Simulation

Loukia Meligkotsidou,
National and Kapodistrian University of Athens

MSc in Statistics and Operational Research,
Department of Mathematics

Motivation

Many applications of simulation are based on the idea of using samples from a distribution in order to approximate characteristics of this distribution. As a trivial example, suppose that we are interested in the mean of a real random variable X with distribution function F and probability density function f, i.e.

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

However, it might be difficult or impossible to perform the above integration. Suppose that a random sample $x_1, \ldots, x_n$ was available where each x_i is a realisation of $X_i \sim F$, $i = 1, \ldots, n$. Then we could approximate the mean of X by the observed sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

This result, which is routinely used in statistical estimation, will be shown to be related with the so-called Monte Carlo integration.

Some limit theorems

The **Law of Large Numbers** is the main result which validates using sample averages to estimate population means.

Some limit theorems

The **Law of Large Numbers** is the main result which validates using sample averages to estimate population means.

Let $X_1, \ldots, X_n$ be a sequence of random variables, and $\theta_n = \theta_n(X_1, \ldots, X_n)$ be a real function of them (e.g $\theta_n = n^{-1} \sum_{i=1}^{n} X_i$). We say that θ_n converges in mean square sense to a fixed value θ if

$$E(\theta_n - \theta)^2 \to 0, \text{ as } n \to \infty.$$

Some limit theorems

The **Law of Large Numbers** is the main result which validates using sample averages to estimate population means.

Let $X_1, \ldots, X_n$ be a sequence of random variables, and $\theta_n = \theta_n(X_1, \ldots, X_n)$ be a real function of them (e.g $\theta_n = n^{-1} \sum_{i=1}^{n} X_i$). We say that θ_n converges in mean square sense to a fixed value θ if

$$E(\theta_n - \theta)^2 \to 0, \text{ as } n \to \infty.$$

Theorem1. A Law of Large Numbers. Let $X_1, \ldots, X_n$ be a sequence of independent random variables with common means $E(X_i) = \theta$ and variances $\text{Var}(X_i) = \sigma^2 < \infty$. If $\theta_n = n^{-1} \sum_{i=1}^{n} X_i$ then

$$E(\theta_n - \theta)^2 = \sigma^2/n \to 0, \text{ as } n \to \infty.$$

Some limit theorems

Although Theorem 1 suggests that sample averages will converge to the population mean (with rate $1/n$), the following theorem gives more information about the (asymptotic) error of θ_n in estimating θ.

Some limit theorems

Although Theorem 1 suggests that sample averages will converge to the population mean (with rate $1/n$), the following theorem gives more information about the (asymptotic) error of θ_n in estimating θ.

Theorem 2. A Law of Large Numbers. Let $X_1, \ldots, X_n$ be a sequence of independent random variables with common means $E(X_i) = \theta$ and variances $\text{Var}(X_i) = \sigma^2$. If $\theta_n = n^{-1} \sum_{i=1}^{n} X_i$ then

$$\frac{\sqrt{n}}{\sigma}(\theta_n - \theta)$$

is approximately distributed as a $N(0, 1)$ r.v., as $n \to \infty$.

Some limit theorems

Although Theorem 1 suggests that sample averages will converge to the population mean (with rate $1/n$), the following theorem gives more information about the (asymptotic) error of θ_n in estimating θ.

Theorem 2. A Law of Large Numbers. Let $X_1, \ldots, X_n$ be a sequence of independent random variables with common means $E(X_i) = \theta$ and variances $\text{Var}(X_i) = \sigma^2$. If $\theta_n = n^{-1} \sum_{i=1}^{n} X_i$ then

$$\frac{\sqrt{n}}{\sigma}(\theta_n - \theta)$$

is approximately distributed as a $N(0, 1)$ r.v., as $n \to \infty$.

Therefore, we wish to have a small σ in order to obtain accurate estimates of θ by θ_n.

Some limit theorems

Although Theorem 1 suggests that sample averages will converge to the population mean (with rate $1/n$), the following theorem gives more information about the (asymptotic) error of θ_n in estimating θ.

Theorem 2. A Law of Large Numbers. Let $X_1, \ldots, X_n$ be a sequence of independent random variables with common means $E(X_i) = \theta$ and variances $\text{Var}(X_i) = \sigma^2$. If $\theta_n = n^{-1} \sum_{i=1}^{n} X_i$ then

$$\frac{\sqrt{n}}{\sigma}(\theta_n - \theta)$$

is approximately distributed as a $N(0, 1)$ r.v., as $n \to \infty$.

Therefore, we wish to have a small σ in order to obtain accurate estimates of θ by θ_n.

Similar results show that sample quantiles converge to population quantiles as the sample size increases.

Issues in Simulation

It is reasonable to use samples from a distribution in order to approximate characteristics of that distribution. Of key importance becomes **how** to generate samples from a specified distribution.

Issues in Simulation

It is reasonable to use samples from a distribution in order to approximate characteristics of that distribution. Of key importance becomes **how** to generate samples from a specified distribution.

Using basic results from probability theory, we will give some answers to the following two questions:

1. How to do it; and
2. How to do it efficiently.

Issues in Simulation

It is reasonable to use samples from a distribution in order to approximate characteristics of that distribution. Of key importance becomes **how** to generate samples from a specified distribution.

Using basic results from probability theory, we will give some answers to the following two questions:

1. How to do it; and
2. How to do it efficiently.

The **raw material** for any simulation is **uniformly distributed random numbers**. Transformation or other manipulation can then be applied to simulate from more complex distributions.

Issues in Simulation

It is reasonable to use samples from a distribution in order to approximate characteristics of that distribution. Of key importance becomes **how** to generate samples from a specified distribution.

Using basic results from probability theory, we will give some answers to the following two questions:

1. How to do it; and
2. How to do it efficiently.

The **raw material** for any simulation is **uniformly distributed random numbers**. Transformation or other manipulation can then be applied to simulate from more complex distributions.

The standard approach to generating uniform random numbers is to use a **deterministic** algorithm which, however, produces a sequence of numbers which do not exhibit any obvious pattern: they can pass all the statistical tests of randomness (**pseudo-random numbers**).

Inversion Method

Suppose we wand to simulate draws from a distribution with distribution function F.

The Inversion method is the simplest of all procedures, and is nothing more than a straightforward application of the probability integral transform:

if $X \sim F$, then $F(X) \sim U[0,1]$, so by inversion if $U \sim U[0,1]$, then $F^{-1}(U) \sim F$.

Inversion Method

Suppose we wand to simulate draws from a distribution with distribution function F.

The Inversion method is the simplest of all procedures, and is nothing more than a straightforward application of the probability integral transform:

if $X \sim F$, then $F(X) \sim U[0,1]$, so by inversion if $U \sim U[0,1]$, then $F^{-1}(U) \sim F$.

Thus, defining $x_i = F^{-1}(u_i)$, generates a sequence of independent realizations from F.

Inversion Method

Suppose we wand to simulate draws from a distribution with distribution function F.

The Inversion method is the simplest of all procedures, and is nothing more than a straightforward application of the probability integral transform:

if $X \sim F$, then $F(X) \sim U[0,1]$, so by inversion if $U \sim U[0,1]$, then $F^{-1}(U) \sim F$.

Thus, defining $x_i = F^{-1}(u_i)$, generates a sequence of independent realizations from F.

For example, to simulate from the exponential distribution we have $F(x) = 1 - \exp(-\lambda x)$, so $F^{-1}(u) = -\lambda^{-1} \log(1-u)$.

Transformation of variables

Suppose that we want to simulate a random variable $X \sim F$, and we know that it can be written as $X = t(Y)$, where $Y \sim G$ is another random variable and $t(\cdot)$ is some function. Then, defining $x_i = t(y_i)$ where y_i has been generated from G, generates a sequence of independent realizations from F.

Transformation of variables

Suppose that we want to simulate a random variable $X \sim F$, and we know that it can be written as $X = t(Y)$, where $Y \sim G$ is another random variable and $t(\cdot)$ is some function. Then, defining $x_i = t(y_i)$ where y_i has been generated from G, generates a sequence of independent realizations from F.

As an example, suppose that $X \sim Exp(\theta)$, that is X has the exponential distribution with mean $1/\theta > 0$, then $X = \theta^{-1}Y$, where $Y \sim Exp(1)$. Therefore, we can simulate X by first simulating a standard exponential random variable, for example using the inversion method, and then dividing it by θ.

Transformation of variables

Suppose that we want to simulate a random variable $X \sim F$, and we know that it can be written as $X = t(Y)$, where $Y \sim G$ is another random variable and $t(\cdot)$ is some function. Then, defining $x_i = t(y_i)$ where y_i has been generated from G, generates a sequence of independent realizations from F.

As an example, suppose that $X \sim Exp(\theta)$, that is X has the exponential distribution with mean $1/\theta > 0$, then $X = \theta^{-1}Y$, where $Y \sim Exp(1)$. Therefore, we can simulate X by first simulating a standard exponential random variable, for example using the inversion method, and then dividing it by θ.

The function $t(\cdot)$ can be a many-to-one function. For example, suppose that $X \sim Gamma(m, \theta)$, that is X has the gamma distribution with mean m/θ and variance m/θ^2, where m is some integer. It is known that $X = t(Y_1, \ldots, Y_m) = \sum_{i=1}^{m} Y_i$, where the Y_is are i.i.d $Exp(\theta)$ random variables. Thus, X can be simulated by simulating m independent Y_is and then summing them up.

Rejection Sampling

The idea in rejection sampling is to simulate from a distribution which is easy to sample from, but then to only accept that simulated value with some probability p. By choosing p correctly, we can ensure that the sequence of accepted simulated values are from the **desired distribution**.

Rejection Sampling

The idea in rejection sampling is to simulate from a distribution which is easy to sample from, but then to only accept that simulated value with some probability p. By choosing p correctly, we can ensure that the sequence of accepted simulated values are from the **desired distribution**.

It can be proved that points can be simulated uniformly inside a set C (e.g a star-shaped set) by simulating uniformly points on a superset $C \subset B$ (e.g the unit square) and rejecting those which fall outside C. The hope is that it is easier to simulate points uniformly inside B than inside C. In our example, it is trivial to simulate a point uniformly inside B, simply take (U, V), where U, V are independent $U[0, 1]$ variables. Notice that if B is much larger than C then most of the simulated points will be rejected, thus it is desirable to find B so that to match closely C.

Rejection Sampling

We now turn to the problem of simulating from a given density f, and link this problem with the rejection idea described above. Suppose that f is a probability density function on the real line. Moreover, let (X, Y) be a uniformly distributed point under the density f. Therefore, if $h_{(X,Y)}(x, y)$ denotes the probability density function of (X, Y), then

$$h_{(X,Y)}(x, y) = 1, \quad -\infty < x < \infty, 0 < y < f(x).$$

It is easy to show (check!) that the marginal density of X is exactly f!

Rejection Sampling

We now turn to the problem of simulating from a given density f, and link this problem with the rejection idea described above. Suppose that f is a probability density function on the real line. Moreover, let (X, Y) be a uniformly distributed point under the density f. Therefore, if $h_{(X,Y)}(x, y)$ denotes the probability density function of (X, Y), then

$$h_{(X,Y)}(x, y) = 1, \quad -\infty < x < \infty, 0 < y < f(x).$$

It is easy to show (check!) that the marginal density of X is exactly f!

The rejection sampling idea can be used to generate samples uniformly under f!

Rejection Sampling

Suppose that we want to simulate from $f(x)$ and that $g(x)$ is another density function, which is easy to sample from, and K a constant such that

$$f(x) \leq Kg(x), \text{ for all } x,$$

therefore the set of points under f lies entirely within the set of points under Kg.

Rejection Sampling

Suppose that we want to simulate from $f(x)$ and that $g(x)$ is another density function, which is easy to sample from, and K a constant such that

$$f(x) \leq Kg(x), \text{ for all } x,$$

therefore the set of points under f lies entirely within the set of points under Kg.

A point (X^*, Y^*) can be simulated uniformly under Kg, by sampling X^* from g, and then simulating Y^* given X^* from a uniform distribution $U[0, Kg(X^*)]$. If we keep all such points that fall under f, i.e all those (X^*, Y^*) such that $Y^* < f(X^*)$, we will obtain a sample of points uniformly distributed under f. The X^*s of this sample are an independent sample from f.

The Algorithm

1. Simulate x^* from $g(x)$.
2. Simulate y^* from $U(0, Kg(x^*))$.
3. Accept x^* if $y^* \leq f(x^*)$.
4. Continue.

Furthermore, $\Pr(X \text{ accepted}) = 1/K$.

The Algorithm

1. Simulate x^* from $g(x)$.
2. Simulate y^* from $U(0, Kg(x^*))$.
3. Accept x^* if $y^* \leq f(x^*)$.
4. Continue.

Furthermore, $\Pr(X \text{ accepted}) = 1/K$.

Note that we only need to know f up to a normalizing constant! This is a big advantage of the method since in many statistical applications we need to simulate from densities for which we cannot compute the normalizing constant.

The efficiency of the procedure depends on the agreement between f and the envelope Kg since if a large value of K is necessary, then the acceptance probability is low, so that large numbers of simulations are needed to achieve a required sample size.

Example. Logistic Distribution.

Simulate from the logistic distribution with pdf

$$f(x) = \frac{e^{-x}}{(1+e^{-x})^2}, x \in R.$$

Example. Logistic Distribution.

Simulate from the logistic distribution with pdf

$$f(x) = \frac{e^{-x}}{(1+e^{-x})^2}, x \in R.$$

The logistic distribution is symmetric around 0. We will consider simulation from the truncated at zero logistic distribution,

$$f(x) = 2\frac{e^{-x}}{(1+e^{-x})^2}, x > 0.$$

Then, simulating the symmetric, left part of the density is straightforward.

Example. Logistic Distribution.

Simulate from the logistic distribution with pdf

$$f(x) = \frac{e^{-x}}{(1+e^{-x})^2}, x \in R.$$

The logistic distribution is symmetric around 0. We will consider simulation from the truncated at zero logistic distribution,

$$f(x) = 2\frac{e^{-x}}{(1+e^{-x})^2}, x > 0.$$

Then, simulating the symmetric, left part of the density is straightforward.

We will use the envelope $g(x) = e^{-x}$, $x > 0$. Then

$$f(x) \leq Kg(x) \Rightarrow \frac{2}{(1+e^{-x})^2} \leq K.$$

The best value of K is $K = \lim_{x\to\infty}(1+e^{-x})^2 = 2$.

Example. The Algorithm.

1. Simulate x^* from $g(x) = e^{-x}$, i.e the Exp(1) distribution.
2. Simulate y^* from $U(0, Kg(x^*))$, i.e from $U(0, 2e^{(} - x))$.
3. Accept x^* if $y^* \leq f(x^*)$.
4. Continue.

Equivalently,

Example. The Algorithm.

1. Simulate x^* from $g(x) = e^{-x}$, i.e the Exp(1) distribution.
2. Simulate y^* from $U(0, Kg(x^*))$, i.e from $U(0, 2e^{(} - x))$.
3. Accept x^* if $y^* \leq f(x^*)$.
4. Continue.

Equivalently,

1. Simulate x^* from $g(x) = e^{-x}$.
2. Simulate u from $U(0, 1)$.
3. Accept x^* if $u \leq \frac{f(x^*)}{Kg(x^*)} \Rightarrow u \leq \frac{1}{(1+e^{-x})^2}$.
4. Continue.

Example. The Algorithm.

1. Simulate x^* from $g(x) = e^{-x}$, i.e the Exp(1) distribution.
2. Simulate y^* from $U(0, Kg(x^*))$, i.e from $U(0, 2e^{(} - x))$.
3. Accept x^* if $y^* \leq f(x^*)$.
4. Continue.

Equivalently,

1. Simulate x^* from $g(x) = e^{-x}$.
2. Simulate u from $U(0, 1)$.
3. Accept x^* if $u \leq \frac{f(x^*)}{Kg(x^*)} \Rightarrow u \leq \frac{1}{(1+e^{-x})^2}$.
4. Continue.

Note. If $U \sim U(0, 1)$, then $aU \sim U(0, a)$.
If $Y \sim U(0, a)$, then $U = \frac{Y}{a} \sim (0, 1)$.
$y \leq b \Rightarrow u \leq \frac{b}{a}$. Here, $a = Kg(x^*)$ and $b = f(x^*)$.

Monte Carlo Integration

The reason for simulation can often be evaluating an integral. We have seen that we can use sample means of the form

$$\hat{\theta}_n(f) = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i), \tag{1}$$

where the x_is are a sample from the distribution of X with density f, in order to approximate expectations

$$\theta = \int \phi(x)f(x)dx = E_f(\phi(X)), \tag{2}$$

where $E_f(\phi(X))$ denotes the expectation of $\phi(X)$ with respect to the density f. The notation $\hat{\theta}_n(f)$ reflects that it is an estimate of θ which depends both on the sample size n and the fact that the x_is have been simulated from f.

Monte Carlo Integration

The idea behind Monte Carlo integration is to express an integral we wish to compute as an expectation of a random variable and then simulate samples from that random variable to approximate the integral by the appropriate sample mean.

Monte Carlo Integration

The idea behind Monte Carlo integration is to express an integral we wish to compute as an expectation of a random variable and then simulate samples from that random variable to approximate the integral by the appropriate sample mean.

As an example of this, suppose we wish to calculate $P(X < 1, Y < 1)$ where (X, Y) are bivariate Standard Normal with correlation 0.5. This can be written as

$$\int I_A(x, y) f(x, y) dx dy$$

where f is the bivariate normal density, and I_A is the indicator function on $A = \{(x, y) : \ x < 1, y < 1\}$. Provided we can simulate from the bivariate normal, we can estimate the probability as

$$n^{-1} \sum_{i=1}^{n} I_A(x_i, y_i)$$

which is simply the proportion of simulated points falling in A.

Importance Sampling

Suppose that we want to estimate the expectation of several different functions with respect to a distribution f, but we cannot obtain samples from f by any of the available methods (inverse CDF, rejection sampling, etc). Suppose, however, that instead we can sample from a density g defined on the same space as f, which dominates f, in the sense that $g(x) = 0 \implies f(x) = 0$. For a given function ϕ, we re-write (2) as,

$$\theta = \int \phi(x) \frac{f(x)}{g(x)} g(x) dx = E_g(\phi(X) w(X)), \quad \text{where } w(x) = \frac{f(x)}{g(x)};$$

$w(X)$ is known as the *importance weight* associated to the sampled point (*particle*) X, and g is called the *importance density*.

Importance Sampling

The above expression suggests two different so-called *importance sampling* estimators:

$$\hat{\theta}_n(g) = \frac{1}{n}\left\{\sum_{i=1}^{n} \phi(x_i)w(x_i)\right\},$$

and

$$\hat{\theta}_n^b(g) = \frac{\sum_{i=1}^{n} \phi(x_i)w(x_i)}{\sum_{i=1}^{n} w(x_i)},$$

where the x_is are iid draws from g.

Importance Sampling

The above expression suggests two different so-called *importance sampling* estimators:

$$\hat{\theta}_n(g) = \frac{1}{n}\left\{\sum_{i=1}^{n} \phi(x_i)w(x_i)\right\},$$

and

$$\hat{\theta}_n^b(g) = \frac{\sum_{i=1}^{n} \phi(x_i)w(x_i)}{\sum_{i=1}^{n} w(x_i)},$$

where the x_is are iid draws from g.

Generally $\hat{\theta}_n^b(g)$ is a biased estimator of θ (thus the superscript "b"), but in many cases it might have a smaller mean square error than $\hat{\theta}_n(g)$. However, the main advantage of the biased estimator over the unbiased is that the former does not require knowledge of the normalizing constants of f and g in order to be computed.

Example

Use (naive) Monte Carlo integration and Importance Sampling to estimate the integral

$$I = \int_0^1 [x(1-x^2)(2-x)]^{1/2} dx.$$

Example

Use (naive) Monte Carlo integration and Importance Sampling to estimate the integral

$$I = \int_0^1 [x(1-x^2)(2-x)]^{1/2} dx.$$

Monte Carlo (naive):

$$I = \int_0^1 [x(1-x^2)(2-x)]^{1/2} f(x) dx, \quad \text{where} \quad f(x) = 1.$$

Then, simulate $x_1, \ldots, x_n$ draws from the $U(0,1)$ distribution and use the estimate

$$\hat{I}_n(f) = \frac{1}{n} \sum_{i=1}^{n} [x_i(1-x_i^2)(2-x_i)]^{1/2}.$$

Example

Importance Sampling: Use the $Beta(\frac{3}{2}, \frac{3}{2})$ distribution as importance density $g(x)$. Then

$$\begin{aligned} I &= \int_0^1 [x(1-x)(1+x)(2-x)]^{1/2} \frac{B(\frac{3}{2}, \frac{3}{2})}{x^{\frac{3}{2}-1}(1-x)^{\frac{3}{2}-1}} g(x)dx \\ &= \int_0^1 [(1+x)(2-x)]^{1/2} \frac{B(\frac{3}{2}, \frac{3}{2})}{g}(x)dx \end{aligned}$$

Then, simulate $x_1, \ldots, x_n$ draws from the $Beta(\frac{3}{2}, \frac{3}{2})$ distribution and use the estimate

$$\hat{I}_n(g) = \frac{1}{n} \sum_{i=1}^{n} [(1+x_i)(2-x_i)]^{1/2} B(\frac{3}{2}, \frac{3}{2}).$$