Big Data Analysis
(MA60306)

Bibhas Adhikari

Spring 2022-23, IIT Kharagpur

Lecture 21
March 22, 2023

Sampling methods

Generation of uniformly distributed random variables and methods from transforming it to other distributions

Sampling methods

Generation of uniformly distributed random variables and methods from transforming it to other distributions

$\rightarrow$ digital computer cannot generate random numbers

Sampling methods

Generation of uniformly distributed random variables and methods from transforming it to other distributions

$\rightarrow$ digital computer cannot generate random numbers

$\rightarrow$ this is not a disadvantage if there is some source of pseudorandom numbers, samples of which seem to be randomly drawn from some known distribution

Sampling methods

Generation of uniformly distributed random variables and methods from transforming it to other distributions

$\rightarrow$ digital computer cannot generate random numbers

$\rightarrow$ this is not a disadvantage if there is some source of pseudorandom numbers, samples of which seem to be randomly drawn from some known distribution

$\rightarrow$ there are two issues: randomness and knowledge of the distribution

Sampling methods

Generation of uniformly distributed random variables and methods from transforming it to other distributions

→ digital computer cannot generate random numbers

→ this is not a disadvantage if there is some source of pseudorandom numbers, samples of which seem to be randomly drawn from some known distribution

→ there are two issues: randomness and knowledge of the distribution

Relevance to cryptography - dynamic concept of randomness

→ a process is "random" if the known conditional probability of the next event, given the previous history (or any other information, for that matter) is no different from the known unconditional probability.

Sampling methods

Generation of uniformly distributed random variables and methods from transforming it to other distributions

→ digital computer cannot generate random numbers

→ this is not a disadvantage if there is some source of pseudorandom numbers, samples of which seem to be randomly drawn from some known distribution

→ there are two issues: randomness and knowledge of the distribution

Relevance to cryptography - dynamic concept of randomness

→ a process is "random" if the known conditional probability of the next event, given the previous history (or any other information, for that matter) is no different from the known unconditional probability.

→ This kind of definition leads to the concept of a "one-way function"

Sampling methods

Generation of uniformly distributed random variables and methods from transforming it to other distributions

→ digital computer cannot generate random numbers

→ this is not a disadvantage if there is some source of pseudorandom numbers, samples of which seem to be randomly drawn from some known distribution

→ there are two issues: randomness and knowledge of the distribution

Relevance to cryptography - dynamic concept of randomness

→ a process is "random" if the known conditional probability of the next event, given the previous history (or any other information, for that matter) is no different from the known unconditional probability.

→ This kind of definition leads to the concept of a "one-way function"

→ A one-way function is a function f, such that for any x in its domain, $f(x)$ can be computed in polynomial time, and given $f(x)$, x cannot be computed in polynomial time

Sampling methods

$\rightarrow$ The existence of a one-way function has not been proven

Sampling methods

→ The existence of a one-way function has not been proven

→ In random number generation, the function of interest yields a stream of "unpredictable" numbers, that is,

$$x_i = f(x_{i-1}, x_{i-2}, \ldots, x_{i-k})$$

is easily computable; but x_{i-1}, given $x_i, x_{i-2}, \ldots, x_{i-k}$, is not easily computable

Sampling methods

$\rightarrow$ The existence of a one-way function has not been proven

$\rightarrow$ In random number generation, the function of interest yields a stream of "unpredictable" numbers, that is,

$$x_i = f(x_{i-1}, x_{i-2}, \ldots, x_{i-k})$$

is easily computable; but x_{i-1}, given $x_i, x_{i-2}, ..., x_{i-k}$, is not easily computable

$\rightarrow$ In our context, random numbers simulate realizations of random variables

(pseudo)Random number generation - a mechanism for generating a sequence of random variables $U_1, U_2, \ldots$ with the property that

each U_i is uniformly distributed between 0 and 1

the U_i are mutually independent i.e. the value of U_i should not be predictable from $U_1, \ldots, U_{i-1}$

Sampling methods

$\rightarrow$ The existence of a one-way function has not been proven

$\rightarrow$ In random number generation, the function of interest yields a stream of "unpredictable" numbers, that is,

$$x_i = f(x_{i-1}, x_{i-2}, \ldots, x_{i-k})$$

is easily computable; but x_{i-1}, given $x_i, x_{i-2}, ..., x_{i-k}$, is not easily computable

$\rightarrow$ In our context, random numbers simulate realizations of random variables

(pseudo)Random number generation - a mechanism for generating a sequence of random variables $U_1, U_2, \ldots$ with the property that

each U_i is uniformly distributed between 0 and 1

the U_i are mutually independent i.e. the value of U_i should not be predictable from $U_1, \ldots, U_{i-1}$

Sampling methods

Uniform distribution over the unit interval $(0, 1)$ with the pdf

$$f(x) = \begin{cases} 1, \text{ if } 0 < x < 1 \\ 0, \text{ otherwise} \end{cases}$$

Sampling methods

Uniform distribution over the unit interval $(0, 1)$ with the pdf

$$f(x) = \begin{cases} 1, \text{ if } 0 < x < 1 \\ 0, \text{ otherwise} \end{cases}$$

We denote it as $U(0, 1)$

Sampling methods

Uniform distribution over the unit interval $(0, 1)$ with the pdf

$$f(x) = \begin{cases} 1, \text{ if } 0 < x < 1 \\ 0, \text{ otherwise} \end{cases}$$

We denote it as $U(0, 1)$There are two basic methods:

1. congruential methods
2. feedback shift register methods

Sampling methods

Uniform distribution over the unit interval (0, 1) with the pdf

$$f(x) = \begin{cases} 1, \text{ if } 0 < x < 1 \\ 0, \text{ otherwise} \end{cases}$$

We denote it as $U(0, 1)$There are two basic methods:

1. congruential methods
2. feedback shift register methods

Usually random integers over some fixed range are first generated and then scaled into the interval $(0, 1)$

Sampling methods

Uniform distribution over the unit interval (0, 1) with the pdf

$$f(x) = \begin{cases} 1, \text{ if } 0 < x < 1 \\ 0, \text{ otherwise} \end{cases}$$

We denote it as $U(0, 1)$There are two basic methods:

1. congruential methods
2. feedback shift register methods

Usually random integers over some fixed range are first generated and then scaled into the interval $(0, 1)$

Goal produce a finite sequence of numbers $u_1, \ldots, u_K$ in the unit interval

Uniformity - If the K is large then the fraction of values falling in any subinterval of the unit interval should be approximately the length of the subinterval

Sampling methods

Uniform distribution over the unit interval (0, 1) with the pdf

$$f(x) = \begin{cases} 1, \text{ if } 0 < x < 1 \\ 0, \text{ otherwise} \end{cases}$$

We denote it as $U(0, 1)$There are two basic methods:

1. congruential methods
2. feedback shift register methods

Usually random integers over some fixed range are first generated and then scaled into the interval $(0, 1)$

Goal produce a finite sequence of numbers $u_1, \ldots, u_K$ in the unit interval

Uniformity - If the K is large then the fraction of values falling in any subinterval of the unit interval should be approximately the length of the subinterval

Independence - there should be no discernible pattern among the values i.e. statistical test for independence should not easily reject segments of the sequence $u_1, \ldots, u_K$

Sampling methods

Modular arithmetic Two integers a, b are said to be congruent or equivalent modulo m if $a - b$ is divisible by m, we write it as

$$a \equiv b \bmod m$$

Sampling methods

Modular arithmetic Two integers a, b are said to be congruent or equivalent modulo m if $a - b$ is divisible by m, we write it as

$$a \equiv b \bmod m$$

This is an equivalence relation.

Sampling methods

Modular arithmetic Two integers a, b are said to be congruent or equivalent modulo m if $a - b$ is divisible by m, we write it as

$$a \equiv b \bmod m$$

This is an equivalence relation.

Reduction modulo method For a given b, find a such that $a \equiv b \bmod m$ and $0 \leq a < m$. If a satisfies these two conditions, a is called the residue of b modulo m.

Sampling methods

Modular arithmetic Two integers a, b are said to be congruent or equivalent modulo m if $a - b$ is divisible by m, we write it as

$$a \equiv b \bmod m$$

This is an equivalence relation.

Reduction modulo method For a given b, find a such that $a \equiv b \bmod m$ and $0 \leq a < m$. If a satisfies these two conditions, a is called the residue of b modulo m.
Reduction of b modulo m can be defined as:

$$a = b - \lfloor b/m \rfloor m.$$

Sampling methods

Sampling methods

Linear (multiplicative) congruential generator

$$\begin{aligned} x_{i+1} &= a x_i \bmod m \\ u_{i+1} &= x_{i+1}/m \end{aligned} \tag{1}$$

the integers a and m determine the values generated, given an initial value $1 \leq x_0 \leq m-1$, specified by the user.

Sampling methods

Linear (multiplicative) congruential generator

$$\begin{aligned} x_{i+1} &= a x_i \bmod m \\ u_{i+1} &= x_{i+1}/m \end{aligned} \tag{1}$$

the integers a and m determine the values generated, given an initial value $1 \leq x_0 \leq m-1$, specified by the user.
If a, m are properly chosen then u_i's 'look like' they are randomly and uniformly distributed. The recurrence relation (1) is equivalent to the recurrence

$$u_i \equiv a u_{i-1} \bmod 1 \quad \text{with } 0 < u_i < 1$$

Sampling methods

Note that the operation

$$y \bmod m = y - \left\lfloor \frac{y}{m} \right\rfloor m$$

For example, $7 \bmod 5 = 2$.

Sampling methods

Note that the operation

$$y \bmod m = y - \left\lfloor \frac{y}{m} \right\rfloor m$$

For example, $7 \bmod 5 = 2$. Also note that $0 \leq u_i \leq (m-1)/m$.

Sampling methods

Note that the operation

$$y \bmod m = y - \left\lfloor \frac{y}{m} \right\rfloor m$$

For example, $7 \bmod 5 = 2$. Also note that $0 \leq u_i \leq (m-1)/m$.

Observation

$\rightarrow$ $x_{i+1} = f(x_i), u_{i+1} = g(x_{i+1})$

Sampling methods

Note that the operation

$$y \bmod m = y - \left\lfloor \frac{y}{m} \right\rfloor m$$

For example, $7 \bmod 5 = 2$. Also note that $0 \leq u_i \leq (m-1)/m$.

Observation

$\rightarrow$ $x_{i+1} = f(x_i), u_{i+1} = g(x_{i+1})$

$\rightarrow$ Suppose $a = 6$ and $m = 11$. Then starting from $x_0 = 1$, the linear congruential generator produces

$$1, 6, 3, 7, 9, 10, 5, 8, 4, 2, 1, 6, \ldots$$

Sampling methods

Note that the operation

$$y \bmod m = y - \left\lfloor \frac{y}{m} \right\rfloor m$$

For example, $7 \bmod 5 = 2$. Also note that $0 \leq u_i \leq (m-1)/m$.

Observation

$\rightarrow$ $x_{i+1} = f(x_i), u_{i+1} = g(x_{i+1})$

$\rightarrow$ Suppose $a = 6$ and $m = 11$. Then starting from $x_0 = 1$, the linear congruential generator produces

$$1, 6, 3, 7, 9, 10, 5, 8, 4, 2, 1, 6, \ldots$$

$\rightarrow$ Once a value repeated, the entire sequence repeats

Sampling methods

Note that the operation

$$y \bmod m = y - \left\lfloor \frac{y}{m} \right\rfloor m$$

For example, $7 \bmod 5 = 2$. Also note that $0 \leq u_i \leq (m-1)/m$.

Observation

$\rightarrow$ $x_{i+1} = f(x_i), u_{i+1} = g(x_{i+1})$

$\rightarrow$ Suppose $a = 6$ and $m = 11$. Then starting from $x_0 = 1$, the linear congruential generator produces

$$1, 6, 3, 7, 9, 10, 5, 8, 4, 2, 1, 6, \ldots$$

$\rightarrow$ Once a value repeated, the entire sequence repeats

$\rightarrow$ (Homework) What is your observation for different choices of x_0?

Sampling methods

General considerations

$\rightarrow$ Period length - The generator of the above form eventually repeat itself. The longest possible period for a linear congruent generator with *mod* m is $m-1$, and with full period the gaps between the values u_i are $1/m$. Thus the larger m is more closely the values can approximate a uniform distribution

Sampling methods

General considerations

→ Period length - The generator of the above form eventually repeat itself. The longest possible period for a linear congruent generator with *mod* m is $m-1$, and with full period the gaps between the values u_i are $1/m$. Thus the larger m is more closely the values can approximate a uniform distribution

→ Portability - An algo for random number generation should produce the same sequence of values on all computing platforms

Sampling methods

General considerations

→ Period length - The generator of the above form eventually repeat itself. The longest possible period for a linear congruent generator with *mod* m is $m-1$, and with full period the gaps between the values u_i are $1/m$. Thus the larger m is more closely the values can approximate a uniform distribution

→ Portability - An algo for random number generation should produce the same sequence of values on all computing platforms

→ Randomness - theoretical properties and statistical test

Sampling methods

General considerations

→ Period length - The generator of the above form eventually repeat itself. The longest possible period for a linear congruent generator with *mod* m is $m-1$, and with full period the gaps between the values u_i are $1/m$. Thus the larger m is more closely the values can approximate a uniform distribution

→ Portability - An algo for random number generation should produce the same sequence of values on all computing platforms

→ Randomness - theoretical properties and statistical test

Full period - A linear congruential generator is said to have full period if it produces all $m-1$ distinct values before repeating

Sampling methods

(D. H. Lehmer, 1948) A general linear (mixed) congruential generator:

$$\begin{aligned} x_{i+1} &= (ax_i + c) \bmod m \\ u_{i+1} &= x_{i+1}/m \end{aligned} \tag{2}$$

a, m, c are integers (a is called 'multiplier, c is called the 'increment', and m is called the 'modulus' of the generator)

Sampling methods

(D. H. Lehmer, 1948) A general linear (mixed) congruential generator:

$$\begin{aligned} x_{i+1} &= (ax_i + c) \bmod m \\ u_{i+1} &= x_{i+1}/m \end{aligned} \tag{2}$$

a, m, c are integers (a is called 'multiplier, c is called the 'increment', and m is called the 'modulus' of the generator)

Produced values: When can the generator have full period i.e. the number of distinct values generated from any seed x_0 is $m - 1$?

c and m are relatively prime

Sampling methods

(D. H. Lehmer, 1948) A general linear (mixed) congruential generator:

$$\begin{aligned} x_{i+1} &= (ax_i + c) \bmod m \\ u_{i+1} &= x_{i+1}/m \end{aligned} \tag{2}$$

a, m, c are integers (a is called 'multiplier, c is called the 'increment', and m is called the 'modulus' of the generator)

Produced values: When can the generator have full period i.e. the number of distinct values generated from any seed x_0 is $m - 1$?

- c and m are relatively prime
- every prime number that divides m divides $a - 1$

Sampling methods

(D. H. Lehmer, 1948) A general linear (mixed) congruential generator:

$$\begin{aligned} x_{i+1} &= (ax_i + c) \bmod m \\ u_{i+1} &= x_{i+1}/m \end{aligned} \tag{2}$$

a, m, c are integers (a is called 'multiplier, c is called the 'increment', and m is called the 'modulus' of the generator)

Produced values: When can the generator have full period i.e. the number of distinct values generated from any seed x_0 is $m - 1$?

- c and m are relatively prime
- every prime number that divides m divides $a - 1$
- $a - 1$ is divisible by 4 if m is

Sampling methods

(D. H. Lehmer, 1948) A general linear (mixed) congruential generator:

$$\begin{aligned} x_{i+1} &= (ax_i + c) \bmod m \\ u_{i+1} &= x_{i+1}/m \end{aligned} \tag{2}$$

a, m, c are integers (a is called 'multiplier, c is called the 'increment', and m is called the 'modulus' of the generator)

Produced values: When can the generator have full period i.e. the number of distinct values generated from any seed x_0 is $m - 1$?

- c and m are relatively prime
- every prime number that divides m divides $a - 1$
- $a - 1$ is divisible by 4 if m is
- If m is a power of 2, the generator has full period if c is odd and $a = 4n + 1$ for some integer n

Sampling methods

Question How large should the value of m be?

Sampling methods

Question How large should the value of m be? Literature:

Modulus m	Multiplier a
$2^{31} - 1$	16807
$(= 2147483647)$	
	1226874159
2147483399	40692
	40014

Sampling methods

Question How large should the value of m be? Literature:

Modulus m	Multiplier a
$2^{31} - 1$	16807
$(= 2147483647)$	
	1226874159
2147483399	40692
	40014

Note Currently the numbers used as moduli in production random number generators are often primes in particular, Mersenne primes, which have the form $2^p - 1$. Numbers of this form for $p \leq 31$ are prime except for the three values: $p = 11, 23$, and 29. Most larger values of p do not yield prime (Write a program and check!), however $p = 859433$ does give a prime.

Sampling methods

For random number generator to be useful in most practical application, the period must be of the order of at least 10^9 or so, which means that the 9 modulus in a linear congruential generator must be at least that large.

Sampling methods

For random number generator to be useful in most practical application, the period must be of the order of at least 10^9 or so, which means that the 9 modulus in a linear congruential generator must be at least that large.

Tests of linear congruential generator form any transformation on subsequences of a produced sequence of numbers, determine the distribution of the transformation under the null hypothesis of independent uniformity of the sequence, and perform a goodness-of-fit test of that distribution

Sampling methods

For random number generator to be useful in most practical application, the period must be of the order of at least 10^9 or so, which means that the 9 modulus in a linear congruential generator must be at least that large.

Tests of linear congruential generator form any transformation on subsequences of a produced sequence of numbers, determine the distribution of the transformation under the null hypothesis of independent uniformity of the sequence, and perform a goodness-of-fit test of that distribution

What is the issue - avoid overflow

Sampling methods

For random number generator to be useful in most practical application, the period must be of the order of at least 10^9 or so, which means that the 9 modulus in a linear congruential generator must be at least that large.

Tests of linear congruential generator form any transformation on subsequences of a produced sequence of numbers, determine the distribution of the transformation under the null hypothesis of independent uniformity of the sequence, and perform a goodness-of-fit test of that distribution

What is the issue - avoid overflow

Some other methods

Multiple Recursive Generators

Sampling methods

For random number generator to be useful in most practical application, the period must be of the order of at least 10^9 or so, which means that the 9 modulus in a linear congruential generator must be at least that large.

Tests of linear congruential generator form any transformation on subsequences of a produced sequence of numbers, determine the distribution of the transformation under the null hypothesis of independent uniformity of the sequence, and perform a goodness-of-fit test of that distribution

What is the issue - avoid overflow

Some other methods

- Multiple Recursive Generators
- Lagged Fibonacci

Sampling methods

For random number generator to be useful in most practical application, the period must be of the order of at least 10^9 or so, which means that the 9 modulus in a linear congruential generator must be at least that large.

Tests of linear congruential generator form any transformation on subsequences of a produced sequence of numbers, determine the distribution of the transformation under the null hypothesis of independent uniformity of the sequence, and perform a goodness-of-fit test of that distribution

What is the issue - avoid overflow

Some other methods

- Multiple Recursive Generators
- Lagged Fibonacci
- Inversive Congruential Generators
- Nonlinear Congruential Generators
- Matrix Congruential Generators and many more including Monte Carlo methods

Sampling methods

Sampling from a nonuniform distribution - usually done by applying a transformation to uniform sampler or from a sequence of uniform samplers, other methods use a random walk sequence, a Markov chain

Sampling methods

Sampling from a nonuniform distribution - usually done by applying a transformation to uniform sampler or from a sequence of uniform samplers, other methods use a random walk sequence, a Markov chain

The performance of the algorithms is judged by - in speed, in accuracy, in storage requirements, and in complexity of coding

Sampling methods

Now we assume that there is a good way to generate independent samples of the uniform random variable U on the interval $(0, 1)$

Sampling methods

Now we assume that there is a good way to generate independent samples of the uniform random variable U on the interval $(0, 1)$

Sampling finite and discrete rvs

Bernoulli random variable: If

$$X = \begin{cases} 1 \text{ if } U \leq p \\ 0 \text{ otherwise} \end{cases}$$

then $X \sim Ber(p)$ since 1 will be sampled with probability p, and 0 will be sampled with probability $1 - p$.

Sampling methods

Inverse transform technique: Let $X = \{x_1, \ldots, x_n\}$ be a random variable with probability distribution p, and where $x_1 \leq \ldots \leq x_n$. Then define

$$q_i = P(X \leq x_i) = \sum_{j=1}^{i} p(x_j)$$

Sampling methods

Inverse transform technique: Let $X = \{x_1, \ldots, x_n\}$ be a random variable with probability distribution p, and where $x_1 \leq \ldots \leq x_n$. Then define

$$q_i = P(X \leq x_i) = \sum_{j=1}^{i} p(x_j)$$

Then the following is the sampling formula for X as follows:

$$X = \begin{cases} x_1 \text{ if } U < q_1 \\ x_2 \text{ if } q_1 \leq U < q_2 \\ \vdots \\ x_{n-1} \text{ if } q_{n-2} \leq U < q_{n-1} \\ x_n \textit{ otherwise} \end{cases}$$

Note that $X = x_i$ in the event that $q_{i-1} \leq U < q_i$, which has probability $p = q_i - q_{i-1} = p(x_i)$.

Sampling methods

Geometric rv For the uniform distribution U,

$$X = \left\lfloor \frac{\ln U}{\ln q} \right\rfloor + 1 \sim G(p)$$

with parameter $p = 1 - q$

Sampling methods

Geometric rv For the uniform distribution U,

$$X = \left\lfloor \frac{\ln U}{\ln q} \right\rfloor + 1 \sim G(p)$$

with parameter $p = 1 - q$

Proof First sample $U(0, 1)$ and then return k, where

$$\sum_{n=1}^{k-1} (1-p)^{n-1} p \leq U < \sum_{n=1}^{k} (1-p)^{n-1} p,$$

which implies

$$1 - (1-p)^{k-1} \leq U < 1 - (1-p)^k \Rightarrow (1-p)^k < 1 - U \leq (1-p)^{k-1}$$

using

$$\sum_{n=1}^{k} ar^{n-1} = a\frac{r^k - 1}{r - 1}$$

Sampling methods

Taking log both side and dividing by the negative number $\ln(1-p)$ then

$$k-1 \leq \frac{\ln(1-U)}{\ln(1-p)} < k \Rightarrow k = \left\lfloor \frac{\ln(1-U)}{\ln(1-p)} \right\rfloor + 1.$$

Sampling methods

Taking log both side and dividing by the negative number $\ln(1-p)$ then

$$k - 1 \leq \frac{\ln(1-U)}{\ln(1-p)} < k \Rightarrow k = \left\lfloor \frac{\ln(1-U)}{\ln(1-p)} \right\rfloor + 1.$$

Finally, setting $q = 1 - p$, and noting that $1 - U$ is also uniformly distributed over $[0, 1,]$ we have

$$k = \left\lfloor \frac{\ln U}{\ln q} \right\rfloor + 1$$

Sampling methods

Continuous rv - Inverse Transform Method Let X be a continuous random variable with cdf $F(x)$ which possesses an inverse F^{-1}. Let $Y = F^{-1}(U)$, then Y has the same distribution as X.

Sampling methods

Continuous rv - Inverse Transform Method Let X be a continuous random variable with cdf $F(x)$ which possesses an inverse F^{-1}. Let $Y = F^{-1}(U)$, then Y has the same distribution as X.

Proof It is sufficient to show that Y has the same cdf as X. Let F and F_Y denote the cdfs of X and Y respectively. Then

$$\begin{aligned} F_Y(x) &= P(Y \leq x) = P(F^{-1}(U) \leq x) \\ &= P(F(F^{-1}(U) \leq F(x))) \\ &= P(U \leq F(x)) = F(x) \end{aligned}$$

where the third-to-last equality follows from the fact that F is strictly increasing.

Sampling methods

Sampling from standard distributions

Uniform: $X \sim U(a, b) : X = a + U(b - a)$

Sampling methods

Sampling from standard distributions

Uniform: $X \sim U(a, b) : X = a + U(b - a)$

Exponential: $X \sim Exp(\lambda) : X = -\ln(U)/\lambda$

Weibull: $X \sim W(\alpha, \beta, \nu) : X = \nu + \alpha[-\ln(U)]^{1/\beta}$

Cauchy: $X \sim C(\mu, \sigma^2) : X = \mu + \sigma \tan \pi(U - \frac{1}{2})$

Sampling methods

Sampling from standard distributions

Uniform: $X \sim U(a, b) : X = a + U(b - a)$

Exponential: $X \sim Exp(\lambda) : X = -\ln(U)/\lambda$

Weibull: $X \sim W(\alpha, \beta, \nu) : X = \nu + \alpha[-\ln(U)]^{1/\beta}$

Cauchy: $X \sim C(\mu, \sigma^2) : X = \mu + \sigma \tan \pi(U - \frac{1}{2})$

Empirical CDF Suppose $x_1 \leq x_2 \leq \ldots \leq x_n$ is a collection of n data points, with $x_i \in [a, \infty]$ for some $a \in \mathbb{R}$. Then the empirical CDF $F(x)$ with linear interpolation is defined as follows:

1. Given $x \in \{x_1, \ldots, x_n\}$, let i be the largest index for which $x = x_i$ then $F(x) = \frac{i}{n}$

Sampling methods

Sampling from standard distributions

Uniform: $X \sim U(a, b) : X = a + U(b - a)$

Exponential: $X \sim Exp(\lambda) : X = -\ln(U)/\lambda$

Weibull: $X \sim W(\alpha, \beta, \nu) : X = \nu + \alpha[-\ln(U)]^{1/\beta}$

Cauchy: $X \sim C(\mu, \sigma^2) : X = \mu + \sigma \tan \pi(U - \frac{1}{2})$

Empirical CDF Suppose $x_1 \leq x_2 \leq \ldots \leq x_n$ is a collection of n data points, with $x_i \in [a, \infty]$ for some $a \in \mathbb{R}$. Then the empirical CDF $F(x)$ with linear interpolation is defined as follows:

1. Given $x \in \{x_1, \ldots, x_n\}$, let i be the largest index for which $x = x_i$ then $F(x) = \frac{i}{n}$
2. $F(x) = 0$ for all $x \leq a$

Sampling methods

Sampling from standard distributions

Uniform: $X \sim U(a, b) : X = a + U(b - a)$

Exponential: $X \sim Exp(\lambda) : X = -\ln(U)/\lambda$

Weibull: $X \sim W(\alpha, \beta, \nu) : X = \nu + \alpha[-\ln(U)]^{1/\beta}$

Cauchy: $X \sim C(\mu, \sigma^2) : X = \mu + \sigma \tan \pi(U - \frac{1}{2})$

Empirical CDF Suppose $x_1 \leq x_2 \leq \ldots \leq x_n$ is a collection of n data points, with $x_i \in [a, \infty]$ for some $a \in \mathbb{R}$. Then the empirical CDF $F(x)$ with linear interpolation is defined as follows:

1. Given $x \in \{x_1, \ldots, x_n\}$, let i be the largest index for which $x = x_i$ then $F(x) = \frac{i}{n}$
2. $F(x) = 0$ for all $x \leq a$
3. $F(x) = 1$ for all $x \geq x_n$

Sampling methods

Sampling from standard distributions

Uniform: $X \sim U(a, b) : X = a + U(b - a)$

Exponential: $X \sim Exp(\lambda) : X = -\ln(U)/\lambda$

Weibull: $X \sim W(\alpha, \beta, \nu) : X = \nu + \alpha[-\ln(U)]^{1/\beta}$

Cauchy: $X \sim C(\mu, \sigma^2) : X = \mu + \sigma \tan \pi(U - \frac{1}{2})$

Empirical CDF Suppose $x_1 \leq x_2 \leq \ldots \leq x_n$ is a collection of n data points, with $x_i \in [a, \infty]$ for some $a \in \mathbb{R}$. Then the empirical CDF $F(x)$ with linear interpolation is defined as follows:

1. Given $x \in \{x_1, \ldots, x_n\}$, let i be the largest index for which $x = x_i$ then $F(x) = \frac{i}{n}$
2. $F(x) = 0$ for all $x \leq a$
3. $F(x) = 1$ for all $x \geq x_n$
4. if $x \in (a, x_1)$ then $F(x) = \frac{F(x_1)}{x_1 - a}(x - a)$

Sampling methods

Sampling from standard distributions

Uniform: $X \sim U(a, b) : X = a + U(b - a)$

Exponential: $X \sim Exp(\lambda) : X = -\ln(U)/\lambda$

Weibull: $X \sim W(\alpha, \beta, \nu) : X = \nu + \alpha[-\ln(U)]^{1/\beta}$

Cauchy: $X \sim C(\mu, \sigma^2) : X = \mu + \sigma \tan \pi(U - \frac{1}{2})$

Empirical CDF Suppose $x_1 \leq x_2 \leq \ldots \leq x_n$ is a collection of n data points, with $x_i \in [a, \infty]$ for some $a \in \mathbb{R}$. Then the empirical CDF $F(x)$ with linear interpolation is defined as follows:

1. Given $x \in \{x_1, \ldots, x_n\}$, let i be the largest index for which $x = x_i$ then $F(x) = \frac{i}{n}$
2. $F(x) = 0$ for all $x \leq a$
3. $F(x) = 1$ for all $x \geq x_n$
4. if $x \in (a, x_1)$ then $F(x) = \frac{F(x_1)}{x_1 - a}(x - a)$
5. if $x \in (x_i, x_{i+1})$ then $F(x) = F(x_i) + \frac{(x - x_i)[F(x_{i+1}) - F(x_i)]}{x_{i+1} - x_i}$

Sampling methods

Sampling from empirical CDFs Procedure for sampling a value from an empirical CDF $F(x)$

1. Sample from U
2. if $U = 0$ return a
3. else if $U = F(x_i)$ for some $1 \leq i \leq n$ then return x_i
4. else if $U < F(x_1)$ then return

$$a + (x_1 - a)\frac{U}{F(x_1)}$$

5. else if $F(x_i) < U < F(x_{i+1})$ then return

$$x_i + (x_{i+1} - x_i)\frac{U - F(x_i)}{F(x_{i+1}) - F(x_i)}$$