**Project #1**

Generate a large (how big?) sequence of random numbers between [0,1] using the rand function in Matlab.

A) Compute basic statistics to test uniformity

- Calculate sample mean and sample variance; compare to expected.

- Can you make a statistical statement about these results?

1.Theoretical Analysis

Let *N* be the length of the sequence of random numbers between [0,1]. Suppose that *X* 1 …, *X N* are these independent random variables in this sequence. The mean of this sample is

and the sample variance *s2*:

2.Description of Program

Because at first there isn’t an accurate number of the variables in the sequence of random numbers between [0,1], I set the sample size from 1 to 5000. It means that there are 5000 samples. And then I get the mean values and variance values of these 5000 samples by using the mean function and variance function. At last, I plot the relationship of between sample size and mean values in the figure 1-1 by using plot function.

3.Simulation Results

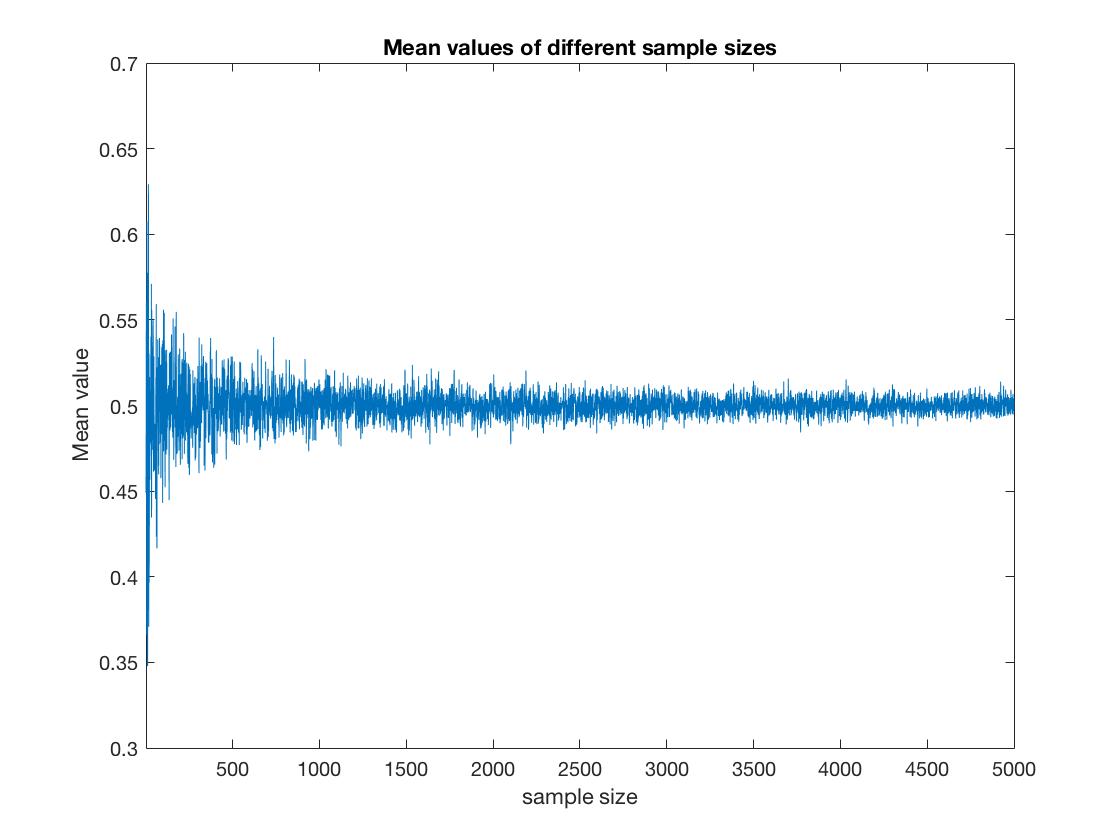


Figure 1 Mean values of different sample sizes

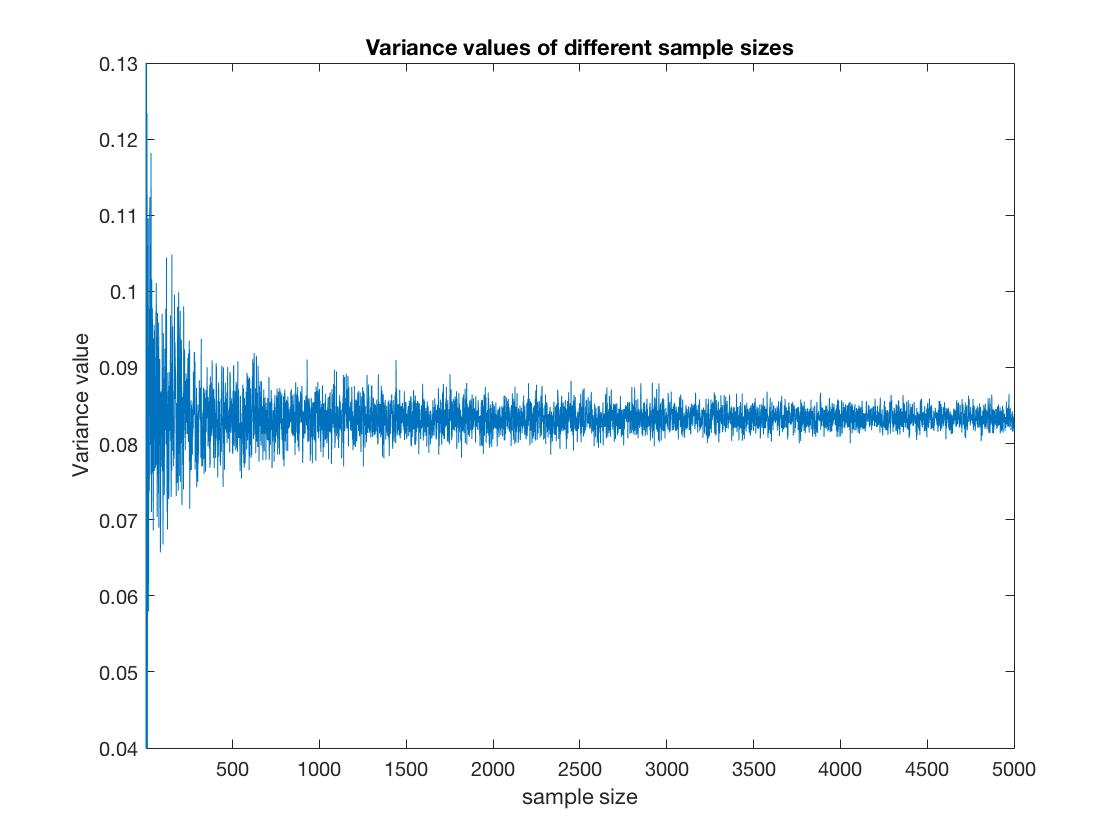


Figure 2 Variance values of different sample sizes

4.Statical Statement

For a sample of uniform random variables. The mean for uniform random variables is given by

when a is 0 and b is 1, then we will get

The variance of the random variable X is defined by

when a is 0 and b is 1, then we will get

From the figure1, we can see that the mean values float up and down at 0.5. And as the sample size becomes larger, the range of the floating of the mean values at 0.5 becomes smaller. We can expect that when the sample size is approaching infinity, the mean value will also approach 0.5, and it will be 0.5 at last. For a sample including uniform random variables between [0,1], the theoretical mean value of the sample is 0.5. So we can get that the results of simulation is in line with the theoretical result.

From the figure2, the variance values float up and down at 0.83. And as the sample size becomes larger, the range of the floating becomes smaller. We can expect that when the sample size is approaching infinity, the variance value will also approach at 0.83, and it will be at 0.83 at last. For a sample including uniform random variables between [0,1], the theoretical variance value of the sample is 0.83. So we can find that the results of simulation is in line with the theoretical result.

B) Independence Tests

- Compute the (sample) covariance *COV(XiXi+k)* for a few different values of *K*. If *Xi* and *Xi+k* are independent what should the covariance be?

1.Theoretical Analysis

The covariance of two random variables *X* and *Y*, denoted Cov( *X*, *Y*), is defined by

Where ,.

We can get a useful expression for Cov(*X*, *Y*) by expanding the right side of the above equation and then making use of the linearity of expectation. This yields

and if *X* and *Y* are independent, then the COV(*X*,*Y*)=0.

2.Description of Program

In the code program, I firstly create an array of 1\*5000 as the sample of *Xi*. And then I create an second array of 100\*5000. The first row of the second array is the sample of *Xi+*1, the second row of the second array is the sample of *Xi+*2 , and the k row of the second array is the sample of *Xi+*k .Then I compute the value of COV(*Xi Xi+k*) with k from 1 to 100 separately. At last, I plot the relationship of between the results of COV(*Xi Xi+k*) and k value in the figure by using plot function.

3.Simulation Results

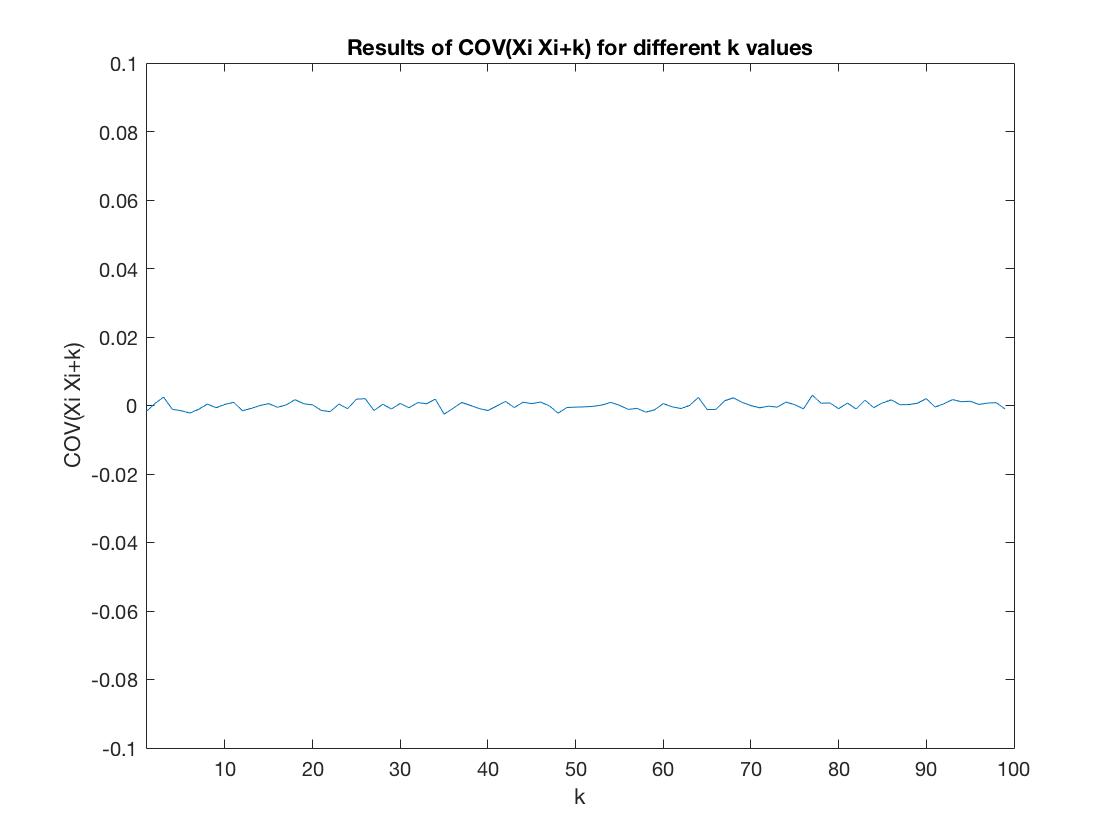


Figure 3 Results of COV(XIXI+K) for different K values

4.Statical Statement

From the figure, we can see that value of COV(*Xi Xi+k*) float up and down at 0. And the floating range is very small. We can approximate value of every COV(*Xi Xi+k*) as 0. At the same time, we know that *X* and *Xi+k* are independent. Theoretically, if *X* and *Y* are independent, then the COV(*X*,*Y*)=0. So we can get that the results of simulation is in line with the theoretical result.

C) Uniform Distribution

Use the data to generate a uniformly distributed discrete random variable that takes on values from 0,1,2,…,9 and generate a histogram (frequency distribution) of the results. How well does this compare to what you expect?

<Optional for bonus points> Can you make a statistical statement about this (Hint: use a goodness of fit test – see text pages 247-249, particularly example 11a)

1.Theoretical Analysis

From the question, we know that it is a sample including uniformly distributed discrete random variables that takes on values from 0,1,2,…,9. So the probability for each value occurs should be 1/10.

2.Description of Program

Firstly, I use a rand function to create an array of 1\*5000, and then multiply each variable with 10. Then I use fix function to make each value take on integers from 0 to 9. Afterwards, I use the hist function to show the frequency distribution of the results. And then, I use the formula of the Chi-Square Goodness of Fit Test to calculate the T value.

3.Simulation Results

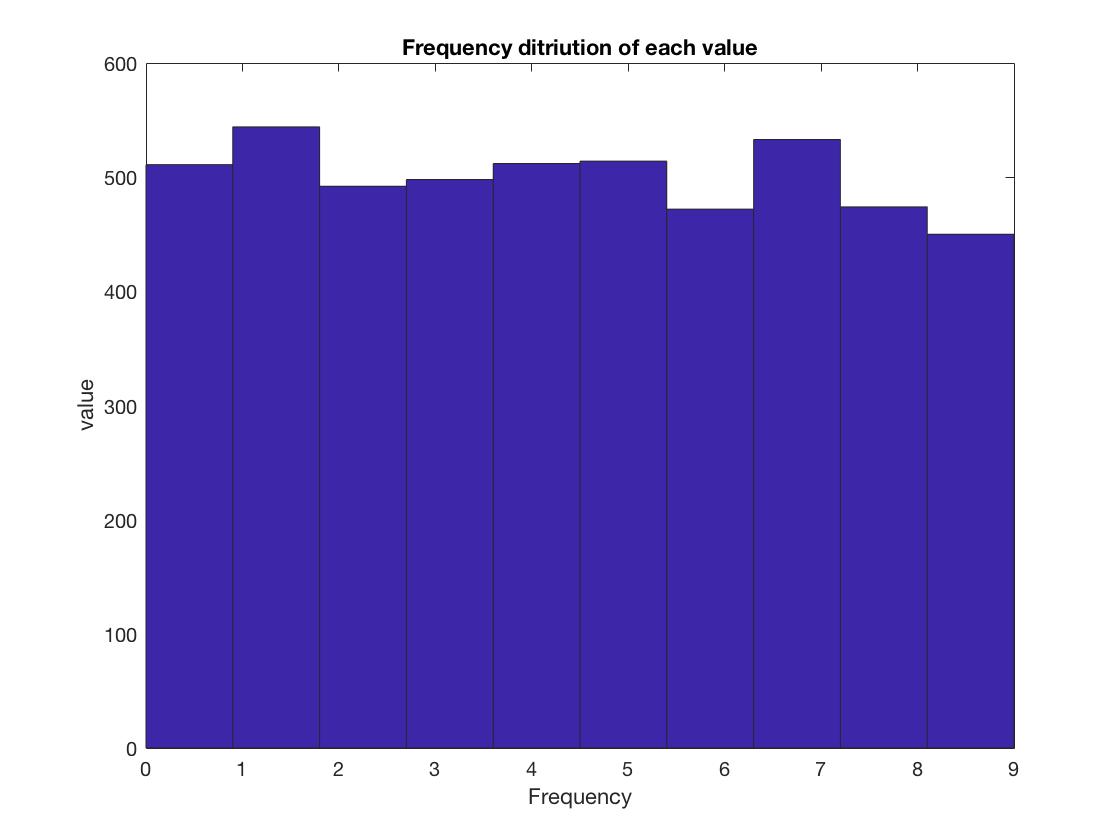


Figure 4 Frequency distribution of each value

Table 1 The experimental frequency and expected frequency

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| expected | 500 | 500 | 500 | 500 | 500 | 500 | 500 | 500 | 500 | 500 |
| experiment | 511 | 544 | 492 | 498 | 512 | 514 | 472 | 533 | 474 | 450 |

4.Statical Statement

We can use a goodness of fit test to compare the expected frequency with the real frequency. By using the formula:

so here n is 5000, and *pi* is1/10; By running the program, we get the T-value is 15.02. and we also know that:

By searching the Table of values vs p-value, we know that the P-value is 0.091. We see that the observed results of occur with a probability roughly 0.091.

Source Code

1.Problem A

Mean\_sequence=zeros(1,5000);

Variance\_sequence=zeros(1,5000);

for i=1:5000 % computing the mean values and variance values

Sequence=rand(1,i);

Mean\_sequence(i)=mean(Sequence);

Variance\_sequence(i)=var(Sequence);

end

figure(1);

plot(Mean\_sequence(1,1:5000));

axis([1 5000 0.3 0.7])

title('Mean values of different sample sizes');

ylabel('Mean value');

xlabel('sample size');

figure(2);

plot(Variance\_sequence(1,1:5000));

axis([1 5000 0.04 0.13])

title('Variance values of different sample sizes');

ylabel('Variance value');

xlabel('sample size');

2. Problem B

X=rand(1,5000);

Y=rand(100,5000);

result=zeros(1,100);

for k=1:100

m=cov(X,Y(k,:)); % computing the COV(XiXi+k)

result(1,k)=m(1,2);

end

figure(3);

plot(result(1,2:100));

axis([1 100 -0.1 0.1])

title('Results of COV(Xi Xi+k) for different k values');

ylabel('COV(Xi Xi+k)');

xlabel('k');

3. Problem C

X=rand(1,5000)\*10;

Sequence=zeros(1,5000);

Y=zeros(1,10);

% make the original random values become integers from 0 to 9;

for i=1:5000

Sequence(i)=fix(X(i)/1);

End

%get the frequency distribution of each value

hist(Sequence);

title('Frequency ditriution of each value');

ylabel('value');

xlabel('Frequency');

% Use the formula to calculate the T-value;

for i=1:5000

Y(Sequence(i)+1)=Y(Sequence(i)+1)+1;

end

sum=0;

for i=1:10

sum=sum+(Y(i)-5000\*(1/10)).^2;

end

T=sum/(5000\*(1/10))