University of Southern California

EE511

Simulation Methods for Stochastic Systems

Project #2

Samples and statistics

BY

Akshay Deepak Hegde

USC ID: 8099460970

[hegdeaks@usc.edu](mailto:hegdeaks@usc.edu)

Question 1:

1. Simulate sampling uniformly (how many?) on the interval [-3 2]

a. Generate a histogram of the outcomes.

b. Compute the sample mean and sample variance for your samples.

How do these values compare to the theoretical values?

If you repeat the experiment will you compute a different sample mean or sample variance?

c. Compute the bootstrap confidence interval (what width?)  for the sample mean and sample standard deviation.

Description:

(a) N random samples in the range [0 1] are generated using rand() function. To sample the data in the interval [x y], the range(y-x) is multiplied with each number and scaled by adding x to it. I get the histogram of the data set in the interval [-3 2] using histogram() function.

(b) Mean() function will give the sample mean and var() function will give the sample variance of my samples.

Let us calculate the sample mean and variance theoretically,

If x={x1,x2,……..xn} is a sequence of N values, then the sample mean is given by,

Mean=(1/N)\*(sum(x));

The sample variance represents the measure of the spread of data with respect to mean. It is

given by, Variance=(1/(N-1))\*(sum(x(i)- Mean)^.2)

Where x(i) is the value of the ith sample and i runs from 1 to N.

For a random variable X,

Mean(X)=(a+b)/2 and Variance(X)= (b-a).^2/12

So, in our case, a= -3 b=2. Hence, the theoretical values of mean and variance will be

Mean = (-3+2)/2 = -0.5

Variance = (2-(- 3))^2/12 = 2.5

And, values from the MATLAB code are as below,

Mean = -0.50544

Variance = 2.0703

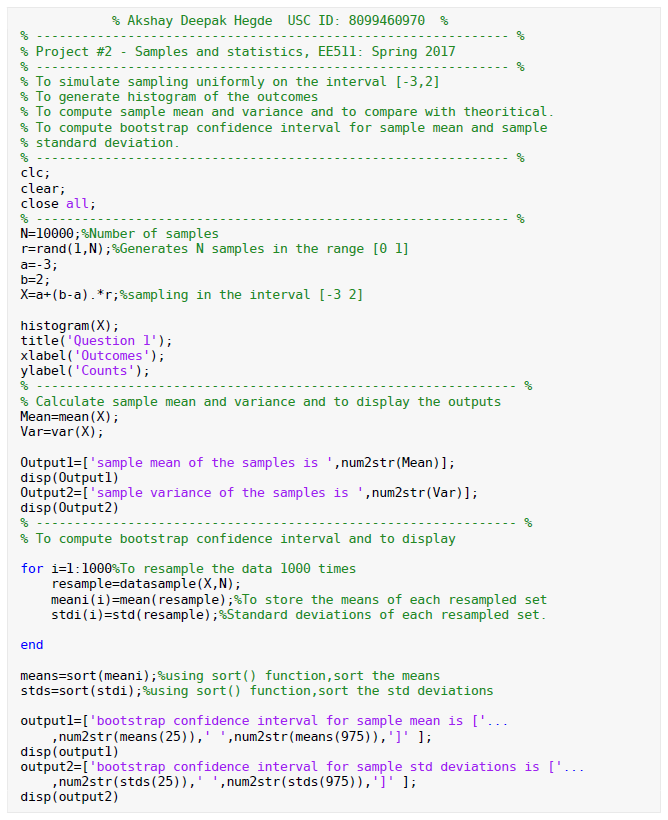
Comment: The theoretical values are close to the observed ones and hence I can say that I have simulated the sample properly. Also, if I repeat the experiment multiple times, I get different mean and variance since the random numbers generated are different. But, as the number of samples increases(N), the mean and variance tend to meet the theoretical values.

(c) Bootstrap confidence Interval:

Bootstrap is a statistical technique where in we perform the computations on the data set itself to estimate the variation of the statistics. It is nothing but sampling with replacement. A confidence interval is the interval which is likely to contain the parameter of interest. 95% of confidence interval would mean that, if the same data is resampled number of times and intervals are estimated each time, then the resulting interval would contain the parameter of our interest 95% of chances.

datasample() function is used to resample the data. A loop is run 1000 times and resampling is done on each occasion. Mean and standard deviation of each resampled data is stored in arrays. The resulting arrays are sorted using sort() function. Since the 95% of confidence level is set, the confidence interval is calculated between 25th and 975th mean value.

MATLAB Code:



Output:

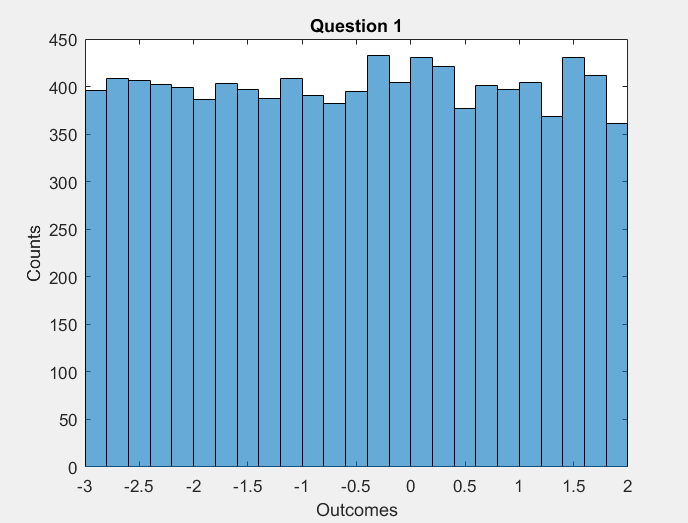
sample mean of the samples is -0.5022

sample variance of the samples is 2.0727

bootstrap confidence interval for sample mean is [-0.52875 -0.47435]

bootstrap confidence interval for sample std deviations is [1.4262 1.4527]

Histogram:



Question 2:

Produce a sequence X by drawing samples from a standard uniform random variable.

a. Compute Cov[Xk,Xk+1]. Are Xk and Xk+1 uncorrelated? what can you

conclude about the independence of Xk and Xk+1?

b. Compute a new sequence Y where: Y[k]= X[k]-2.X[k-1]+0.5.X[k-2]-X[k-3]. Assume X[k]=0 for k<=0. Compute Cov(Xk,Yk). Are they uncorrelated?

Description:

Covariance is a statistical calculation that helps you understand how two sets of data are related to each other. Covariance is a measure of how changes in one variable are associated with changes in a second variable. Specifically, covariance measures the degree to which two variables are linearly associated.

Let Xand Y be random variables (discrete or continuous!) with means μX and μY. The **covariance** of Xand Y, denoted ***Cov*(*X*,*Y*)** or ***σXY*,** is defined as:

Cov(X,Y)=σXY=E[(X−μX)(Y−μY)]

(a) N random variable samples are generated using rand() function and stored as X sequence. To get the time shifted sequence of the sequence, a zero is padded as the first element of the sequence and first N samples are obtained to store it as Xk+1 sequence. Covariance of both the sequences is calculated using inbuilt cov() function and displayed using disp() function.

Comment:

X and Y are said to be uncorrelated if Cov(X,Y)=0.

In our case, covariance is 0.01022. Since it is almost equal to zero, I can say that the two sequences are uncorrelated.

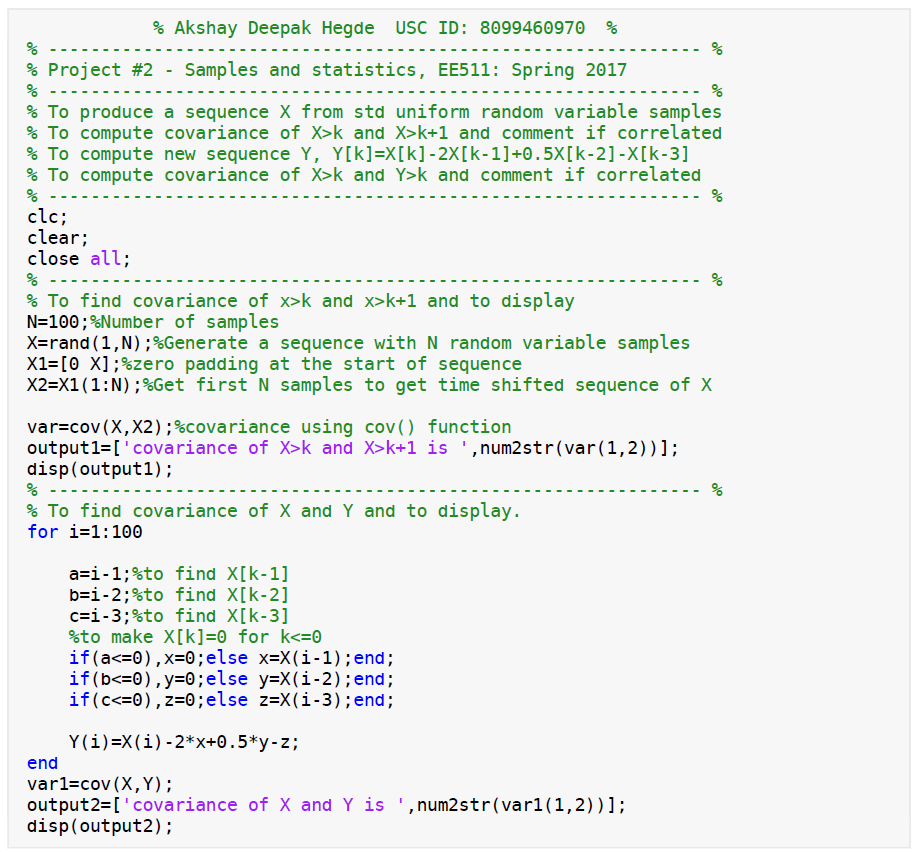
Also, if two variables are [independent](http://en.wikipedia.org/wiki/Independence_%28probability_theory%29), their covariance is 0. But, having a covariance of 0 does not imply the variables are independent. Since one sequence is time shifted version of another, I feel they are not independent.

(b) A new sequence Y is calculated with the given conditions. A loop is run and X[k]=0 for k<=0 is obtained. Covariance of X and Y are calculated using cov() function and output is displayed.

Comment:

The covariance value obtained is .080354. Since it is not equal to zero, I can conclude that they are correlated.

MATLAB Code:



Output:

covariance of X>k and X>k+1 is 0.01022

covariance of X and Y is 0.080354

Question 3:

Let M = 10.  Simulate (uniform) sampling with replacement from the outcomes 0, 1, 2, 3, …, M‐1.

a. Generate a histogram of the outcomes.

b.Perform a statistical goodnesfit test to conclude at the 95% confidence level if your data fits samples from a discrete uniform distribution 0, 1, 2, …, 9.

c.Repeat (b) to see if your data(the same data from b) instead fit an alternate uniform distribution 1, 2, 3,.. , 10.

Description:

Goodness of fit test:

If we have an observed set of data and want to know how well it reflects the data, that is, how close are the observed values to those which would be expected, we can use the chi-square goodness of fit test. Chi-square goodness-of-fit is very appropriate when the sampling method is random sampling.

For a random variable X it is given by

X^2=(sum(observed values-expected values)^2)/expected values

(a) datasample() function is used to generate and sample from the data with replacement. Since I have to sample from the outcomes on 0,1,..9, (0:M-1) range is mentioned in the command where M=10. Replace argument is given true to sample N sequences with replacement. Histogram of the outcomes is generated using histogram() function.

(b) To perform statistical goodness-of-fit, the expected value is calculated considering equal probability. So 1/M is multiplied with N samples and (1,M) array with ones to get the expected value sequence. To get the distribution from 0,1,...,9, ranges of bin is taken as 1:9 and using histc() function counts of values in associated with each x/ histogram bin counts are calculated. These values are stored in an array as observed values.

(c) To get the distribution from 1,2,..10, ranges of bin is specified as 1:10. histc() is used to count the histogram bin counts and the values are stored in an array as set of observed values.

Comment:

Chi-Square Threshold is calculated using chi2inv() function. 95% of confidence interval is set with degrees of freedom 9(as the total number of categories is 10). The Chi-square threshold value thus calculated comes to be 16.9190.

Now, Chi-squared values are calculated using observed values and expected value using,

X^2=(sum(observed values-expected values)^2)/expected values

For case (b), this value is 13.9040

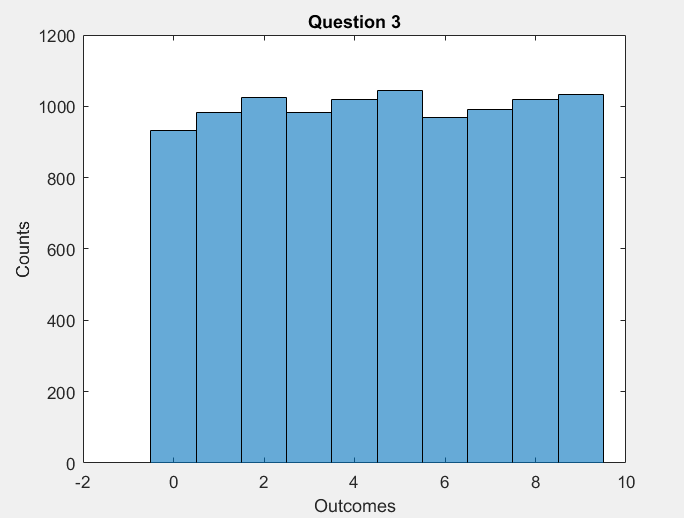
For case (c), this value is 1013.9

Since 13.9040<16.9190, the data fits sample from a uniform distribution 0,1,2,..,9

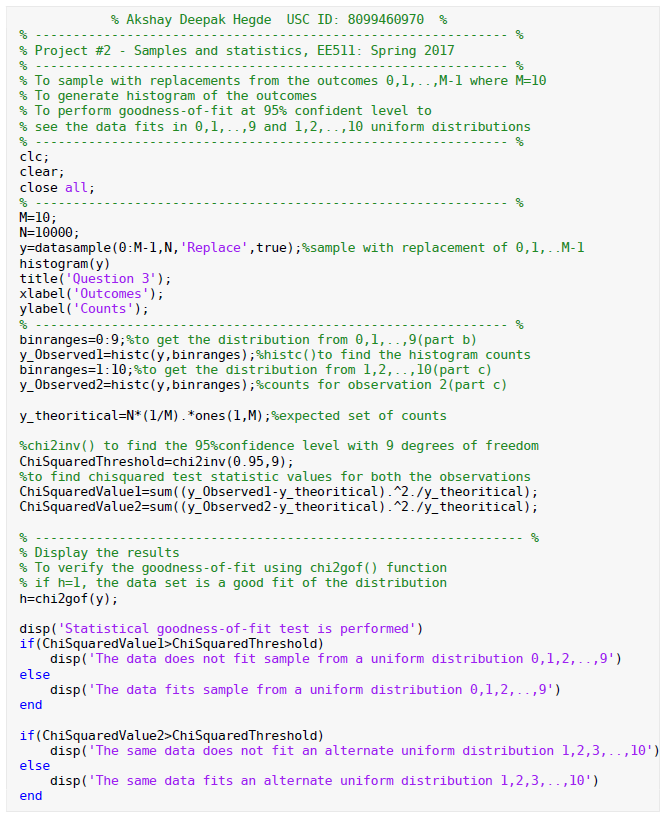
And as 1013.9>16.9190, the same data does not fit an alternate uniform distribution 1,2,3,..,10

Also, we can use chi2gof() function to check the goodness-of-fit. If it returns 1, the data fits sample from particular distribution. It does not return if it returns 0.

Histogram:



MATLAB Code:



Output:

Statistical goodness-of-fit test is performed

The data fits sample from a uniform distribution 0,1,2,..,9

The same data does not fit an alternate uniform distribution 1,2,3,..,10