University of Southern California EE511

Samples and Statistics

Project #2

Wu Jiawei
2017-1-24

Abstract

In the project of samples and statistics, three experiments are conducted using Matlab. The core method of the project is to sample randomly, obtain the statistic characteristics of samples through Matlab simulation, and analysis the simulation result using different mathematical techniques. The experiment outcomes are shown in diagrams and calculations to reflect the characteristics of the total population.

Introduction

Three experiments were conducted in the lab. All the samples are generated by random selection from the giving interval. The goal of the first trial is computing the bootstrap confidence interval for the sample mean and sample standard deviation. The aim of the second experiment is to detect the sample independence by calculating the correlation coefficient. The objective of the final trial is performing a statistical goodness-of-fit test using chi-square test method to test whether at the 95% confidence level the experiment data fit the standards.

Methodology & Results

Experiment No.1

Question:

Simulate sampling uniformly 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 for the sample mean and sample standard deviation.

Code:

```
function f=pro(N)%N represents the number of a sample
alpha=0.05;
arr_Mu=zeros(1,N);
arr_Std=zeros(1,N);%Create two arrays to store means and standard deviations
of different samples
Rand_Num=randi([-3 2],1,N);
x=-3:1:2;
figure(1);
hist(Rand_Num,x);%Generate a histogram of the sampling outcome
title('Histogram of Sampling Outcome')
xlabel('Numbers in the Sample');
ylabel('Quantity of Numbers')
theo_Mu=(-3+2)/2
sample_Mu1=mean(Rand_Num)
theo_var=(2-(-3))^2/12
sample_var=var(Rand_Num)%Calculate experimental values
for i=1:N
Rand_Ind=randi([1 N],1,N);%Every loop, different individuals of different
```

```
order are chosen with replacement from the original sample.
Rand_Sel=Rand_Num(:,Rand_Ind);%Resample with same amount and different
individuals
sample_Mu=mean(Rand_Sel);
arr_Mu(i)=sample_Mu;
sample_Std=std(Rand_Sel);
arr_Std(i)=sample_Std;%Store different values of mean and std in arrays
end

c=sort(arr_Mu);
Mu_X1=prctile(arr_Mu,2.5)
Mu_X2=prctile(arr_Mu,97.5)%show the width of bootstrap confidence interval
for sample mean
CI = prctile(c,[100*alpha/2,100*(1-alpha/2)]);%Define the confidence interval
figure(2);
hist(c);
hold on;
T1=plot([CI(1),CI(1)], get(gca, 'YLim'), '-r', 'LineWidth', 3)
T2=plot([CI(2),CI(2)], get(gca, 'YLim'), '-y', 'LineWidth', 3)
hold off;
title('Bootstrap Confidence Interval for Sample Mean');
xlabel('Mean of Samples');
ylabel('Quantity of Mean');
legend([T1,T2],'2.5% Sample Mean','97.5% Sample Mean');

d=sort(arr_Std);
Std_X1=prctile(arr_Std,2.5)
Std_X2=prctile(arr_Std,97.5)%show the width of bootstrap confidence interval
for the sample standard deviation
CII = prctile(d,[100*alpha/2,100*(1-alpha/2)]);
figure(3);
hist(d);
hold on;
P1=plot([CII(1),CII(1)], get(gca, 'YLim'), '-r', 'LineWidth', 2)
P2=plot([CII(2),CII(2)], get(gca, 'YLim'), '-y', 'LineWidth', 2)
hold off;
title('Bootstrap Confidence Interval for Sample Standard Deviation');
xlabel('Standard Deviation of Samples');
ylabel('Quantity of Standard Deviation');
legend([P1,P2],'2.5% Sample Std','97.5% Sample Std');
```

Simulation Result:

Command Window

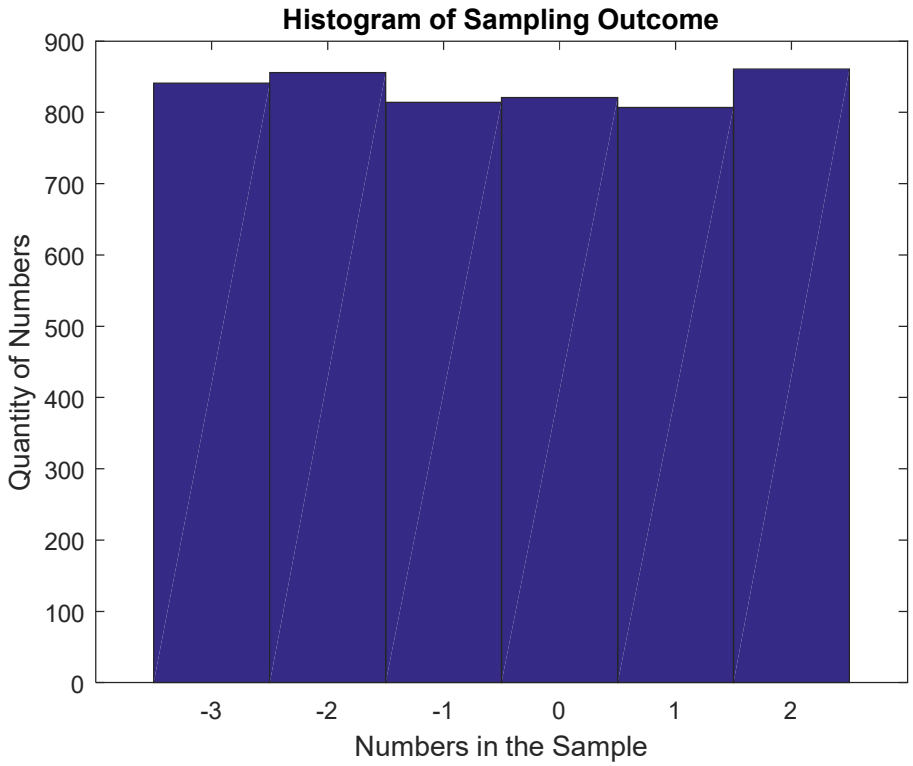
```
>> pro(5000)

theo_Mu =

   -0.5000


sample_Mu1 =

   -0.5040


theo_var =

    2.0833


sample_var =

    2.9582
```

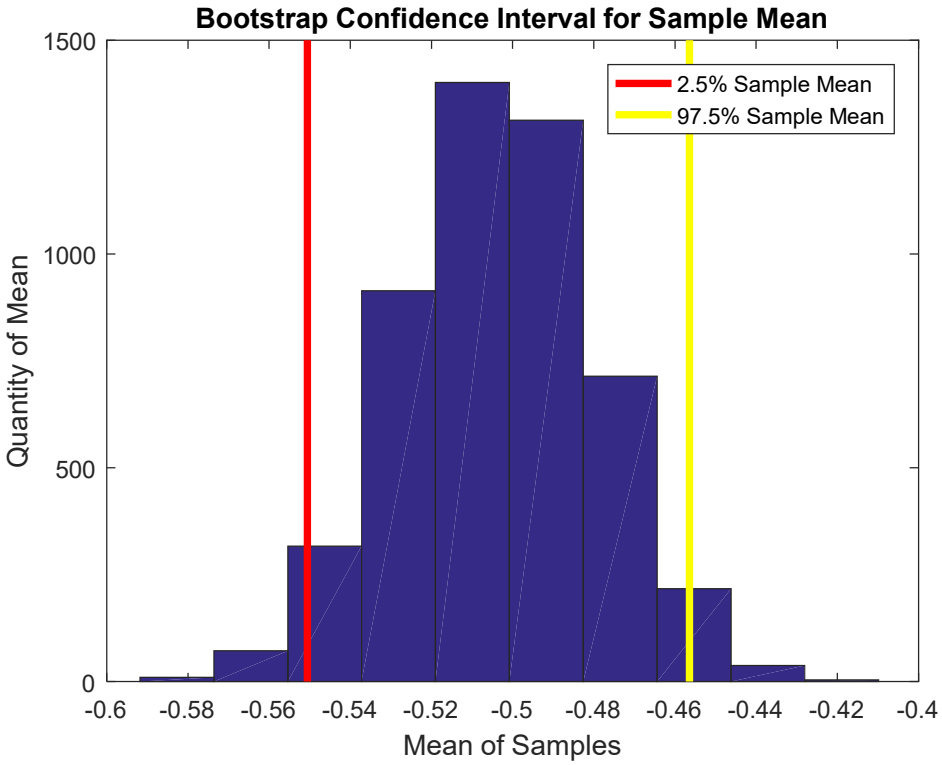
Command Window

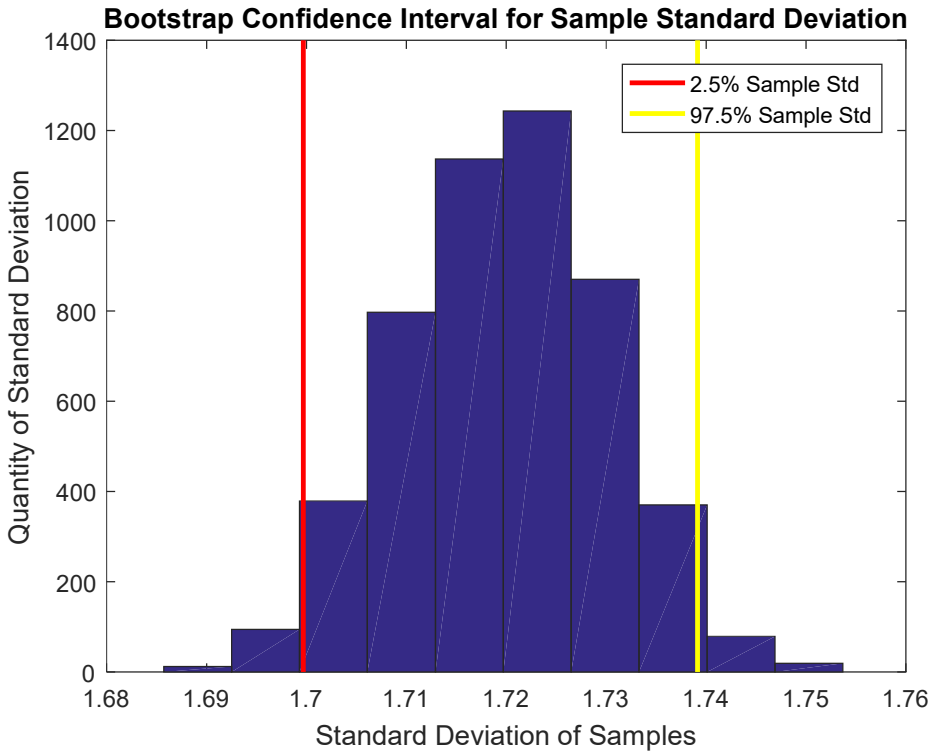
```
Mu_X1 =

   -0.5506


Mu_X2 =

   -0.4564
```

Command Window

```
Std_X1 =

    1.6997


Std_X2 =

    1.7391
```


Histogram of Sampling Outcome
Quantity of Numbers
Numbers in the Sample
0
100
200
300
400
500
600
700
800
900
-3
-2
-1
0
1
2


Bootstrap Confidence Interval for Sample Mean
2.5% Sample Mean
97.5% Sample Mean
Quantity of Mean
Mean of Samples
0
500
1000
1500
-0.6
-0.58
-0.56
-0.54
-0.52
-0.5
-0.48
-0.46
-0.44
-0.42
-0.4



Comment:

1). In the first trial, experimental data are sampled uniformly on the interval [-3,2] and the sampling outcomes are recorded by generating a histogram. When the amount of sample is 5000, it can be seen from the diagram that each of number ranging from -3 to 2 all appears in the samples, and the quantity of each number -3,-2,-1,…,2 is approximately equal to 800. Therefore, the sample meet the criterion of "uniformly sampling", so the amount 5000 is a fair and enough sample amount in this experiment.

2). Function "var" and "mean" are used to calculate the experimental value of sample variance and sample mean. The theoretical value of sample mean and variance are obtained through mathematical calculation. From the result of simulation, the experimental sample mean is -0.5040, which is approximately equal to the theoretical mean -0.5000. Similarly, the experimental sample variance is 2.9582, which is nearly same as the theoretical variance value 2.0833 while there exists small number of difference in the decimal place. If I repeat the experiment multiple times, the experiment would compute different values of sample mean and sample variance.

3). The process of computing the bootstrap confidence interval is completed through programming instead of using the function in Matlab directly. I create the variable "Rand_Ind" to record the order (index) of different numbers in the original sample array, and simulate the processing of resampling by randomly selecting individual with replacement , and the new samples generated in this process are of same amount with the original sample. Thus, the resampling process are done with same amount but different individuals from the original samples.

4). The "for" loop is used in the programming to count the mean and standard deviation of new samples N times. The value of means and standard deviations for each loop are stored into two arrays named "arr_Mu" and "arr_Std". Then two histograms are created to show the quantity of different value of means and standard deviations. Finally, I used function "prctile" to calculate the value of 2.5% and

97.5% of the distribution. Two histograms are generated to show the bootstrap confidence interval for the sample mean and sample standard deviation.

Experiment No.2

Question:

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

a. Compute $\mathrm{Cov}[X_k, X_{k+1}]$. Are X_k and X_{k+1} uncorrelated? What can you conclude about the independence of X_k and X_{k+1}?

Code:

```
function f=count(N)
Rand_Num=rand(1,N);
X_k=zeros(1,N-1);
for i=1:N-1
X_k(i)=Rand_Num(i);%Sequence X_k contains numbers ranging from the first to
the penultimate of the array "Rand_Num"
end
X_k1=zeros(1,N-1);%Sequence X_k1 contains numbers ranging from the second to
the last of the array "Rand_Num"
for i=2:N
X_k1(i-1)=Rand_Num(i);%Sequence X_k and X_k1 have the same length
end
Cov_Num=cov(X_k,X_k1);
Cov_Num(1,2)%Calculate the correlation coefficent
```

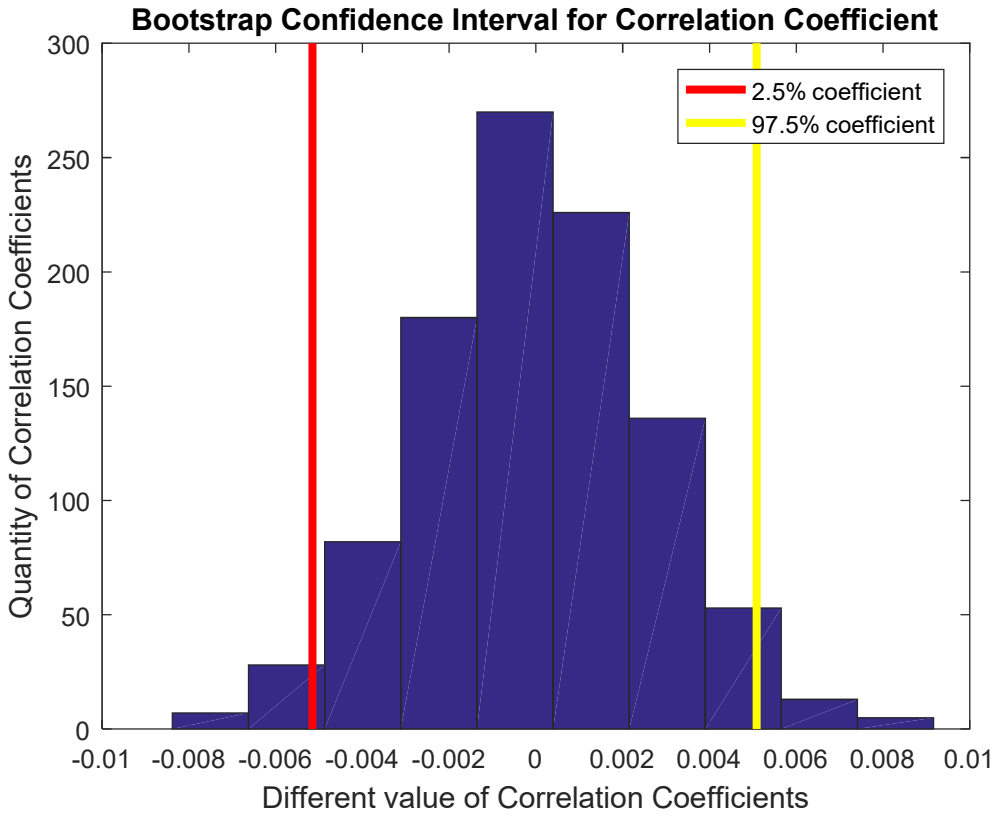
Simulation Result:

Comment:

1). In the first question, the 1*N rand array is generated by sampling randomly from a standard uniform random variable (interval [0,1]). The first sequence is consisted of the elements from the first to the penultimate; and the second sequence is consisted of the elements from the second to the last. The function "cov" is used to count the correlation coefficient of two sequences.

2). From the experimental outcome, the value of correlation coefficient is 0.00083089, which is approximately close to zero. However, when I repeat the experiment multiple times such as 1000 times, the value of correlation coefficient varies from negative number to positive number. In order to decide whether the sequence X_k and X_{k+1} are correlated, I use the method in the first

question by computing the bootstrap confidence interval. It can been seen from the histogram that 95% of correlation coefficient are ranging between -0.0052 to +0.0051; and the majority of correlation coefficient calculated is zero which can be achieved in ideal selection condition. Therefore, the conclusion can be reached that the sequence X_k and X_{k+1} are uncorrelated. The independence between X_k and X_{k+1} can not been decided, since uncorrelated is not necessarily independent.



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 $k \leq 0$. Compute $Cov[X_k, Y_k]$. Are X_k and Y_k uncorrelated?

Code:

```
function y=digui(N)
x=rand(1,N);
for i=1:N
    if N==1
        y(1)=x(1);
        y=[y(1)];
    elseif N==2
        y(1)=x(1);
        y(2)=x(2)-2*x(1);
        y=[y(1),y(2)];
    elseif N==3
        y(1)=x(1);
        y(2)=x(2)-2*x(1);
```

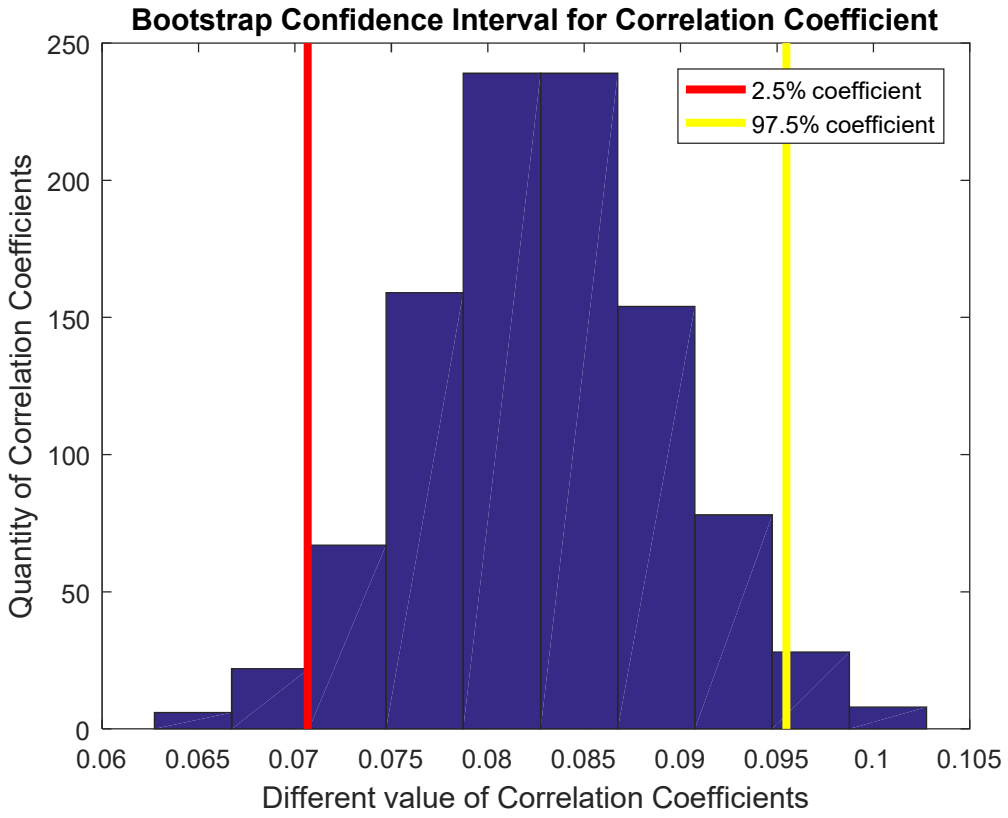
```
        y(3)=x(3)-2*x(2)+0.5*x(1);
        y=[y(1),y(2),y(3)];
    else
        y(1)=x(1);
        y(2)=x(2)-2*x(1);
        y(3)=x(3)-2*x(2)+0.5*x(1);
        Y=zeros(1,N-3);
        for i=4:N
            y(i-3)=x(i)-2*x(i-1)+0.5*x(i-2)-x(i-3);
            Y(i-3)=y(i-3);
        end
    y=[y(1),y(2),y(3),Y];%Generate the array to store the value of y
    end
end
Cov_Num=cov(x,y);
Cov_Num(1,2)
```

Simulation Result:

Comment:

1). In the second question, the sequence Y[k] is generated based on the giving formula, and the value of Y[k] is depend on the value of X[k], X[k-1], X[k-2] and X[k-3]. Therefore, I hypothesis that the two sequences X[k] and Y[k] are correlated based on the recursive function. From the simulation result, when the amount of sample is 10000, the correlation coefficient of two sequences is 0.0832.

2). When I repeat the experiment multiple times such as 1000 times, different values of correlation coefficients are calculated. In order to decide whether the sequence X_k and Y_k are correlated, I use the method in the first question by computing the bootstrap confidence interval. It can been seen from the histogram that 95% of correlation coefficient are ranging between +0.07 to +0.095, the majority of correlation coefficient calculated is 0.085, and all of the correlation coefficients are nonzero. Therefore, the conclusion can be reached that the sequence X_k and X_{k+1} are correlated, which proves the hypothesis.



Experiment No.3

Question:

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

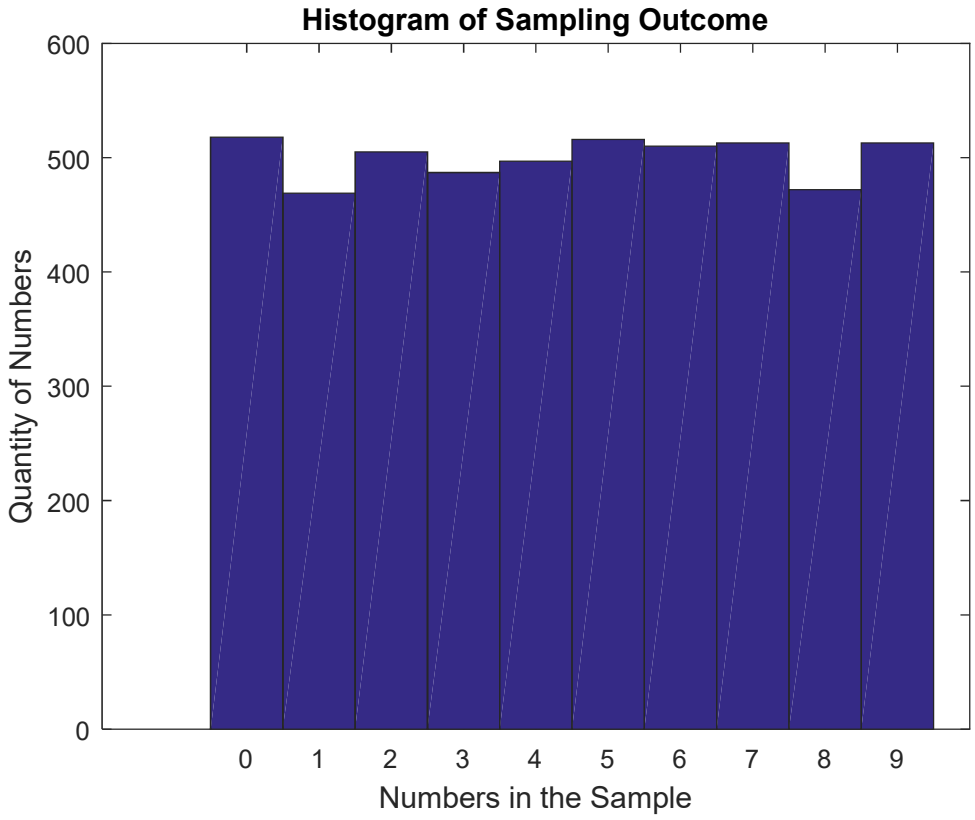
a. Generate a histogram of the outcomes.
b. Perform a statistical goodness-of-fit test to conclude at the 95% confidence level if your data fits samples from a discrete uniform distribution 0,1,2,…,9.
c. Repeat (b) to see if your data (the same data from b) instead fit an alternate uniform distribution 1,2,3,…,10.

Code:

```
function f=kafang(N)
x = randi([0 9],1,N);
t=0:1:9;
hist(x,t);%Generate a histogram to show the sampling result
x_theo = N*[1/10, 1/10, 1/10, 1/10, 1/10, 1/10, 1/10, 1/10, 1/10, 1/10];
a=zeros(0,9);
for i=1:10
    a(i)=sum(x==(i-1));%Calculate the quantity of each number appears in the
rand array "x"
end
Chi_Test = sum((a-x_theo).^2./x_theo)%calculate the experimental Chi-square
test statistic value
Chi_Threshold = chi2inv(0.95,9)%Return the standard value of Chi-square test
```

```
with probability of 95% and 9-degrees freedom.

b=a(1,2:10);
c=[b 0];%Replace the sample to judge whether the same data fit the
alternative uniform distribution ranging from 1 to 10
Chi_TestC = sum((c-x_theo).^2./x_theo)
```

Simulation Result:



Comment:

1). In the third experiment, experimental data are generated by simulating sampling uniformly and with replacement from the interval [0,9]. The histogram is created to show the result. As It is shown in the histogram, when the amount of sample is 5000, all different numbers from 0 to 9 appear in the sample, and the quantity of each number is approximately same and closely to 500. Therefore, the sample meet the criterion of "uniformly sampling", so the amount 5000 is a fair and enough sample amount in this experiment.

2). The theoretical value for each number appeared is stored in the array "x_theo". the Chi-Square Test formula is used to calculate the experimental value of X^2. Then I use function "chi2inv" to calculate the standard value of X^2 when choose the degree of freedom as 9 and the probability as 95%.

3). The simulation result shows that the theoretical value of X^2 is 16.9190 and the experimental value of Chi-Square-Test is 5.932. The experimental value of Chi-Square-Test is less than the critical value. Therefore, with the probability of 95%, the data fits samples from a discrete uniform distribution 0,1,2,…,9. Similarly, since the value of Chi-Square-Test for the number 1,2,3,…,10 is 505.2840, which is much larger than the critical value of X^2 16.9190. Therefore, the data do not fit an alternative uniform distribution 1,2,3,…,10.

Conclusion

Overall, the simulation results of three experiments reflect the statistics property of the whole population by calculating the mathematic characteristics of samples. The first experiment obtains the sample mean and sample standard deviation by computing the bootstrap confidence interval. The second experiment proves the relations between two sequences by calculating the correlation coefficient. The third experiment conducts the statistical goodness-of-fit tests for data. In conclusion, the project uses mathematical methods to process samples, and reflect the statistic property of the whole population.