Project 1: Probability and random numbers generator

EE 511 – Section Thursday 9 am

Name: Junquan Yu

Student ID #: 3372029142

1. Problem Statement

(1) Let ~ U(0,1), evaluate the mean, μ, and variance, σ_X^2.

(2) Generate a sequence of N=100 random numbers between [0,1] and compute the sample mean $m = \frac{\sum_{i=1}^{N} X_i}{N}$ and sample variance $s^2 = \frac{\sum_{i=1}^{N}(X_i - m)^2}{N-1}$ and compare to μ and σ^2. Also, estimate the (sample) variance of the sample mean. Repeat for N=10000.

(3) The Central Limit Theorem says that $m = \frac{\sum_{i=1}^{n} X_i}{n} \rightarrow N(\mu, \sigma^2/n)$. Repeat the experiment in (2) (for N=100) 50 times to generate a set of sample means $\{m_j$, j=1..50$\}$. Do they appear to be approximately normally distributed values with mean μ and variance σ^2/n ?

(4) We want to check whether there is any dependency between X_i and X_{i+1}. Generate a sequence of N+1 random numbers that are ~U (0,1) for N=1,000.

Compute:

$$Z = [\frac{\sum_{i=1}^{N} X_i X_{i+1}}{N}] - [\frac{\sum_{i=1}^{N} X_i}{N}][\frac{\sum_{j=2}^{N+1} X_j}{N}]$$

Comment on what you expect and what you find.

2. Theoretical Exploration or Analysis

For the problem (1):

As is provided in the problem (1), the random variable X_i obeys the uniform distribution U(0,1), so we know that the probability density function of random variable X_i is as follows:

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

The mean of X_i is as follows:

$$\mu = E(X_i) = \int_{-\infty}^{\infty} f(x)dx = \int_0^1 x dx = \frac{1}{2}$$

The variance of X_i is as follows:

$$\sigma_X^2 = D(X_i) = E(X_i^2) - E(X_i)^2 = \int_0^1 x^2 dx - (\frac{1}{2})^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

For the problem (2):

Assuming that random variables $X_1, \dots, X_n$ are independent of each other and obeys the same distribution, and they have mean $E(X_k) = \mu$ and variance $D(X_k) = \sigma^2$ (k=1,2,…), the mean and variance for the sum of these random variables $\sum_{k=1}^{n} X_k$ are as follows:

$$E(\sum_{k=1}^{n} X_k) = n\mu$$

$$D(\sum_{k=1}^{n} X_k) = n\sigma^2$$

The sum of these random variables can be normalized as follows:

$$Z_n = \frac{\sum_{k=1}^{n} X_k - E(\sum_{k=1}^{n} X_k)}{\sqrt{D(\sum_{k=1}^{n} X_k)}} = \frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\sigma}$$

Then the distribution function of Z_n is:

$$\lim_{n\to\infty} F(x) = \lim_{n\to\infty} P(Z_n \le x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

This means that normalized variable Z_n obeys standardized normal distribution approximately when n is sufficiently large, that is:

$$Z_n = \frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\sigma} \sim N(0,1) \qquad \text{(approximately)}$$

And

$$\overline{X} = \frac{\sum_{k=1}^{n} X_k}{n} \sim N(\mu, \sigma^2 / n) \qquad \text{(approximately)}$$

That is to say, arithmetic mean $\overline{X}$ of random variables $X_1,\ldots,X_n$ approximately obeys the normal distribution whose mean is μ and variance is σ^2 / n when n is sufficiently large.

For the problem (3):

From the mathematical derivation, I know that $(\mu - 2\sigma, \mu + 2\sigma)$ is the 95% confidence interval of the normal distribution $N(\mu, \sigma^2 / n)$. I used MATLAB to draw the graph of this normal distribution with mean $\mu = 1/2$ and variance $\sigma^2 / n = 1/1200$. I also drew two reference lines $x_1 = \mu - 2\sigma = 0.471$ and $x_2 = \mu + 2\sigma = 0.529$ to mark the 95% confidence interval of this normal distribution. Then i repeated the experiment in (2) (for N=100) 50 times to generate a set of sample means and plotted them on the x-axis. If most of mean values fall between two reference lines, these sample means appear to be approximately normally distributed with mean $\mu = 1/2$ and variance $\sigma^2 / n = 1/1200$.

For the problem (4):

The dependence of X_i and X_{i+1} can be measure by covariance:

$$COV(X_i, X_{i+1}) = E(X_i X_{i+1}) \text{-} E(X_i)E(X_{i+1})$$

If $COV(X_i, X_{i+1})$=0, X_i and X_{i+1} are independent, or they are dependent.

Simulation Methodology

For the problem (1):

I use the built-in function unifstat( ) to directly return the mean and variance for the continuous uniform distribution U(0,1) with parameters of mu and sig1.

For the problem (2):

I used the built-in function rand( ) to return a 10*10 matrix filled with random numbers on the interval of (0,1). Then I computed the sample mean and sample variance of these N=100 random numbers with the function mean( ) and var( ). In order to estimate the variance of sample means, I inserted the “for” loop construction. Every time through the “for” expression, a matrix filled with N=10000 random numbers in the interval of (0,1) would be generated and the mean of these numbers would be computed. After 100000 cycles with loop structure, I can get a set of 100000 sample means in total. Finally, I estimated and computed the (sample) variance of these 100000 sample means with the function var( ).

For the problem (3):

I created a set of 50 sample means totally in the same way I used in the problem (2) with loop structure. Then I used the built-in function normplot( ) to draw the normal probability plot of these sample means in the figure (1). If a series of points appear along the reference line, it can be roughly estimated that these sample means obey the normal distribution. In order to further test whether these sample means are approximately normally distributed with the mean $\mu = \frac{1}{2}$ and the variance $\frac{\sigma^2}{N} = \frac{1}{1200}$, I used the built-in function normpdf( ) to draw the graph of the probability density function of the normal distribution with the mean $\mu = \frac{1}{2}$ and the variance $\frac{\sigma^2}{N} = \frac{1}{1200}$ in the figure (2). In addition, I also add two reference lines $x3 = \mu + \frac{2\sigma}{\sqrt{N}} = mu + 2*sig2 = 0.558$ and $x4 = \mu - \frac{2\sigma}{\sqrt{N}} = mu - 2*sig2 = 0.442$ on the figure (2) which are the upper and lower bound of 95% confidence interval of the normal distribution with the mean $\mu = \frac{1}{2}$ and the variance $\frac{\sigma^2}{N} = \frac{1}{1200}$. Finally, I plotted the 50 sample means on the x-axis in the form of a series of discrete points. If most of points fall between two reference lines on the x-axis, these sample means appear to be approximately normally distributed with the mean $\mu = \frac{1}{2}$ and the variance $\frac{\sigma^2}{N} = \frac{1}{1200}$.

For the problem (4):

I generated a sequence of N+1=1001 random numbers $X_1, X_2 \dots X_n$ that are $\sim U(0,1)$ with the built-in function unifrnd( ). Then I used the function mean( ) to compute three arithmetic mean values which are $S = \frac{\sum_{i=1}^{1000} X_i}{1000}$, $T = \frac{\sum_{j=2}^{1001} X_j}{1000}$ and $Q = \frac{\sum_{i=1}^{1000} X_i X_{i+1}}{1000}$. In the end, I can get the value of Z from the equation $Z = Q - S*T$.

3. Experiments and Results

For the problem (1):

Because random variable X obeys the uniform distribution U(0,1), I chose the A=0, B=1 as the parameters for the function unifstat(A,B). After running the program, it would return mu=0.5 and sig1=0.833, which are the mean and the variance of random variable X. They match well with the theoretical values from the mathematical derivation which I discussed in the Theoretical Explanation or Analysis Section.

```
>> Project1

mu =

    0.5000


sig1 =

    0.0833
```

For the problem (2):

The problem requires me to generate a sequence of N =100 random numbers between [0,1] and compute their sample mean and variance, and repeat this process for N=10000, so I chose n=10 and n=100 as the parameter for the function rand(n). then I used two arrays, which are currentdata1 and currentdata2, to accommodate 100 and 10000 random numbers respectively. After that, using these two arrays as the parameters, I computed the mean and variance of these 100 and 10000 random numbers with the function mean( ) and var( ). The mean and variance for 100 random numbers are currentmean1=0.5280 and currentvar1=0.0882 whereas the mean and variance for 10000 random numbers are currentmean2=0.4991 and currentvar2=0.0829. They all close to the theoretical values which are mean=0.5 and variance=0.083. It is worth noting, however, that the larger N is, the closer results of stimulation are to the theoretical values.

```
currentmean1 =

    0.5280


currentvar1 =

    0.0882


currentmean2 =

    0.4991


currentvar2 =

    0.0829
```

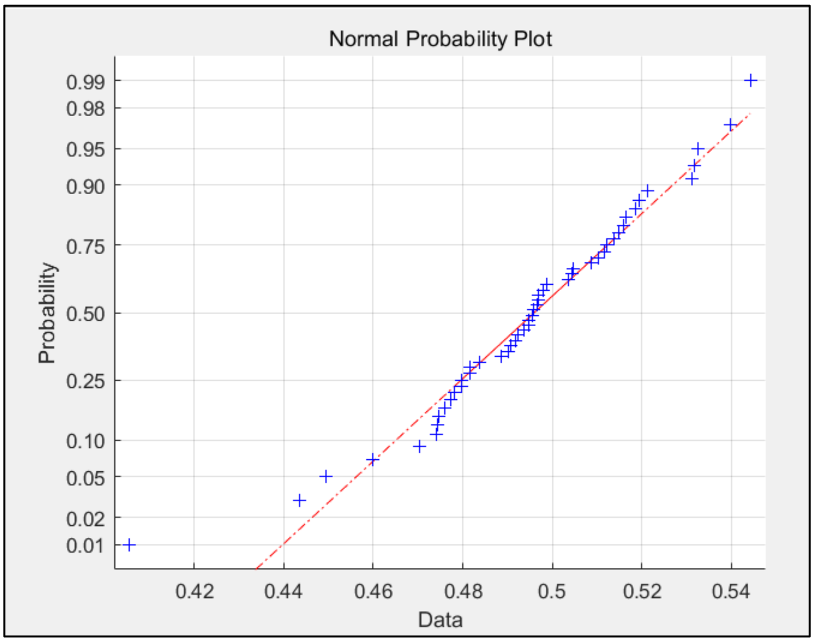
In order to precisely estimate the (sample variance) of the sample mean, I also repeat this process for 100000 times to generate 100000 sample means with the loop structure, each of which is the arithmetic mean value of N=10000 random numbers. Then I computed the variance of these 100000 sample means and assigned it to the parameter $\mathrm{currentva\gamma 3} = 8.365 \times 10^{-6}$. Based on this fact, I can estimate that the variance of sample means of random variable X normally distributed with mean μ and variance σ^2 is $\frac{\sigma^2}{N}$ (In this case, $\mu = \frac{1}{2}$, $\sigma^2 = \frac{1}{12}$ and N=10000). This agrees with the result of theoretical analysis in Theoretical Explanation or Analysis Section.

```
currentvar3 =

   8.3648e-06
```

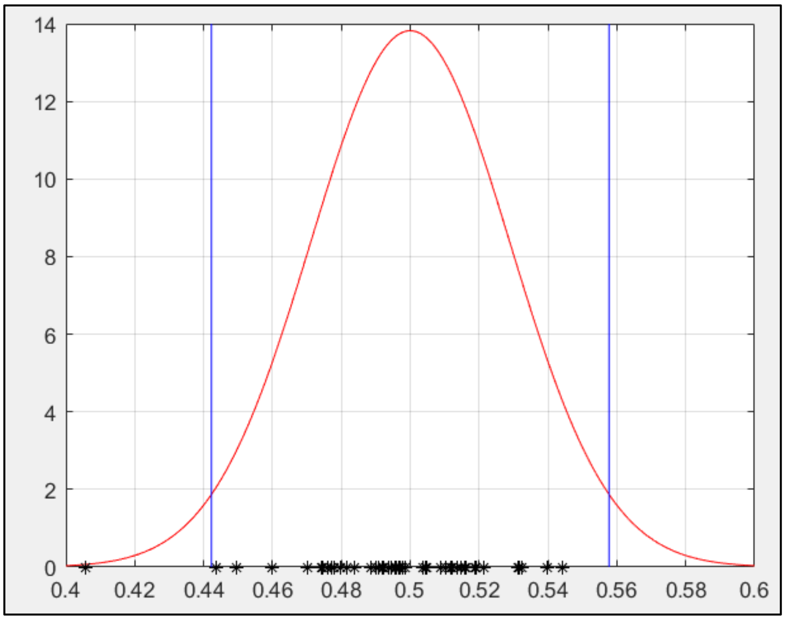
For the problem (3):

I created a set of 50 sample means totally in the same way I used in the problem (2) with loop structure and assigned them to the array currentmean4. Using this whole array as the parameter, I drew the normal probability plot of these 50 sample means with the function normplot( ) in the figure (1).



From this graph, I can see that these blue discrete points are all approximately along the red reference line, which means that these sample means obey the normal distribution approximately. However, I need further examination to decide whether these sample means are normally distributed with specific parameters, like mean $\mu = \frac{1}{2}$ and $\frac{\sigma^2}{N} = \frac{1}{1200}$. For the further examination, I used the function normpdf(x,mu,sigma) to draw the graph of normal distribution with parameters mean

$\text{mu} = \frac{1}{2}$ and variance $\text{sigma} = \sqrt{\frac{1}{1200}} = 0.0289$. In addition, I also add two reference lines $x3 = \mu + \frac{2\sigma}{\sqrt{N}} = 0.558$ and $x4 = \mu - \frac{2\sigma}{\sqrt{N}} = 0.442$ on the figure (2) which are the upper and lower bound of 95% confidence interval of the normal distribution with the mean $\mu = \frac{1}{2}$ and the variance $\frac{\sigma^2}{N} = \frac{1}{1200}$. Then I plotted the 50 sample means on the x-axis in the form of a series of discrete points.



From this graph, we can find that most of discrete points fall between two reference lines on the x-axis, which confirms that these sample means are approximately normally distributed with the mean $\mu = \frac{1}{2}$ and the variance $\frac{\sigma^2}{N} = \frac{1}{1200}$. This is exactly conclusion I can draw from mathematical derivation in the Theoretical Explanation or Analysis Section.

For the problem (4):

Because I need to generate a sequence of N+1=1001 random numbers $X_1, X_2 \ldots X_n$

that are $\sim U(0,1)$, I chose the built-in function unifrnd(A,B,m,n) with the parameters A=0, B=1, m=1, n=1001. After running the program, it would return 1×1001 array currentdata5 filled with 1001 random numbers which are $\sim U(0,1)$. Using all of values in this array as the parameter, I used the function mean( ) to compute three arithmetic mean values which are S=0.4972, T=0.4975 and Q=0.2466, I end up getting the value $Z=-7.705\times10^{-4}$. Because Z is not equal to zero, I think it can be concluded that X_i and X_{i+1} are not independent.

4. References

1. *Alberto Leon-Garcia. (2008). Probability, Statistics, and Random Processes for Electrical Engineering. Upper Saddle River, NJ 07458. Pearson Education, Inc.*

2. *Zhou Sheng, Shiqian Xie, Chengyi Pan. (2008). Probability Theory and Mathematical Statistics. No.4, Dewai Street, Xicheng District, Beijing. Higher Education Press.*

5. Source Code

```
[mu,sig1]=unifstat(0,1)            %evaluate the mean and variance of
                                    uniform random variables

currentdata1=rand(10);             %compute the sample mean and sample
                                    variance of 100 random numbers on the
                                    interval of (0,1)
currentmean1=mean(currentdata1(:))
currentvar1=var(currentdata1(:))
currentdata2=rand(100);            %repeat for 10000 random numbers
currentmean2=mean(currentdata2(:))
currentvar2=var(currentdata2(:))
Nrepeat1=100000;
for k=1:Nrepeat1                   %generate 100000 sample means
    currentdata3=rand(100);
    currentmean3(1,k)=mean(currentdata3(:));
end
currentvar3=var(currentmean3(:))   %estimate the (sample) variance of
                                    these sample means

Nrepeat2=50;
```

```
for k=1:Nrepeat2                          %generate 100 random numbers
                                          between (0,1) and repeat this process
                                          for 50 times
    currentdata4=rand(10);
    currentmean4(1,k)=mean(currentdata4(:));   %compute the sample mean
                                               of 100 random numbers
                                               from each of 50 trials
end
figure(1);                                %estimate roughly whether these 50
                                          sample means apply for normal
                                          distribution with built-in function
                                          normplot( )
normplot(currentmean4(:));
x1=0.40:0.0001:0.60;
sig2=(sig1/100)^(0.5);
y1=normpdf(x1,mu,sig2);
figure(2);                                %draw the graph of normal
                                          distribution with parameters mean
                                          mu=0.5 and variance sig2=0.0289
plot(x1,y1,'r');
hold on;
x2=currentmean4;
y2=zeros(1,50);
plot(x2,y2,'k*');                         %plot the 50 sample means on the X-
                                          axis
y3=0:14;
y4=0:14;
x3=(mu+2*sig2)*ones(1,15);
x4=(mu-2*sig2)*ones(1,15);
plot(x3,y3,'b',x4,y4,'b');                %draw two reference lines
                                          x3=mu+2*sig2 and x4=mu-2*sig2 which
                                          are the upper and lower bound of
                                          confidence interval of 95%
grid on;

currentdata5=unifrnd(0,1,1,1001);
S=mean(currentdata5(1,1:1000));
T=mean(currentdata5(1,2:1001));
for j=1:1000
    c(1,j)=currentdata5(1,j)*currentdata5(1,j+1);
end
Q=mean(c(:));
Z=Q-S*T                                   %compute the value of Z
```