University of Southern California EE511

Samples and Statistics

Project #3

Wu Jiawei
2017-1-24

Abstract

In the project of samples and statistics, five experiments are conducted using Matlab. The core method of the project is to sample randomly, obtain the statistic mean and variance of samples by simulation, and analysis the simulation result through comparing with the theoretical values. The experiment outcomes are shown in histograms and calculations to reflect the characteristics of different probability distribution functions.

Introduction

Five experiments were conducted in the lab. All the samples are generated by randomly and the experiments are repeated multiple times. The goal of the first trial is to simulate sampling from the hypergeometric distribution. The aim of the second experiment is to simulate the property of Poisson distribution. The objective of the fourth trial is run simulation and sampling from geometric distribution, and the goal of the final experiment is using the accept-reject method to sample from the giving distribution.

Methodology & Results

Experiment No.1

Question:

A components manufacturer delivers a batch of 125 microchips to a parts distributor. The distributor checks for lot conformance by counting the number of defective chips in a random sampling (without replacement) of the lot. If the distributor finds any defective chips in the sample it rejects the entire lot. Suppose that there are six defective units in the lot of 125 microchips. Simulate the lot sampling to estimate the probability that the distributor will reject this lot 95% of the time?

Algorithm:

The experimental samples follow the hypergeometric distribution. The probability mass function is giving by

$$\mathrm{P}(\mathrm{X}=\mathrm{k})=\frac{\binom{K}{k}*\binom{N-K}{n-k}}{\binom{N}{n}}$$

Here we are sampling n = 5 items from (N-K) =119 good items and K = 6 bad items without replacement.

$$\mathrm{P}[\text{reject}]=1-\mathrm{p}[\text{accept}]=1-\mathrm{p}[0\ \text{defective}]$$

Thus, I compute a theoretical value for $\mathrm{p}[0\ \text{defective}]$ and determine $\mathrm{P}[\text{reject}]$.

In the first part of the experiment, I run simulations with different number of sample sizes to estimate the value, and compare the simulation result with the theoretical value in the end.

In the second part of the experiment, since there are defective chips in the lot, we want the probability of rejection to be high enough. You need to set a number k which is the number of selected chips and repeat the first part. Find the k that achieves 0.95 probability of rejection.

Description of Method:

The experiment consists of two parts. In the first part, the rejection probability after testing five chips is simulated. The trial uses "Randperm(125,6)<=5" to simulates the process that distributor tests randomly and checks whether any of six defective chips is mixed in five sample. The "for" loop is used to calculate the sum of rejection times in "N" total times and obtain the test probability. The theoretical probability is calculated by using "nchoosek" to simulate the sampling in hypergeometric distribution.

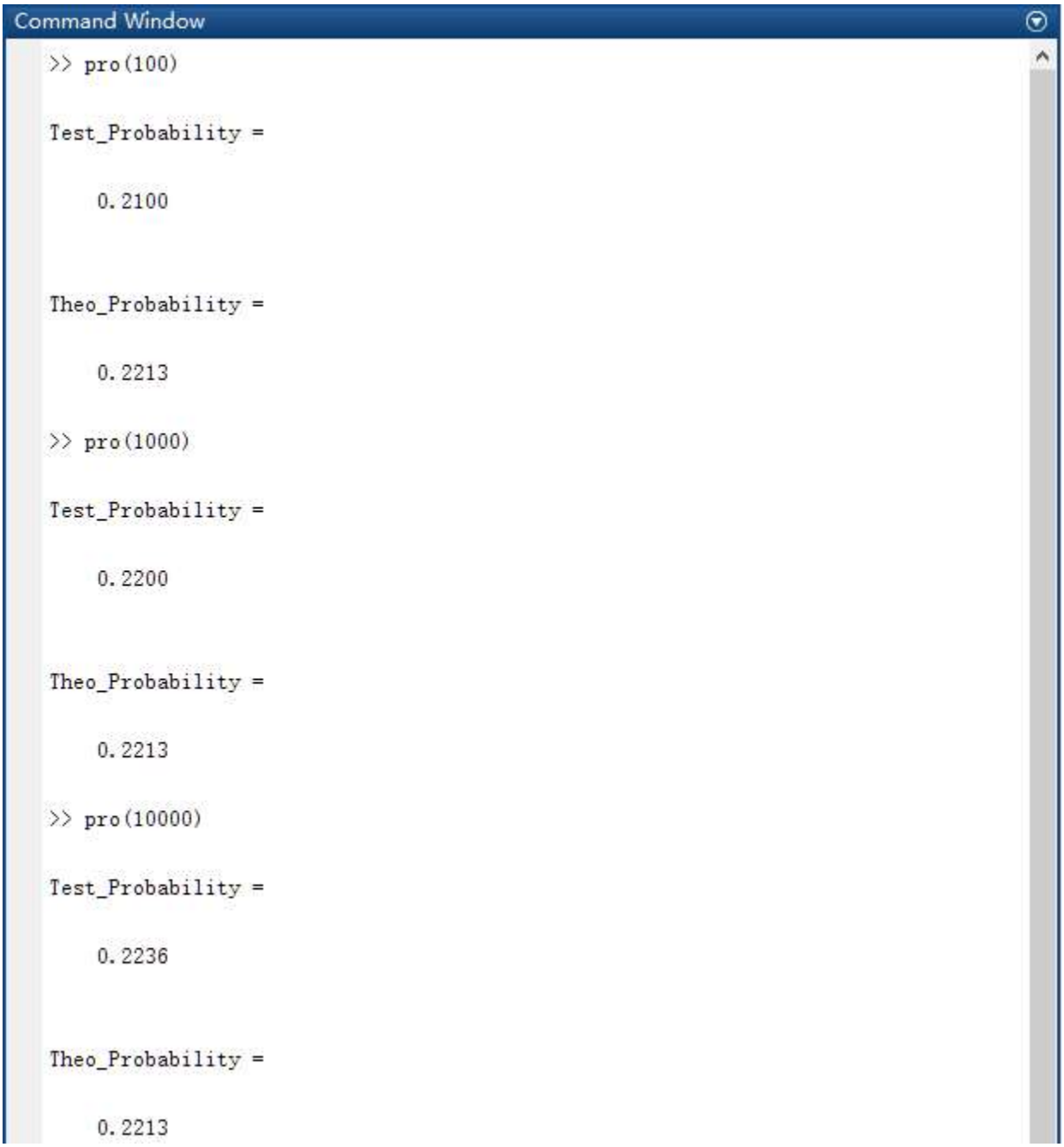
Similarly, in the second part, I increase the amount of selection in each test and count the rejection probability accordingly. A "while" loop is used to return the value of selection amount "K" if the rejection probability is more than 95%. Different sample sizes are set to obtain the fewest number of chips that should be test to reject the lot 95% of times.

Part 1:

Code:

```
function f=pro(N)%N represent the sample size
counts=zeros(1,N);
for a1=1:N
    counts(a1)=sum(randperm(125,6)<=5)>0;
end
probRejection = sum(counts)/N;
Test_Probability=probRejection

c1=nchoosek(119,5);
c2=nchoosek(125,5);
p_accept=c1./c2;
p_reject = 1-p_accept;%Calculate the theoretical value of rejection
probability.
Theo_Probability = p_reject
```

Simulation Result:



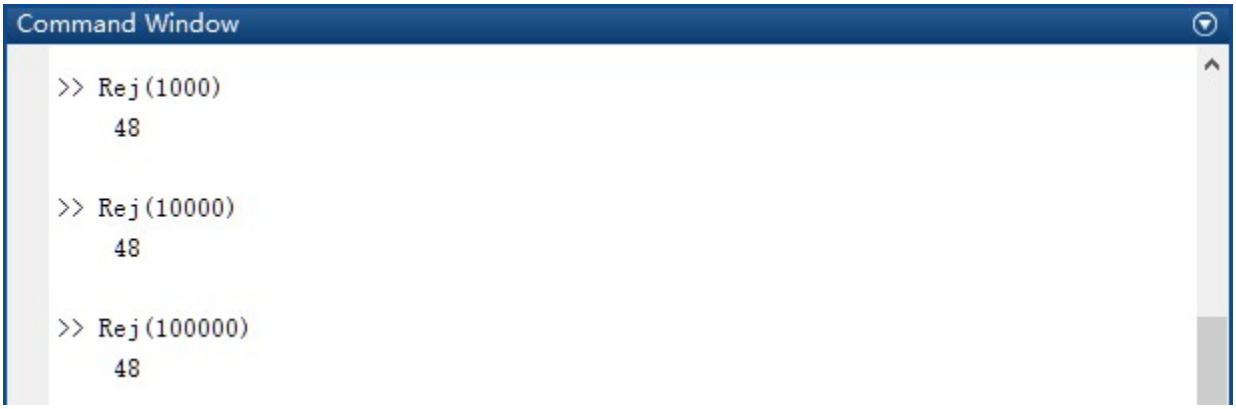
Part2:

Code:

```
function f=Rej(N)
counts=zeros(1,N);
k=1;
while 1
    k=k+1;
for a1=1:N
    counts(a1)=sum(randperm(125,6)<=k)>0;
```

```
end
probRejection = sum(counts)/N;
Test_Value=probRejection;
if Test_Value >= 0.95
    break
end
end
disp(k);
```

Simulation Result:



Finding:

In the first question, I calculate the theoretical value $\mathrm{P}[\mathrm{reject}] = 0.2213$, and run simulations with sample sizes 100, 1000, 10000, and 100000 to estimate the value. The test probability is 0.2100, 0.2270, 0.2199 and 0.2203 respectively. The comparison shows that the test value of experiment is approximately equal to the theoretical value.

In the second question, the fewest number is the number that the distributor need to test every time. Thus, I simulate with sample size 1000, 10000, and 100000. The result shows that the fewest number of microchips that the distributor should test is 48 to reject this lot 95% of the time.

Experiment No.2

Question:

Suppose that 120 cars arrive at a freeway onramp per hour on average. Simulate one hour of arrivals to the freeway onramp: (1) subdivide the hour into small time intervals and then (2) perform a Bernoulli trial to indicate a car arrival within each small time interval. Generate a histogram for the number of arrivals per hour. Repeat the counting experiment by sampling directly from an equivalent Poisson distribution by using the inverse transform method. Generate a histogram for the number of arrivals per hour using this method. Overlay the theoretical p.m.f. on both histograms. Comment on the results.

Algorithm:

The experimental samples follow the Poisson distribution because the events occur completely at random in the time period but with a fixed rate which is denoted as λ. Here the value of λ is 120.

$$P_X(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & \text{if } x = 0,1,2\cdots \\ 0 & \text{otherwise} \end{cases}$$

In the first question, I cut the interval of 1 hour into N=10000 small subintervals (<1 second) and conduct Bernoulli trial in each subinterval with success probability $p = \frac{\lambda}{N}$, which is 0.012. I sample a series of uniformly random variables and count the number of successes to simulate Poisson distribution.

$$P_X(x) = \begin{cases} 1 & \text{if } x < p \\ 0 & \text{if } x \geq p \end{cases}$$

And

$$f(k; p) = p^k(1-p)^{1-k} \qquad \text{for } k \in \{0,1\}$$

In the second question, I sample directly from an equivalent Poisson distribution by using the inverse transform method. Then I generate a histogram and overlay the theoretical p.m.f. on it.

Description of Method:

The second experiment are completed in two methods. Firstly, the one hour interval is divided into "N_interval=100000" subintervals. A "for" loop is used to calculate the sum of result 1 in Bernoulli trials with successful probability "p=lamada/10000". Another "for" loop is used to repeat trial N times and store results in an empty array "arr". The histogram is generated and overlaid by the p.m.f. of Poisson distribution using function "poisspdf".

The inverse transform method is also used to complete the simulation. Uniformly random variables are generated through "r=rand" and the initial Poisson probability is counted using "p=exp(-120)".A "while" loop is used to conduct the comparison between rand variable and cumulative probability. The "for" loop is used to repeat the trial N times and store the results into zero array. The histogram is generated and compared with the figure in first method.

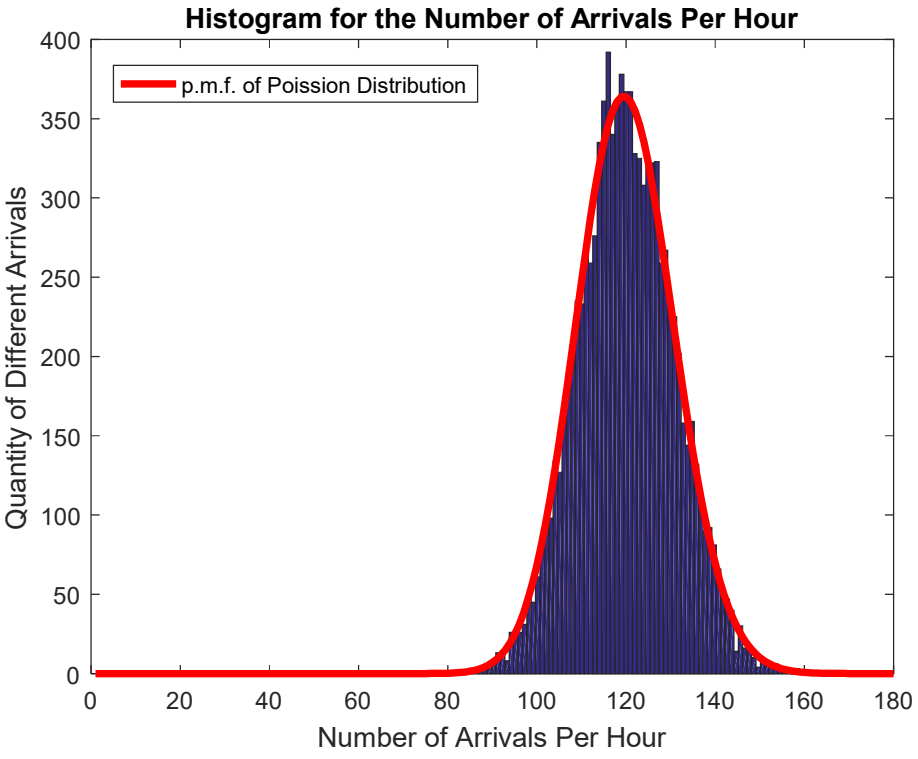
In additional, the function "poissinv" is used in the final part to generate the figure. The comparison is conducted through three histograms to prove the correctness of simulations.

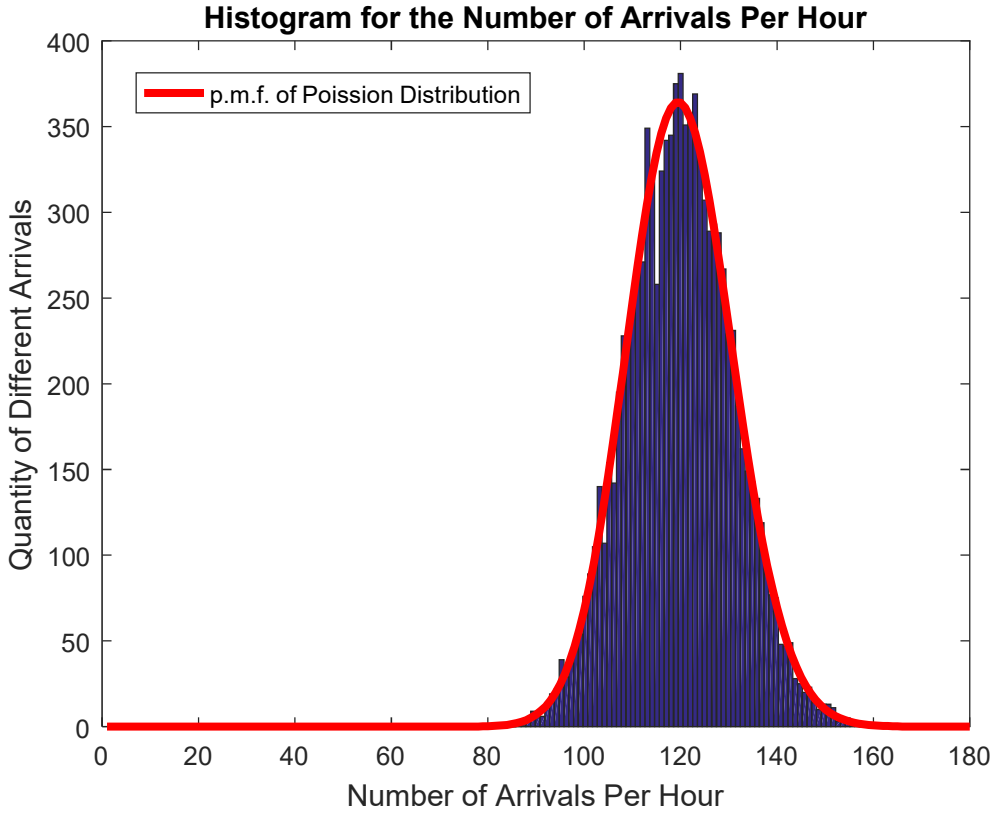
Code:

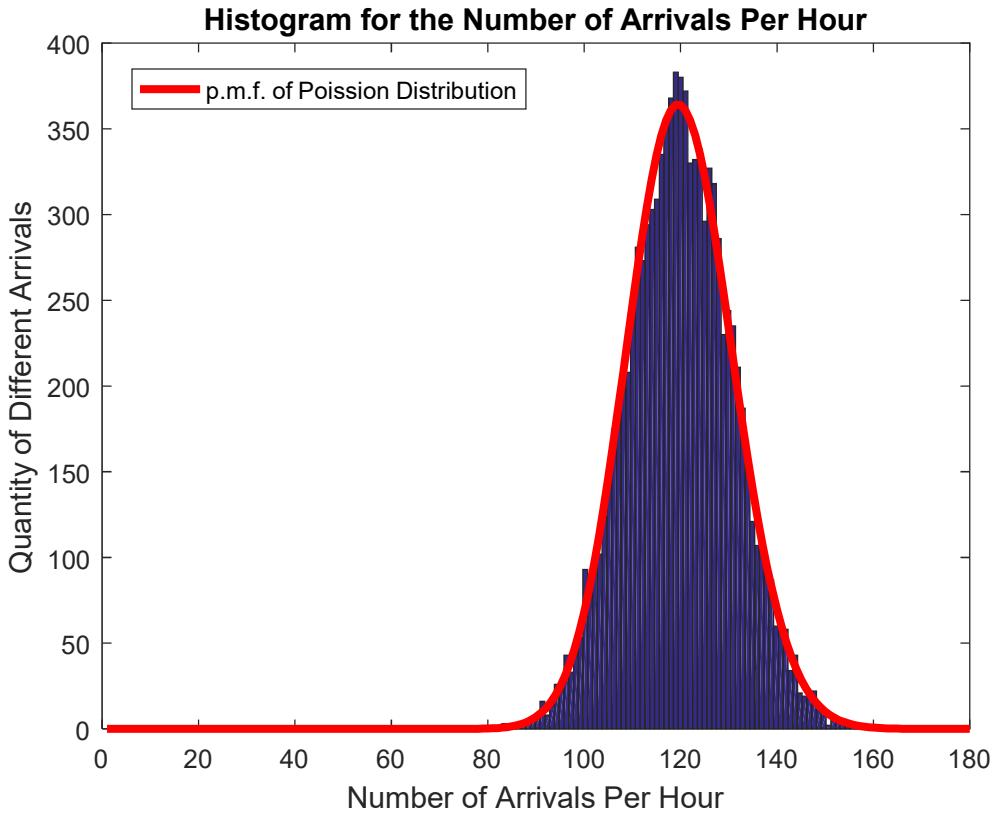
```
function f=poiss(N)
lamada=120;
N_interv=100000;
bernoulliTrials=0;
p=lamada/N_interv;
arr=[];
for j=1:N
bernoulliTrials=0;
for i=1:N_interv;
    if rand < p
        bernoulliTrials=bernoulliTrials+1;
    end
end
arr(j)=bernoulliTrials;
end
figure(1);
MAX=max(arr);
```

```
MIN=min(arr)
hist(arr,(MAX-MIN+1));
hold on
x=1:180;
y=poisspdf(x,120)*N;
T1=plot(x,y,'-r','LineWidth', 3);
hold off
title('Histogram for the Number of Arrivals Per Hour');
xlabel('Number of Arrivals Per Hour');
ylabel('Quantity of Different Arrivals');
legend([T1],'p.m.f. of Poission Distribution');
arr_i=zeros(1,N);
for t=1:N
    r=rand;
    k=0;
    P=exp(-120);
    f=P;
    while r>=f
        P=P*(120)/(k+1);
        f=f+P;
        k=k+1;
    end
        arr_i(t)=k;
end
MAXa=max(arr_i);
MINa=min(arr_i);
figure(2);
hist(arr_i,(MAXa-MINa+1));
hold on
x_a=1:180;
y_a=poisspdf(x,120)*N;
T2=plot(x_a,y_a,'-r','LineWidth', 3);
hold off
title('Histogram for the Number of Arrivals Per Hour');
xlabel('Number of Arrivals Per Hour');
ylabel('Quantity of Different Arrivals');
legend([T2],'p.m.f. of Poission Distribution');
arr_Sample=zeros(1,N);
for z=1:N
    a=rand;
    b=poissinv(a,120);
    arr_Sample(z)=b;
end
figure(3);
MAXb=max(arr_Sample);
MINb=min(arr_Sample)
hist(arr_Sample,(MAXb-MINb+1));
hold on
x_b=1:180;
y_b=poisspdf(x,120)*N;
T3=plot(x_b,y_b,'-r','LineWidth', 3);
hold off
title('Histogram for the Number of Arrivals Per Hour');
xlabel('Number of Arrivals Per Hour');
ylabel('Quantity of Different Arrivals');
legend([T3],'p.m.f. of Poission Distribution');
```

Simulation Result:







Finding:

The second experiment sets the sample size as 1000. The interval of 1 hour is divided into 10000 subintervals to conduct the Bernoulli trials. The simulation results shows nearly 250 times that 120 cars would arrive per hour, nearly 50 times that 100 cars would arrive per hour, nearly 60 times that 140 cars would arrive per hour, and nearly 0 times that 80 cars arrive per hour or 180 cars arrive per hour.

The p.m.f. curve of Poisson distribution reflects the theoretical distribution of random variable $X \sim \text{Poisson}(\lambda)$, λ=120. The peak of p.m.f. curve is nearly equal to 250 and obtained when the value of x-axis is 120. This matches the distribution of experimental samples. In addition, for Poisson distribution, E[X]= λ and Var[X]= λ. As is shown in both figures, the maximum value of histogram is nearly 250 which is the value of λ, and Other values are successively decreasing, and are distributed on both sides of the highest value.

Experiment No.3

Question:

Define the random variable $N = \min\{n: \sum_{i=1}^{n} X_i > 4\}$ as the smallest number of uniform random samples whose sum is greater than four. Generate a histogram using 100, 1000 and 10000 samples for N. Comment on E[N].

Algorithm:

This experiment is required to calculate the value of N

$$N = \min\{n: \sum_{i=1}^{n} X_i > 4\}$$

X_i is random sample, so its range is 0 to 1.

Thus, choosing different sample size to calculate the variable N and its average.

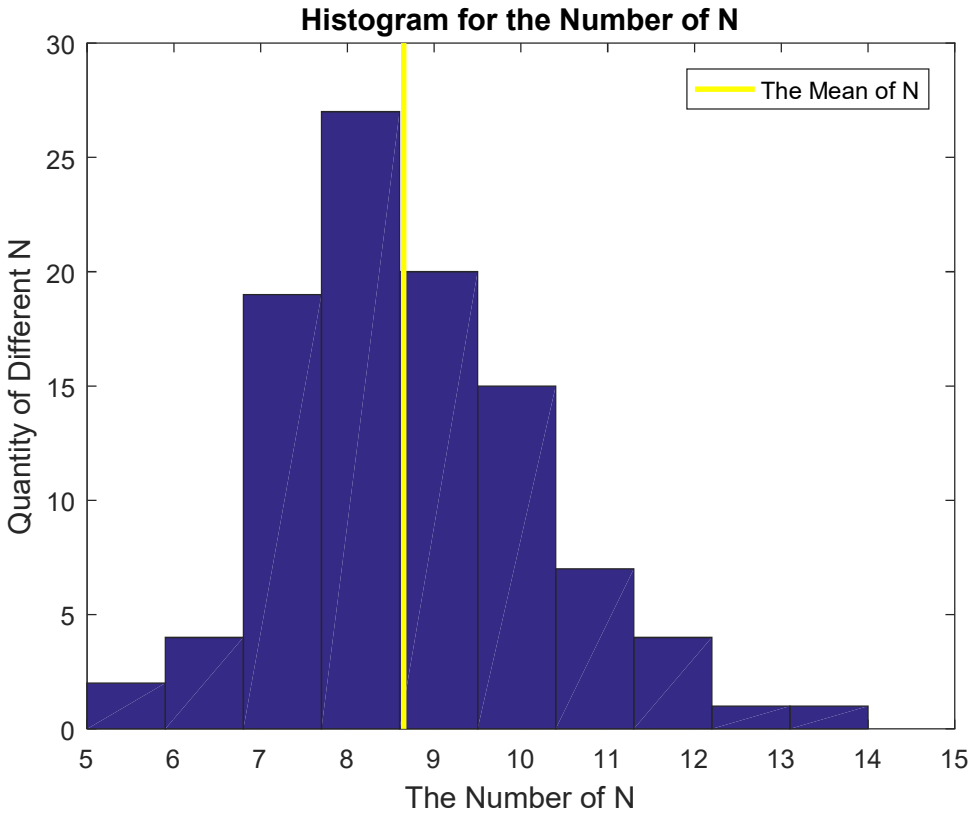
Description of method:

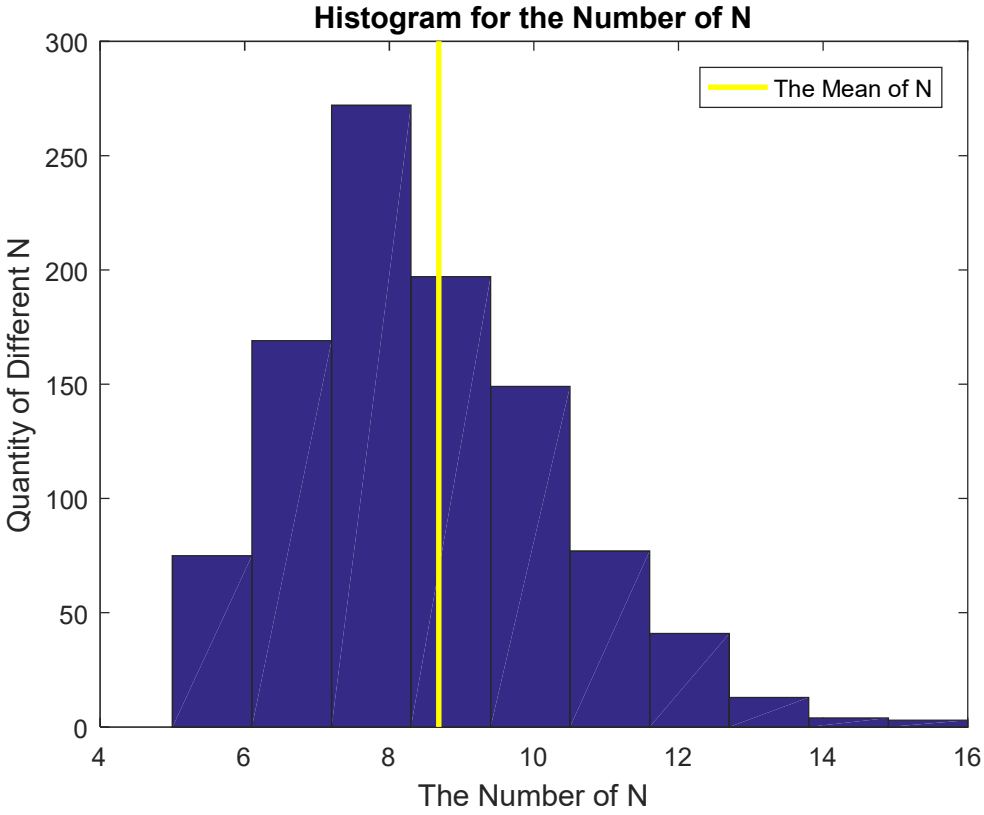
In the third experiment, a "while" loop is used to simulate the process of adding multiple uniformly random variable and terminate the loop if the sum is larger than 4. The "for" loop is used to repeat the trial N times and store the result of each trial into a zero array. The histograms are generated and average is calculated under different sample size.

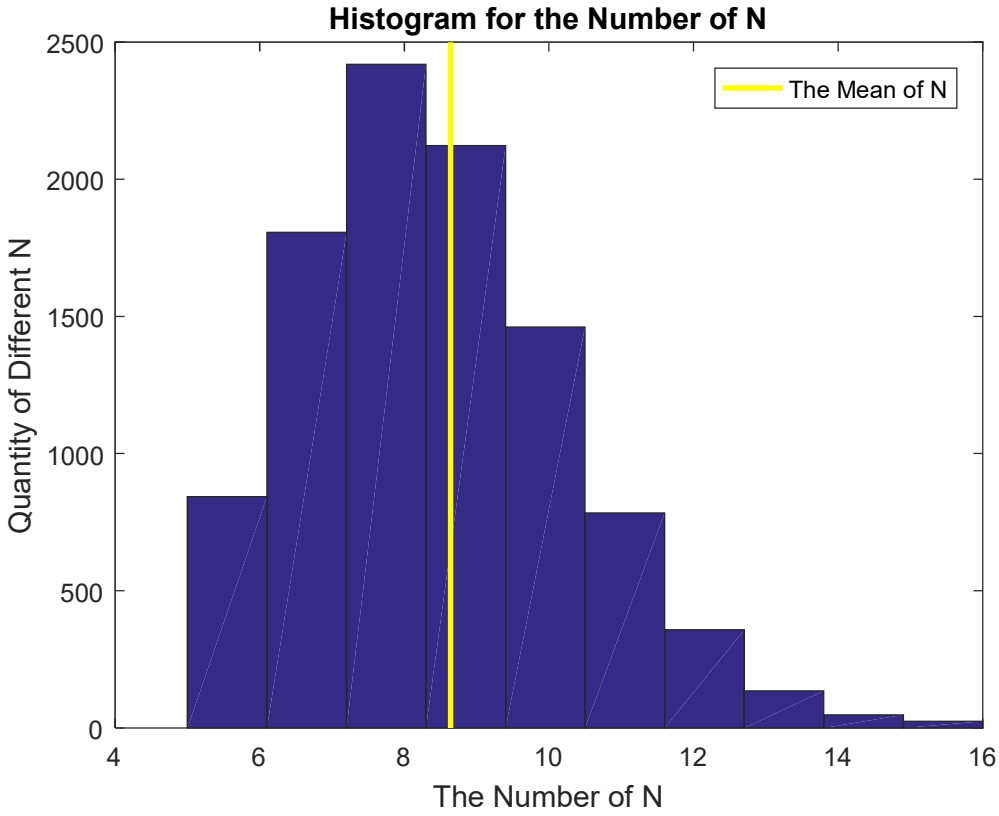
Code:

```
function f=Count_Sample(N)
arr_Sample=zeros(1,N);
for j=1:N
Sum_Smaple=0;
Count=0;
while 1
    Sum_Smaple=Sum_Smaple+rand;
    Count=Count+1;
    if Sum_Smaple > 4
        break
    end
end
arr_Sample(j)=Count;
end
hist(arr_Sample);
Mu_N=mean(arr_Sample);
hold on
plot([Mu_N,Mu_N],ylim,'y-','LineWidth',2);
hold off
disp(Mu_N);
```

Simulation Result:


Histogram for the Number of N
The Mean of N
Quantity of Different N
The Number of N
0
5
10
15
20
25
30
5
6
7
8
9
10
11
12
13
14
15


Histogram for the Number of N
The Mean of N
Quantity of Different N
The Number of N
0
50
100
150
200
250
300
4
6
8
10
12
14
16




```
Command Window
>> Count_Sample(100)
    8.6500
```

```
Command Window
>> Count_Sample(1000)
    8.6820
```

```
Command Window
>> Count_Sample(10000)
    8.6768
```

Finding:

The simulation result shows that the average of N is equal to 8.6500 when the sample size is 100.

The average of N is equal to 8.6820 when the sample size is 1000. The average of N is equal to 8.6768 when the sample size is 10000. Thus, E[N] is larger than 8 and nearly equal to 8.6.

Experiment No.4

Question:

Produce a sequence $\{X_k\}$ where $p_j = \frac{p}{j}$ for j $= 1,2,\cdots,60$ where p is a constant for you to determine. (This is equivalent to spinning the minute hand on a clock and observing the stopping position if $\mathrm{p}[\text{stop on minute j}] = \frac{p}{j}$). Generate a histogram. Define the random variable $N_j = \min\{k: X_k = j\}$. Simulate sampling from N_{60}. Estimate $\mathrm{E}[N_{60}]$ and $\mathrm{Var}[N_{60}]$. Compare you estimates with the theoretical values.

Algorithm:

The samples in the fourth experiments follows the geometric distribution.

$$\mathrm{P(X = k)} = (1-\mathrm{p})^{\mathrm{k}-1}\mathrm{p}$$

The theoretical mean of the distribution is $(1/\mathrm{p})$, and the variance is $(1-\mathrm{p})/\mathrm{p}^2$.

However, each element is the sequence $\{\mathrm{X_k}\}$ can be 1,2,3,…,60 with different probability $\mathrm{p_j} = \mathrm{p}/\mathrm{j}$ for j=1,2,3,…,60. Thus, the normalization constant p is supposed to calculated first.

Because

$$\mathrm{a} * \sum_{\mathrm{j}=1}^{60} \frac{1}{\mathrm{j}} = 1$$

So

$$\mathrm{a} = \frac{1}{\sum_{\mathrm{j}=1}^{60} \frac{1}{\mathrm{j}}}$$

The cumulative distribution function is used to finish sampling under different probability.

1). $\mathrm{U}\sim\mathrm{Uniform}([0,1])$

2). $\mathrm{X_i} = \mathrm{F_X^{-1}}(\mathrm{U_i})$

3). Find smallest $\mathrm{X_i}$, such that $\mathrm{U_i} \le \mathrm{F_X^{-1}}(\mathrm{X_i})$

Description of method：

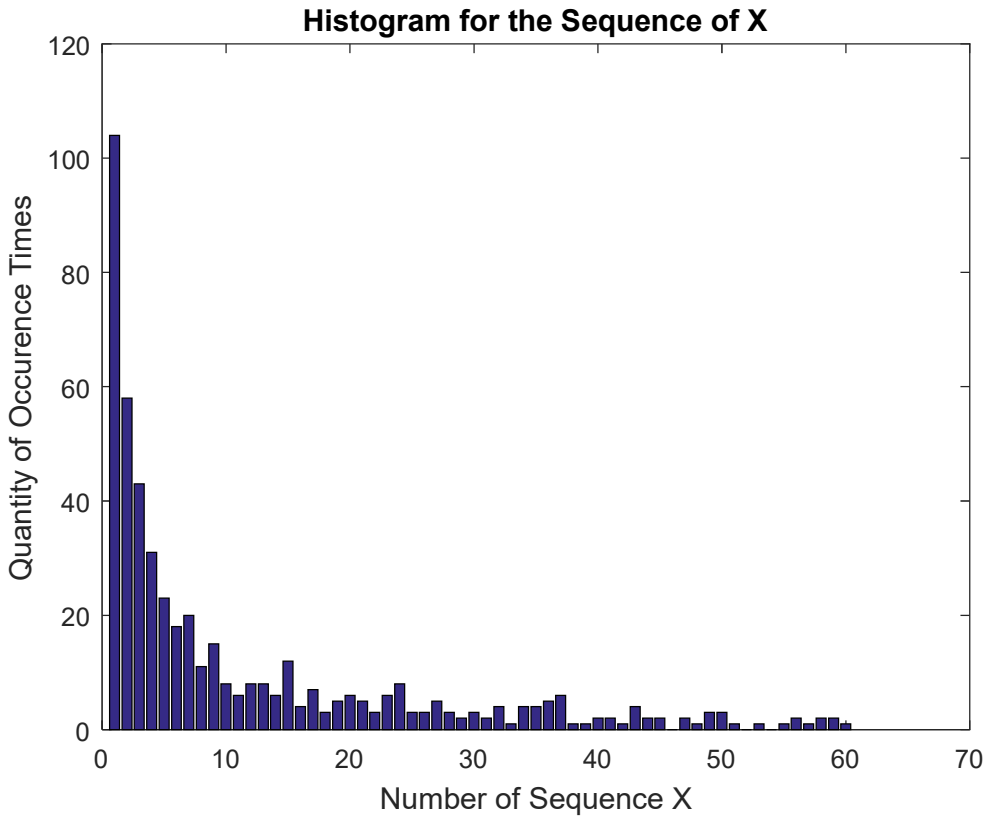
In the four experiment, the constant “a_Value” can be calculated through normalization. The idea of the experiment is to keep generating sample until the first 60 appears. “k” represents the total times of trials that required to generate the 60 for the first time. “X” indicates the sequence of different numbers in the sample until the first 60 occurs. The first two “while” loops are used to create sequence X randomly and judge whether any element of sequence is equal to 60. The third “while” loop is used to count the times that required to generate the 60 for the first time. The “for” loops in the programs in used to repeat the experiment N times and calculate the experimental average and variance of N(60), and compare them with the theoretical mean and variance of geometric distribution.

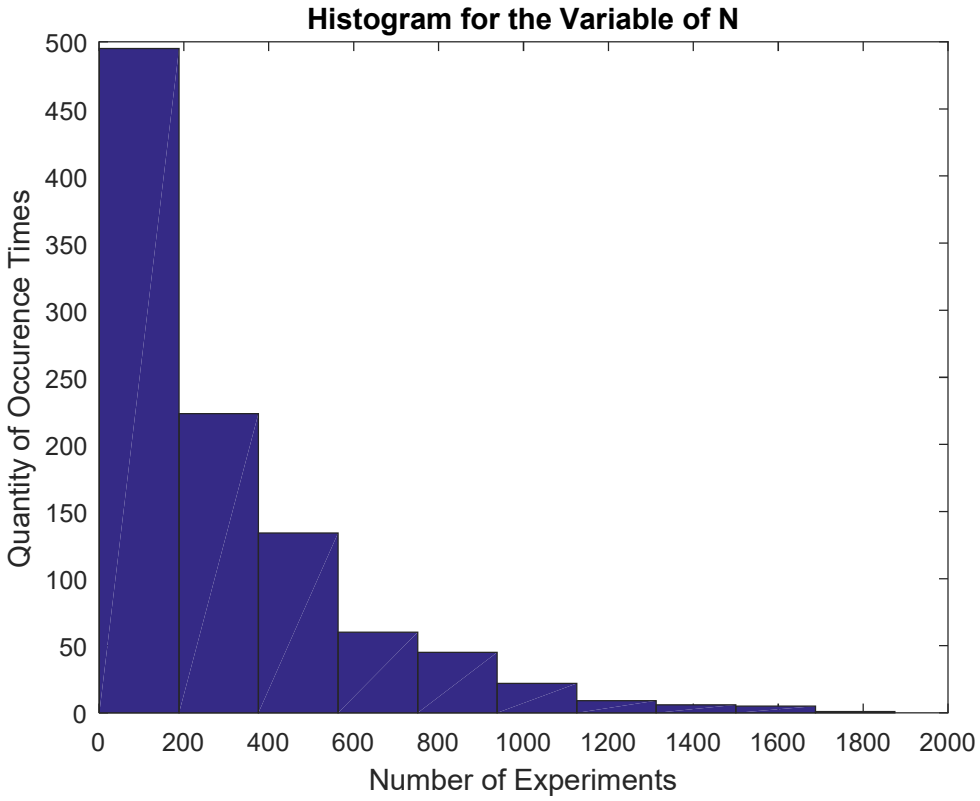
Code:

```
function f=sequ(N)
sum=0;
for i=1:60
    a_a=1/i;
    sum=sum+a_a;
end
a_Value = 1/sum;
count = zeros(1, 60);

stopcond =0;
while stopcond ==0
    data = rand;
    j=1;
    b = a_Value;
    stopcond2 = 0;
    while stopcond2 == 0
        if data < b
            count(j) = count(j) + 1;
            stopcond2 =1;
        else
            j = j+1;
            b=b+0.2137/j;
        end
    end
    if j == 60
        stopcond =1;
    end
end
figure(1);
bar(1:60, count);
title('Histogram for the Sequence of X');
xlabel('Number of Sequence X');
ylabel('Quantity of Occurence Times');

B = zeros(1,60);
arr_Number=zeros(1,N);
p=a_Value/60;
Number=0;
for k=1:N
    B(1) = a_Value;
    stopcond3 = 0;
    while stopcond3 == 0
        Number=Number+1;
        data = rand;
        for j=2:60
            B(j)=B(j-1)+0.2137/j;
            if data < B(j)
                data_label = j;
                break
            end
        end
        if data_label == 60
            stopcond3 = 1;
        end
```

```
    end
    arr_Number(k)=Number;
    Number=0;
end
figure(2);
hist(arr_Number);
title('Histogram for the Variable of N');
xlabel('Number of Experiments');
ylabel('Quantity of Occurence Times');
Test_Mean=mean(arr_Number)
Theo_Mean=1/p
Test_Variance=var(arr_Number)
Theo_Variance=(1-p)/p^2
```

Simulation Result:





Command Window

```
>> sequ(1000)

Test_Mean =

  288.0040


Theo_Mean =

  280.7922


Test_Variance =

   7.8156e+04


Theo_Variance =

   7.8563e+04
```

Finding:

The histogram for the sequence X shows that before the first 60 appears, the quantity of different numbers 1,2,3,…,60 in the sequence. When set the sample size as 1000, the times that number 1 occurs is nearly 105, the times that 2 occurs is nearly 58, the times that 3 occurs is nearly 45, and the occurrence quantity is decreasing from 1 to 60. The simulation result follows the probability model that the P(1)>P(2)>P(3)>P(4)>…>P(60).

The histogram for the Variable of N shows the total times of experiments needed to get the first 60. When set the sample size as 1000, in nearly 500 cases that the total experiment is 0-200 to get the first 60, in nearly 220 cases that the total experiment is 200-400 to get the first 60, and the quantity of cases decreasing as the total number of experiments increase.

It can be seen from the calculation that the test mean is 288.0040 which is close to the theoretical mean 288.7922; and the test variance is 78156 which is close to the theoretical variance 78563 too. The comparison between test and theoretical statistic values prove the simulation.

Experiment No.5

Question:

Use the accept-reject method to sample from the following distribution P_j by sampling from the uniform auxiliary distribution (q_j=0.05 for j=1,…,20):

$$p_1 = p_2 = p_3 = p_4 = p_5 = 0.06, \qquad p_6 = p_9 = 0.15, \qquad p_7 = p_{10} = 0.13, \qquad p_8 = 0.14$$

Generate a histogram and overlay the target distribution P_j. Compute the sample mean and sample variance and compare these values to the theoretical values. Estimate the efficiency of your sampler with the following ratio:

$$\text{Efficiency} = \frac{\#accepted}{\#accepted + \#rejected}$$

Compare your estimate of the efficiency to the theoretical efficiency given your choice for the constant c.

Algorithm:

Because the functional form of distribution P(i) makes it difficult to sampling directly or using inversion methods. Thus, the experiment uses the “Acceptance-rejection methods”.

1). Choose an uniform distribution and the probability mass function q(j)=0.05, for j=1,2,3,4,…,20.

2). Finds the constant c=max(p(i))/0.05= 3 so that p(i)/q(j)<=c all the time.

3). Generate uniformly random variable $\mathrm{u}\sim\mathrm{Uniform}([0,1])$, loop time add one.

4). Select a p(i) randomly.

5). Check if (c*u)<=p(i)/0.05, than accept and return I, reset loop time; otherwise, reject i and go back to step three.

The expression to calculate the expectation and variance are

$$\mathrm{E}[X] = \sum_{k=1}^{N} x * p(x)$$

$$\mathrm{V}[X] = E[(X - E[X])^2]$$

The theoretical efficiency is $1/\mathrm{c}$

The experimental efficiency is $1/\mathrm{mean(loop\ time)}$.

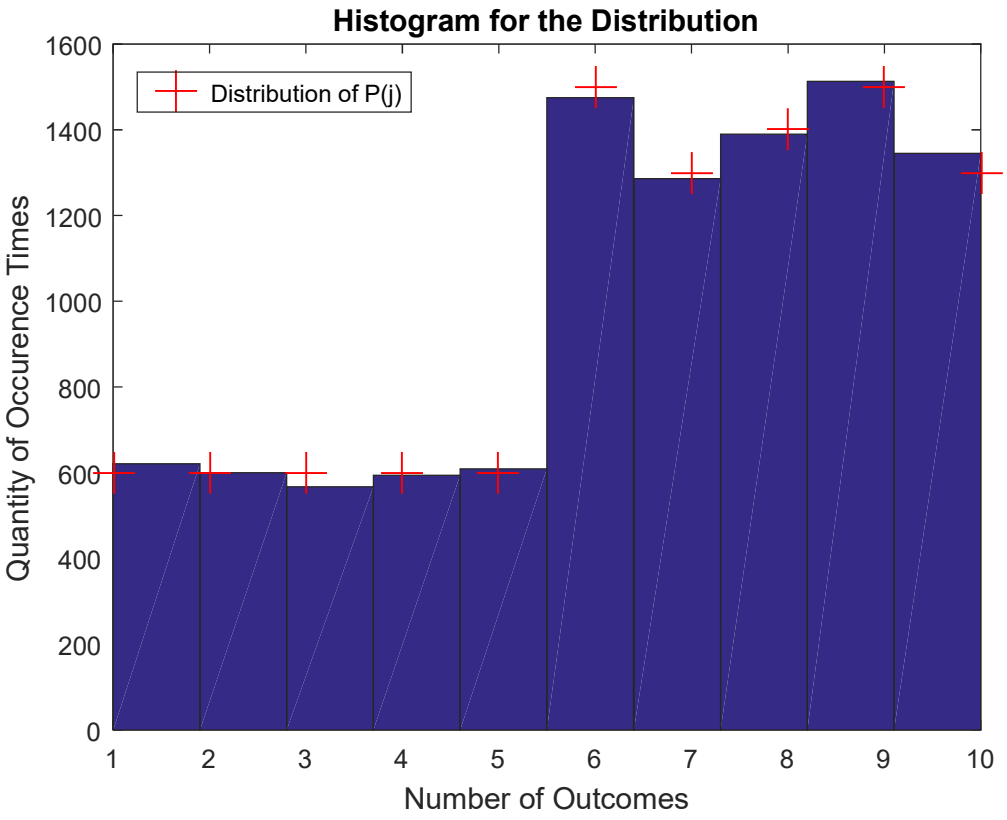
Description of method:

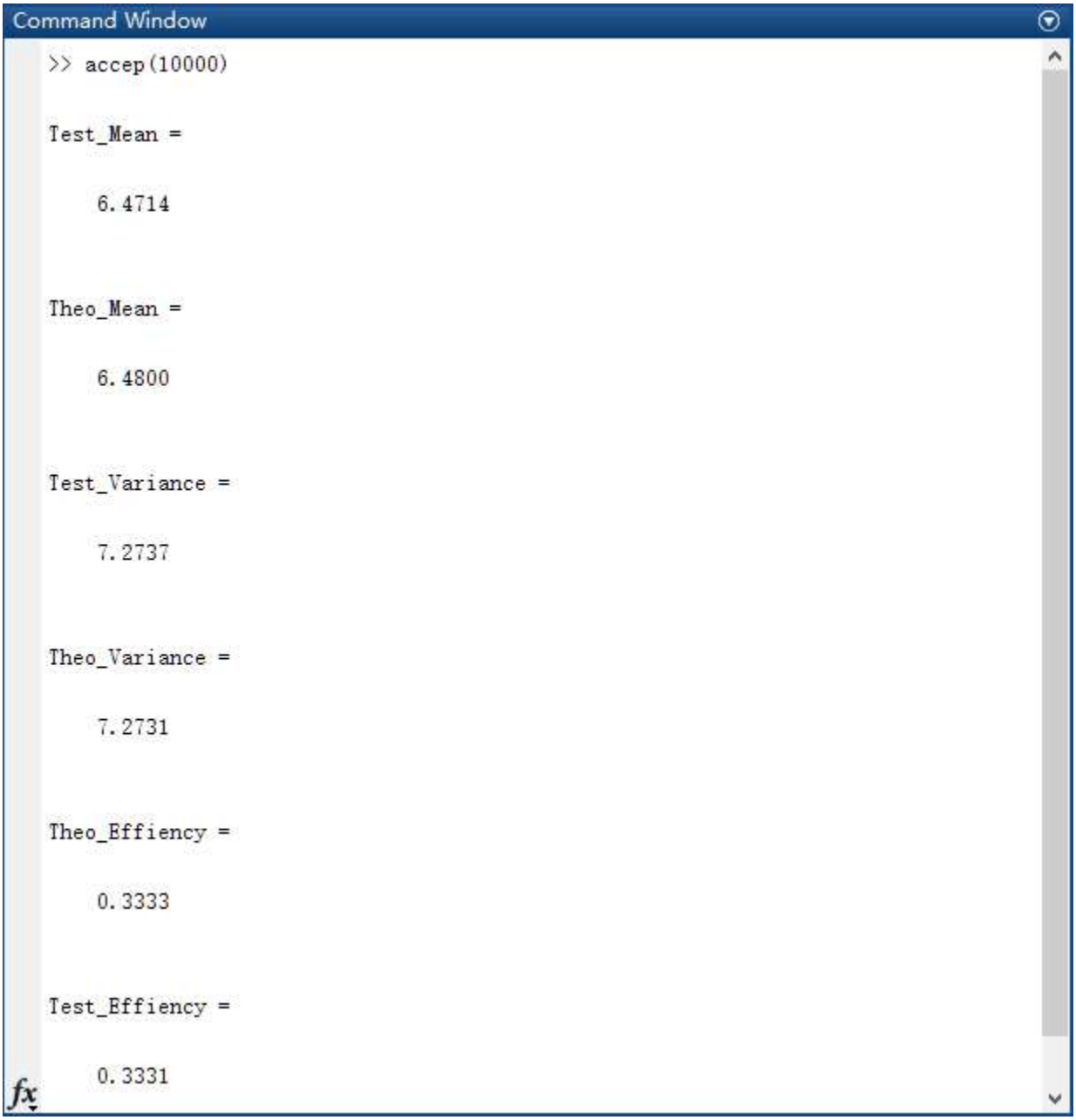
In order to obtain samples follows the distribution of P(i), the experiment simulate selecting randomly from an uniform auxiliary distribution q(j). The "while" loop is used to generate number "j" such as 1,2,3,4,…,10 that accepted in the distribution of p(i), the "for" loop is used to repeat the trial multiple times and record the acceptance number j into an array. The function "mean" and "variance" are used to calculate the test mean and test variance of acceptance numbers. The histogram is generated to show the accepted samples and overlaid with the target distribution p(i).

Code:

```
function f=accep(N)
p = [6 6 6 6 6 15 13 14 15 13]/100;
c=max(p)/0.05;
for i = 1:N, k = 0;
    while 1
        k = k + 1;
        j = 1 + floor(20*rand);% Get Uniform j
        if j<=10
            if 3*rand < p(j)/0.05 % Accept p(j) if U<p(j)/c, q(j)= 0.1
                X(i) = j; C(i) = k;
                break
            end
        end
    end
end
Test_Mean=mean(X)
Theo_Mean=1*0.06+2*0.06+3*0.06+4*0.06+5*0.06+6*0.15+7*0.13+8*0.14+9*0.15+10*0
.13
Test_Variance=var(X)
Theo_Variance=0.0001*sum((X-Theo_Mean).^2)

Mu_C=mean(C);
Theo_Effiency=1/c
Test_Effiency=1/Mu_C

hist(X);
hold on
x=1:10;
T1=plot(x,p*N,'r+','markersize',15);
title('Histogram for the Distribution');
xlabel('Number of Outcomes');
ylabel('Quantity of Occurence Times');
legend([T1],'Distribution of P(j)');
hold off
```

Simulation Result:





Finding:

The histogram the amount of different acceptance numbers ranging from 1 to 10. When set the sample size as 10000, the amount of 1, 2, 3, 4, 5 are all nearly equal to 600, the amount of 6 and 9 are both nearly equal to 1500, the amount of 7 and 10 are both nearly equal to 1300. The amount of different number follows the probability distribution function of p(i), in which p(1)=p(2)=p(3)=p(4)=p(5); p(6)=p(9); and p(7)=p(10).

It is shown from the simulation result that the test mean of acceptance number is 6.4714, which is close to the theoretical mean 6.4800. The test variance is 7.2737, which is close to the theoretical variance 7.2731. The test efficiency is 0.3331, which is close to the theoretical efficiency 0.3333.

Conclusion

The first experiment simulate sampling from the hypergeometric distribution. The second trial simulate sampling from the Poisson distribution using Bernoulli experiment in multiples subintervals and inverse transform method. The fourth experiment simulate sampling from geometric distribution. The fifth trial uses accept-reject method to sample for the distribution. In conclusion, the five experiments delve into different probability distribution functions and calculate the use statistic method to calculate mean or variance to obtain the different property of distributions.