University of Southern California

EE511 Simulation Methods for Stochastic Systems

Project #4 – Integrals and Intervals

BY

Mohan Krishna Thota

USC ID: 6683486728

mthota@usc.edu

Description:

Monte Carlo simulations are used to model the probability of different outcomes in a process that cannot easily be predicted due to the intervention of random variables. Monte Carlo methods are mainly used in three distinct problem classes:

- Optimization.
- Numerical integration.
- generating draws from a probability distribution.

Procedure:

- A set of random numbers are generated using rand() function which are uniformly distributed between 0 and 1.
- At each random number function f(x) is evaluated.
- Then Monte-Carlo estimate of integration Is the function average.
- Rand() function is used to generate random numbers between 0 and 1.
- Substitution method is used such that Integral limits lie in the range of 0 to 1.
- Now the generated random numbers are substituted in the function and the obtained result is stored in 'Result vector'.
- Now Monte-Carlo simulation is given by dividing the sum of all the values in Result vector by number of samples.
- By using MATLAB inbuilt int() and syms() functions, theoretical values are calculated.
- Double precision is used to get the numerical value as the obtained expression is exponential.

1(A). CODE:

```
clc;
clear;
close all;
Number_of_samples=10000;
for j=1:Number_of_samples
    k(j )=rand();
end

for n=1:Number_of_samples
    Result(n)=4*exp(2-12*k(n)+16*(k(n))^2);
end

sum_values=sum(Result);
Monte_estimation=sum_values/Number_of_samples;

syms x;
func=exp(x+x^2);
theoretical=double(int(func,x,-2,2));

disp('Monte Carlo estimated value is:');
disp(Monte_estimation);
disp('Theoretical value is:');
disp(theoretical);
```

OUTPUT:

```
N=10000
Monte Carlo estimated value is:
   94.7398
Theoretical value is:
   93.1628
N=1000
Monte Carlo estimated value is:
   92.2652
Theoretical value is:
   93.1628
N=100
The Monte Carlo estimate is:
   92.0657
The theoretical value of the integral is:
   93.1628
```

1(B). CODE:

CODE:

```
clc;
clear;
Number_of_samples=100;
for n=1:Number_of_samples
    k(n)=rand();
end

for n=1:Number_of_samples
    values(n)=(2*exp(-1-1/(k(n))^2+2/k(n)))/(k(n))^2;
end

sum_values=sum(values);
Monte_estimation=sum_values/Number_of_samples;

syms x;
func=(2*exp(-((1/x)-1)^2))/x^2;
theoretical=double(int(func,x,0,1));

disp('Monte Carlo estimated value is:');
disp(Monte_estimation);
disp('Theoretical value is:');
disp(theoretical);
```

OUTPUT:

```
N=100
Monte Carlo estimated value is:
    1.7711
Theoretical value is:
    1.7725

N=1000
Monte Carlo estimated value is:
    1.7691
Theoretical value is:
    1.7725

N=10000
Monte Carlo estimated value is:
    1.7937

Theoretical value is:
    1.7725
```

1(C)

CODE:

```
clc;
clear;
Number_of_samples=100;
for n=1:Number_of_samples
    k(n)=rand();
end

for n=1:Number_of_samples
    values(n)=exp(-4*(k(n))^2);
end

sum_values=sum(values);
Monte_estimation=sum_values/Number_of_samples;

syms x y;

func=exp(-(x+y)^2);
theoretical=double(int(int(func,x,0,1),y,0,1));

disp('Monte Carlo estimated value is:');
disp(Monte_estimation);
disp('Theoretical value is:');
disp(theoretical);
```

OUTPUT:

```
N=100
Monte Carlo estimated value is:
    0.4082
Theoretical value is:
    0.4118

N=1000
Monte Carlo estimated value is:
    0.4333
Theoretical value is:
    0.4218

N=10000
Monte Carlo estimated value is:
    0.4377
Theoretical value is:
    0.4217
```

ANALYSIS:

- From the above results we could see that both the theoretical and Monte Carlo estimates are similar.
- Monte Carlo estimates are done by averaging the samples. Samples in the range of 100,1000 and 10000 are used.
- Therefore, we could say that Monte Carlo gives a satisfactory value when compared with theoretical values.

Problem 2:

Description:

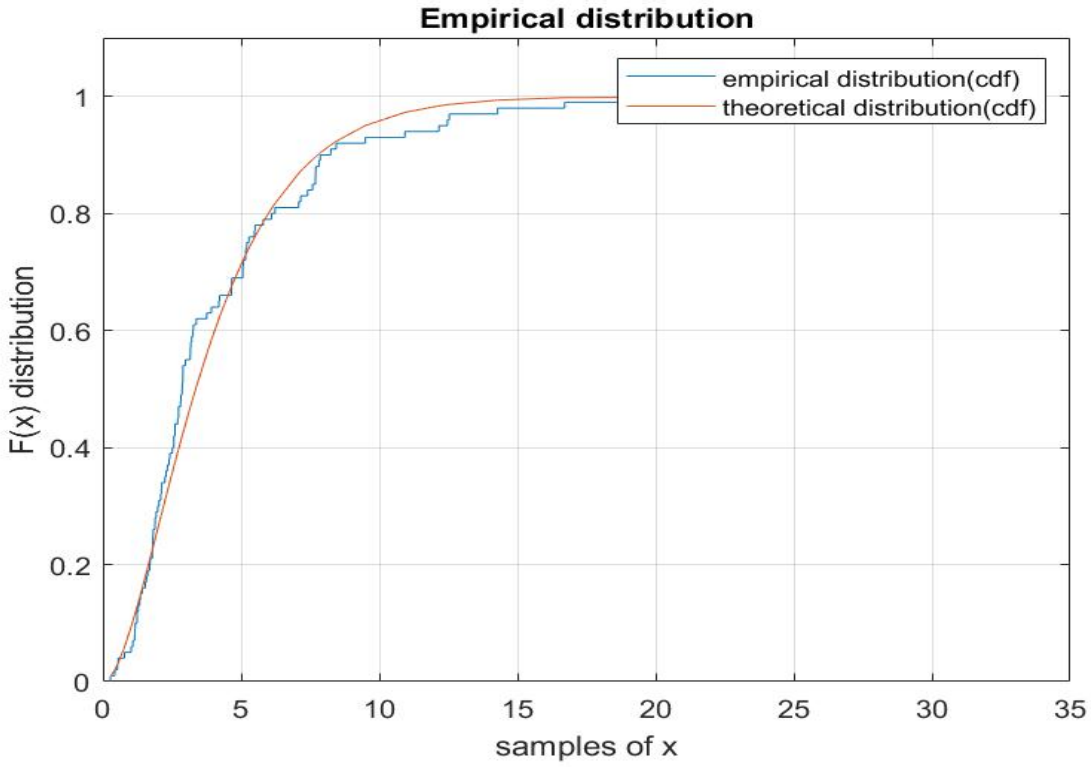
- According to Glivenko-Cantelli theorem, Fn(x) empirical distribution function converges to cumulative distribution function f(x) with probability 1.
- We could see that with the increase in number of samples the estimation is getting better i.e., getting closer to the actual values.
- Empirical distribution function is a step function that jumps up by 1/n for each sample.
- Chi-Squared random Variable Distributions have been generated with degree of freedom = 4.
- The empirical and theoretical distributions of the chi square distributions are overlayed with the help of cdfplot() and chi2cdf() functions.
- As given in the question, firstly 100 elements and later they are increased to 1000.
- Ecdf() function is used which returns the empirical cumulative distribution function evaluated at the points in given distribution.
- Lower bound is given by the difference between of ecdf and theoretical distribution.
- Matlab's prctile() function is used to calculate percentiles.
- As asked in the question 25^{th} ,50^{th} and 90^{th} percentile values are calculated and the obtained values are compared with theoretical.

CODE:

```
Z=1:4;
Number_of_samples=100;
sample_X=zeros(Number_of_samples,1);

for n=1:Number_of_samples
    for k=1:4
        Z(k)=randn();
        sample_X(n) = sample_X(n) + power(Z(k),2);
    end
end

cdfplot(sample_X);
hold on;
grid on;

sample_X=sort(sample_X);

theor_values=chi2cdf(sample_X,4);
F10x=ecdf(sample_X);
plot(sample_X,theor_values);
hold off;

low_bound_est=1:Number_of_samples;
for n=1:Number_of_samples
    low_bound_est(n)=abs(F10x(n)-theor_values(n));
end
lower_bound=max(low_bound_est)
disp('Estimation of the lower bound');
disp(lower_bound);

disp('25th percentile using empirical distribution is');
disp(prctile(F10x,25));
disp('Theoretical Value of 25th percentile is');
disp(prctile(theor_values,25));

disp('50th percentile using empirical distribution is');
disp(prctile(F10x,50));
disp('Theoretical Value of 50th percentile is');
disp(prctile(theor_values,50));

disp('90th percentile using empirical distribution is');
disp(prctile(F10x,90));
disp('Theoretical Value of 90th percentile is');
disp(prctile(theor_values,90));

legend('empirical distribution(cdf)','theoretical distribution(cdf)');
ylim([0 1.1]);
xlabel('samples of x');
ylabel('F(x) distribution');
title('Empirical distribution');
```

OUTPUT:

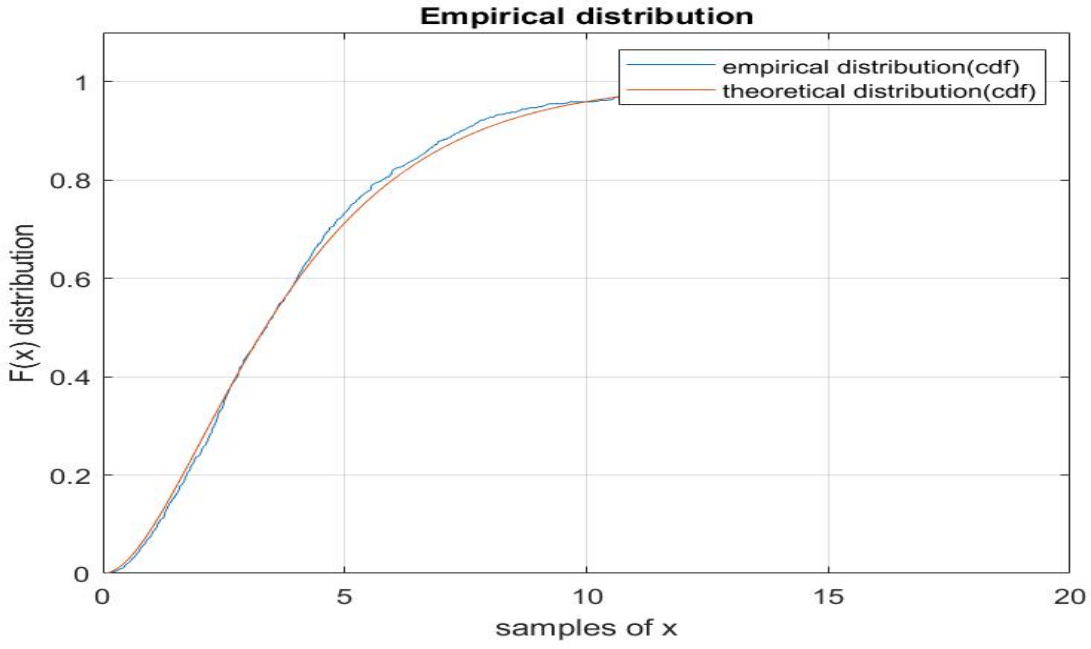
```
>> problem2

lower_bound =
    0.1134

Estimation of the lower bound
    0.1134

25th percentile using empirical distribution is
    0.2475

Theoretical Value of 25th percentile is
    0.2304

50th percentile using empirical distribution is
    0.5000

Theoretical Value of 50th percentile is
    0.4196

90th percentile using empirical distribution is
    0.9040

Theoretical Value of 90th percentile is
    0.9101
```



Number of Samples=1000

```
lower_bound =
    0.0259

Estimation of the lower bound
    0.0259

25th percentile using empirical distribution is
    0.2498

Theoretical Value of 25th percentile is
    0.2442

50th percentile using empirical distribution is
    0.5000

Theoretical Value of 50th percentile is
    0.4746

90th percentile using empirical distribution is
    0.9004

Theoretical Value of 90th percentile is
    0.8897
```

OUTPUT:



ANALYSIS:

- From the above plots we could see that the difference between empirical distribution and theoretical distribution decreases considerably with the increase in the number of samples.
- Therefore, the above result verifies the Glivenko-Cantelli theorem, Empirical function is given by

$$Fn\,(x) \;=\; 1\,n\,\sum I(-\infty, x](Xi)$$

- It means that the empirical function Fn(x) converges to the cumulative distribution function f(x) with the increase in number of samples.

Problem 3:

A geyser is a hot spring characterized by an intermittent discharge of water and steam. Old Faithful is a famous cone geyser in Yellowstone National Park, Wyoming. It has a predictable geothermal discharge and since 2000 it has erupted every 44 to 125 minutes. Refer to the addendum data file that contains waiting times and the durations for 272 eruptions. Compute a 95% statistical confidence interval for the waiting time using data from only the first 15 eruptions. Compare this to a 95% bootstrap confidence interval using the same 15 data samples. Repeat these calculations using all the data samples. Comment on the relative width of the confidence intervals when using only 15 samples vs using all sample

Solution:

- As given in the problem, a file is given which gives waiting numbers and durations of 272 eruptions.
- Matlab's fileread() function is used to read data file and values are scanned to a variable 'file_data'. Prctile() function is used to calculate Upper and lower bound.
- Standard error is calculated using std() and sqrt() functions.
- Tinv() function is used to calculate T-score value. Therefore, 95% statistical confidence is found.
- Bootstrap() function is used to calculate bootstrap confidence interval.

CODE:

```
clc;
clear all;

file = fileread( 'faithful.dat.txt' ) ;
file_data = textscan( file, '%f %f %f%*[^\n]', ...
            'HeaderLines', 3) ;
a = file_data{1};
b = file_data{2};
waiting_272 = file_data{3};
waiting_15 = (waiting_272(1:15));

%statistical confidence interval
mean1=mean(waiting_272);
Std_error = std(waiting_272)/sqrt(length(waiting_272)); %
Standard error calculation
score = tinv([0.025 0.975],length(waiting_272)-1); % T-score
calculation
Conf_Interval = mean1+ score*Std_error; % Confidence Interval
disp('Statistical confidence interval for 272 values');
disp(Conf_Interval);

mean2= mean(waiting_15);
Std_error = std(waiting_15)/sqrt(length(waiting_15));% Standard
error calculation
score = tinv([0.025 0.975],length(waiting_15)-1); % T-score
calculation
Conf_Interval = mean2+ score*Std_error; % Confidence Interval
disp('Statistical confidence interval for all 15 values');
disp(Conf_Interval);

% Bootstrap confidence interval
y = bootstrp(15, @mean, waiting_15);
Sorted = sort(y);
disp('Bootstrap confidence interval for 15 values');
low_conf=prctile(Sorted,2.5);
disp(low_conf);
high_conf=prctile(Sorted,97.5);
disp(high_conf);

y = bootstrp(272, @mean, waiting_272);
Sorted = sort(y);
disp('Bootstrap confidence interval for all 272 values');
low_conf=prctile(Sorted,2.5);
disp(low_conf);
high_conf=prctile(Sorted,97.5);
disp(high_conf);
```

OUTPUT:

```
Statistical confidence interval for all 15 values
   62.5571
    79.3096

Statistical confidence interval for 272 values
   69.2742
   72.5199

Bootstrap confidence interval for 15 values
   63.8667
   77.6000

Bootstrap confidence interval for all 272 values
   69.1805
   72.4680
```

ANALYSIS:

- From the above results, we could say that the confidence interval reduces with the increase in number of samples.
- The relative width is higher, when only 15 samples are taken and corresponding statistical and bootstrap confidence intervals are calculated.
- When all the samples are taken and corresponding bootstrap and statistical confidence intervals are calculated, the width is almost similar.