davies

Part A Simulation and Statistical Programming

Hilary Term 2020

Recap from previous lecture

- Examples of distributions from different fields we might be interested in studying
- Monte Carlo
  - Suppose $X \sim dist$, and we have a method to simulate iid random variables $X_i \sim dist$
  - Then $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i)$ is an unbiased estimator of $\mathbb{E}\left(\phi(X)\right)$
  - We can form a confidence interval for θ using the sample variable $S^2_{\phi(X)}$ using the central limit theorem
- Rest of simulation lectures
  - How do we generate $X \sim dist$ in the real world for increasingly complicated distributions
  - Today: Inversion, the simplest case, when the CDF is well behaved
  - Also today: Transformation, when you can build your distribution from distributions that are well behaved

A quick note about pseudo-random numbers

- We seek to be able to generate complicated random variables and stochastic models.
- Henceforth, we will assume that we have access to a sequence of independent random variables $(U_i, i \geq 1)$ that are uniformly distributed on $(0, 1)$; i.e. $U_i \sim \mathcal{U}[0, 1]$.
- In R, the command `u←runif(100)` return 100 realizations of uniform r.v. in $(0, 1)$.
- Strictly speaking, we only have access to pseudo-random (deterministic) numbers.
- The behaviour of modern random number generators (constructed on number theory) resembles mathematical random numbers in many respects: standard statistical tests for uniformity, independence, etc. do not show significant deviations.

Outline

Inversion Method

Transformation Methods

Recap of CDF definition

- A function $F : \mathbb{R} \to [0, 1]$ is a cumulative distribution function (cdf) if
  - F is increasing; i.e. if $x \leq y$ then $F(x) \leq F(y)$
  - F is right continuous; i.e. $F(x + \epsilon) \to F(x)$ as $\epsilon \to 0$ ($\epsilon > 0$)
  - $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to +\infty$.
- A random variable $X \in \mathbb{R}$ has cdf F if $\mathbb{P}\left(X \leq x\right) = F(x)$ for all $x \in \mathbb{R}$.
- If F is differentiable on $\mathbb{R}$, with derivative f, then X is continuously distributed with probability density function (pdf) f.

The CDF of a random variable has a uniform distribution

- **Proposition**. Let F be a continuous and strictly increasing cdf on $\mathbb{R}$, with inverse $F^{-1} : [0,1] \to \mathbb{R}$. Then the random variable $F(X)$ has a uniform distribution on $[0,1]$.
- Proof. Let $y \in [0,1]$. Then

$$P(F(X) \leq y) = P(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y$$

and so $F(X) \sim \mathcal{U}[0,1]$

The inverse of the CDF applied to uniforms generates random variables from the CDF

- **Proposition**. Let F be a continuous and strictly increasing cdf on $\mathbb{R}$, with inverse $F^{-1} : [0,1] \to \mathbb{R}$. Let $U \sim \mathcal{U}[0,1]$ then $X = F^{-1}(U)$ has cdf F.
- Proof. Let $x \in \mathbb{R}$. Then we have

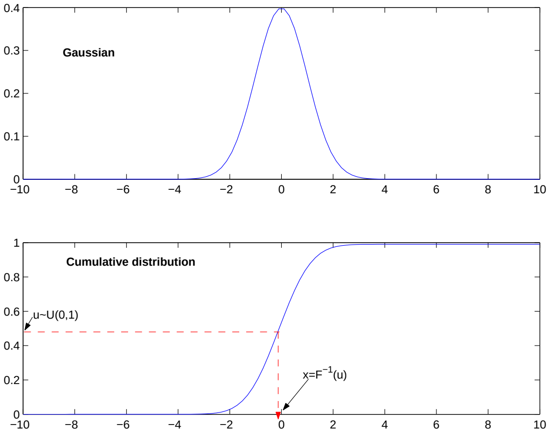
$$\begin{aligned} \mathbb{P}(X \leq x) &= \mathbb{P}\left(F^{-1}(U) \leq x\right) \\ &= \mathbb{P}(U \leq F(x)) \\ &= F(x). \end{aligned}$$

Inversion method

Algorithm 1 Inversion method

- Given CDF F, calculate F^{-1}
- Simulate independent $U_i \sim \mathcal{U}[0,1]$
- Return $X_i = F^{-1}(U_i) \sim F$

Illustrative example of inversion method using Gaussian distribution



Top: pdf of a Gaussian r.v., bottom: associated cdf.

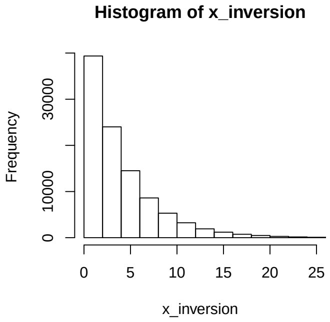
Exponential distribution example

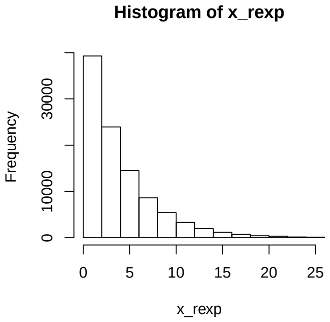
- *Exponential distribution*. Let $\lambda > 0$. Then the exponential CDF is given by
$$F(x) = 1 - e^{-\lambda x}$$

  We calculate
$$\begin{aligned} u &= F(x) \\ u &= 1 - e^{-\lambda x} \\ \implies \log(1-u) &= -\lambda x \\ \implies x &= -\frac{\log(1-u)}{\lambda} \end{aligned}$$

Exponential rvs using the inversion method

```
set.seed(9119)
lambda <- 0.25
n <- 100000
u <- runif(n)
x_inversion <- -log(1 - u) / lambda
x_rexp <- rexp(n = n, rate = lambda)
wilcox.test(x_inversion, x_rexp)$p.value # 0.46
```





Examples

- *Cauchy distribution*. It has pdf and cdf

$$f(x) = \frac{1}{\pi\left(1+x^2\right)},\ F(x) = \frac{1}{2} + \frac{arc\tan x}{\pi}$$

  We have

$$\begin{aligned} u &= F(x) \Leftrightarrow u = \frac{1}{2} + \frac{arc\tan x}{\pi} \\ &\Leftrightarrow x = \tan\left(\pi\left(u - \frac{1}{2}\right)\right) \end{aligned}$$

- *Logistic distribution.* It has pdf and cdf

$$\begin{aligned} f(x) &= \frac{\exp(-x)}{\left(1+\exp(-x)\right)^2},\ F(x) = \frac{1}{1+\exp(-x)} \\ &\Leftrightarrow x = \log\left(\frac{u}{1-u}\right). \end{aligned}$$

- Practice: Derive an algorithm to simulate from an Weibull random variable with rates $\alpha, \lambda > 0$

Definition of the discrete CDF inverse

- **Proposition**. Let F be a cdf on $\mathbb{R}$ and define its generalized inverse $F^{-1} : [0, 1] \rightarrow \mathbb{R}$,

$$F^{-1}(u) = \inf \{x \in \mathbb{R}; F(x) \geq u\}.$$

Let $U \sim \mathcal{U}[0, 1]$ then $X = F^{-1}(U)$ has cdf F.

Discrete $\mathbb{N}-$r.v. CDF

- If X is a discrete $\mathbb{N}-$r.v. with $\mathbb{P}(X = n) = p(n)$, we get $F(x) = \sum_{j=0}^{\lfloor x \rfloor} p(j)$ and $F^{-1}(u)$ is $x \in \mathbb{N}$ such that
$$\sum_{j=0}^{x-1} p(j) < u \leq \sum_{j=0}^{x} p(j)$$
with the LHS$= 0$ if $x = 0$.
- Note: the mapping at the values $F(n)$ are irrelevant (0 probability of getting a single point)
- Note: the same method is applicable to any discrete valued r.v. X, $\mathbb{P}(X = x_n) = p(n)$.

Example code for simple discrete rv

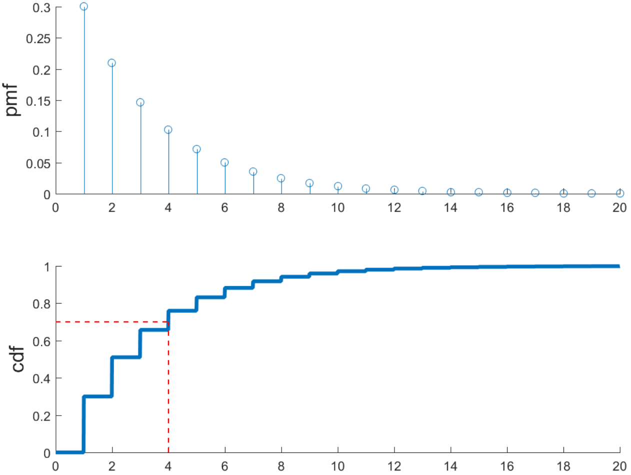
```
p <- c(0.5, 0.3, 0.2) ## pmf
p_norm <- c(0, cumsum(p)) ## 0.0 0.5 0.8 1.0
m <- length(p)
n <- 100000
u <- runif(n)
x <- array(NA, n)
for(i in 1:n) {
    for(j in 1:m) {
        if ((p_norm[j] < u[i]) & (u[i] <= p_norm[j + 1])) {
            x[i] <- j
        }
    }
}
sum(is.na(x)) ## 0
table(x)
##     1     2     3
## 50227 30105 19668
```

Example: Geometric Distribution

- If $0 < p < 1$ and $q = 1 - p$ and we want to simulate $X \sim \mathrm{Geom}(p)$ then
$$p(x) = pq^{x-1}, F(x) = 1 - q^x \qquad x = 1, 2, 3...$$

- The smallest $x \in \mathbb{N}$ giving $F(x) \geq u$ is the smallest $x \geq 1$ satisfying
$$x \geq \log(1-u)/\log(q)$$
and this is given by
$$x = F^{-1}(u) = \left\lceil \frac{\log(1-u)}{\log(q)} \right\rceil$$
where $\lceil x \rceil$ rounds up and we could replace $1 - u$ with u.

Illustration of the Inversion Method: Discrete case



Outline

Inversion Method

Transformation Methods

Transformation Methods

- Suppose we
  - Have a random variable $Y \sim Q$, $Y \in \Omega_Q$, which we **can** simulate (*e.g.*, by inversion)
  - Have a random variable $X \sim P$, $X \in \Omega_P$, which we **wish** to simulate
  - Can find a function $\varphi : \Omega_Q \to \Omega_P$ with the property that if $Y \sim Q$ then $X = \varphi(Y) \sim P$.
- Then we can simulate from X by first simulating $Y \sim Q$, and then set $X = \varphi(Y)$.
- Inversion is a special case of this idea.
- We may generalize this idea to take functions of collections of variables with different distributions.

Transformation method

Algorithm 2 Transformation method

- Find $Y \sim Q$ that you can simulate from, and a function φ such that $X = \varphi(Y) \sim P$
- Simulate independent $Y_i \sim Q$
- Return $X_i = \varphi(Y_i) \sim P$

Exponential to gamma example

- Example: Let Y_i, $i = 1, 2, ..., \alpha$, be iid variables with $Y_i \sim \mathrm{Exp}(1)$ and $X = \beta^{-1} \sum_{i=1}^{\alpha} Y_i$ then $X \sim \mathrm{Gamma}(\alpha, \beta)$.
  Proof: The MGF of the random variable X is

$$\mathbb{E}\left(e^{tX}\right) = \prod_{i=1}^{\alpha} \mathbb{E}\left(e^{\beta^{-1} t Y_i}\right) = (1 - t/\beta)^{-\alpha}$$

  which is the MGF of a $\mathrm{Gamma}(\alpha, \beta)$ variable.
  Incidentally, the $\mathrm{Gamma}(\alpha, \beta)$ density is $f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ for $x > 0$.

Transformation Methods: Box-Muller Algorithm

- **Proposition**. If $R^2 \sim \text{Exp}(\frac{1}{2})$ and $\Theta \sim \mathcal{U}[0, 2\pi]$ are independent then $X = R\cos\Theta$, $Y = R\sin\Theta$ are independent with $X \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{N}(0,1)$.
  Proof: We have $f_{R^2,\Theta}(r^2\theta) = \frac{1}{2}\exp\left(-r^2/2\right)\frac{1}{2\pi}$ and therefore we are interested in

$$f_{X,Y}(x,y) = f_{R^2,\Theta}(r^2(x,y), \theta(x,y))\left|\det\frac{\partial(r^2,\theta)}{\partial(x,y)}\right|$$

  where

$$\left|\det\frac{\partial(r^2,\theta)}{\partial(x,y)}\right| = \left|\begin{array}{cc} \frac{\partial r^2}{\partial x} & \frac{\partial r^2}{\partial y} \\ \frac{\partial \theta}{\partial x} & \frac{\partial \theta}{\partial y} \end{array}\right| = 2$$

$$\Longrightarrow f_{X,Y}(x,y) = \frac{1}{2}e^{-\frac{1}{2}(x^2+y^2)}\frac{1}{2\pi}2 = \left(\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\right)\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}\right)$$

Transformation Methods: Box-Muller Algorithm, applied

- Let $U_1 \sim \mathcal{U}[0,1]$ and $U_2 \sim \mathcal{U}[0,1]$ then

$$\begin{aligned} R^2 &= -2\log(U_1) \sim \mathrm{Exp}\left(\frac{1}{2}\right) \\ \Theta &= 2\pi U_2 \sim \mathcal{U}[0,2\pi] \end{aligned}$$

  and

$$\begin{aligned} X &= R\cos\Theta \sim \mathcal{N}(0,1) \\ Y &= R\sin\Theta \sim \mathcal{N}(0,1), \end{aligned}$$

- Note this still requires evaluating $\log, \cos$ and $\sin$.

Box Muller applied

```
set.seed(913)
n <- 100000
u1 <- runif(n)
u2 <- runif(n)
lambda <- 1 / 2
r2 <- -log(1 - u1) / lambda ## are now Exp(1/2)
theta <- 2 * pi * u2 ## U[0, 2*pi]
r <- sqrt(r2)
x <- r * cos(theta)
y <- r * sin(theta)
round(c(mean(x), var(x)), 3) ##  -0.001  0.998
round(c(mean(y), var(y)), 3) ##  -0.003  1.000
cor(x, y) ## -0.0006317268
```

Simulating Multivariate Normal

- Let consider $X \in \mathbb{R}^d$, $X \sim N(\mu, \Sigma)$ where μ is the mean and Σ is the (positive definite) covariance matrix.

$$f_X(x) = (2\pi)^{-d/2} |\det \Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right).$$

- **Proposition**. Let $Z = (Z_1, ..., Z_d)$ be a collection of d independent standard normal random variables. Let L be a real $d \times d$ matrix satisfying

$$LL^T = \Sigma,$$

and

$$X = LZ + \mu.$$

Then

$$X \sim \mathcal{N}(\mu, \Sigma).$$

Simulating Multivariate Normal proof

- Proof. We have $f_Z(z) = (2\pi)^{d/2} \exp\left(-\frac{1}{2} z^T z\right)$.The joint density of the new variables is

$$f_X(x) = f_Z(z) \left| \det \frac{\partial z}{\partial x} \right|$$

where $\frac{\partial z}{\partial x} = L^{-1}$ and $\det(L) = \det(L^T)$ so $\det(L^2) = \det(\Sigma)$, and $\det(L^{-1}) = 1/\det(L)$ so $\det(L^{-1}) = \det(\Sigma)^{-1/2}$. Also

$$\begin{aligned} z^T z &= (x-\mu)^T \left(L^{-1}\right)^T L^{-1} (x-\mu) \\ &= (x-\mu)^T \Sigma^{-1} (x-\mu) \,. \end{aligned}$$

- If $\Sigma = VDV^T$ is the eigendecomposition of Σ, we can pick $L = VD^{1/2}$.
- Cholesky factorization $\Sigma = LL^T$ where L is a lower triangular matrix.

Recap

- Monte Carlo is useful but requires simulated random variables
- Assume we can always drawn uniform random variables
- **Inversion method** For continuous strictly increasing CDFs we can draw X_i as $F^{-1}(U_i)$
- We can do the same thing for discrete distributions
- **Transformation method** If we can find φ for some distribution Y_i such that $X_i = \varphi(Y_i)$, then we can simulate X_i in that way