MATLAB Project

ECE 131A Probability and Statistics

University of California, Los Angeles

Kimber King

Professor: Vwani Roychowdhury

Table of Contents

Introduction

In this project we will further analyze random variables and their various properties. We will also investigate how are random variables used to model practical systems. Each experiment done had a combination of MATLAB programming, mathematical analysis and technical writing. The three experiments conducted include a simulation of a game of cards, a simulation of a binary transmission system, and a simulation of the central limit theorem.

Experiment 1

Introduction:

The purpose of this experiment is to conduct a simulation of a game of cards. In this scenario, a deck of 52 cards is shuffled to create random permutations of 52 cards. The deck is shuffled until the desired pattern of cards has occurred while recording the number of shuffles it takes. The exact number of shuffles it takes to create the desired pattern is a random variable that is denoted as X.

Definitions:

X: random variable denoting the number of shuffles it takes to create a desired pattern.
m: number of repetitions of this experiment.

Values of X are denoted as $\{x_i : I \in \{1, 2, \dots, m\}$

(a) Verification of randomness: 1st order and 2nd order Repeat the shuffling for 100, 000 times. Record the shuffling results in a 100000 ×52 sized matrix.
   a. Check whether at each position, each card is equally likely to show up. Specifically, within each column, count the number of occurrences of each of the 52 cards, and see whether the numbers are evenly distributed. This should result in 52 sets of 52 count numbers
   b. Similarly, check whether at the following pairs of positions, each of 52 × 51 card pairs is equally likely to show up. (5, 20) (45, 51) (2, 32) In this case, for each pair of positions, you will have 52 × 51 = 2652 count numbers. Please plot one histogram with 50 bins of the count numbers for each pair of positions.

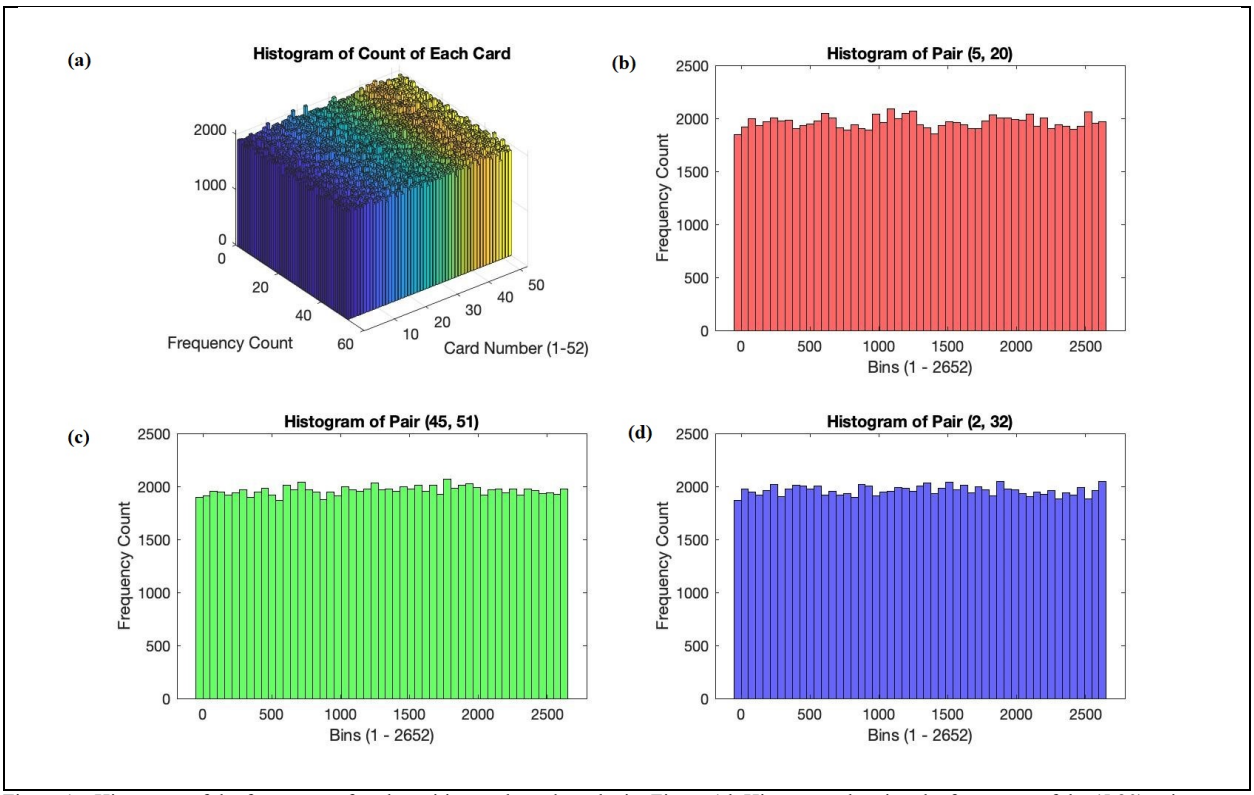


Figure 1.a Histogram of the frequency of each position each card can be in. Figure 1.b Histogram showing the frequency of the (5,20) pair showing up. Figure 1.c Histogram showing the frequency of the (45,51) pair showing up. Figure 1.d Histogram showing the frequency of the (2,32) pair showing up.

Given independent trials for positions (5, 20), (45, 51), and (2, 32), all 52 x 51 = 2652 pairs of cards appeared to be approximately equally likely to appear. This experimental result is what we have expected as one out assume that all cards are equally likely to appear in any given position.

(b) Let m = 100, 000; that is, repeat the experiment 1 for 100, 000 times. Record the x_i 's.

a. Plot the distribution of the xi 's you get. Specifically, calculate the vector N, so that N_n is the number of x_i 's that are equal to n. n ∈ {1, 2, . . . , max $(\{x_i\}_{i=1}^{m})$}.

For example, if we have 5 x_i 's, namely,

$$X_1 = 2, x_2 = 1, x_3 = 4, x_4 = 1, x_5 = 4$$

then

$$N = [N_1, N_2, N_3, N_4] \text{ and } N_1 = 2, N_2 = 1, N_3 = 0, N_4 = 2$$

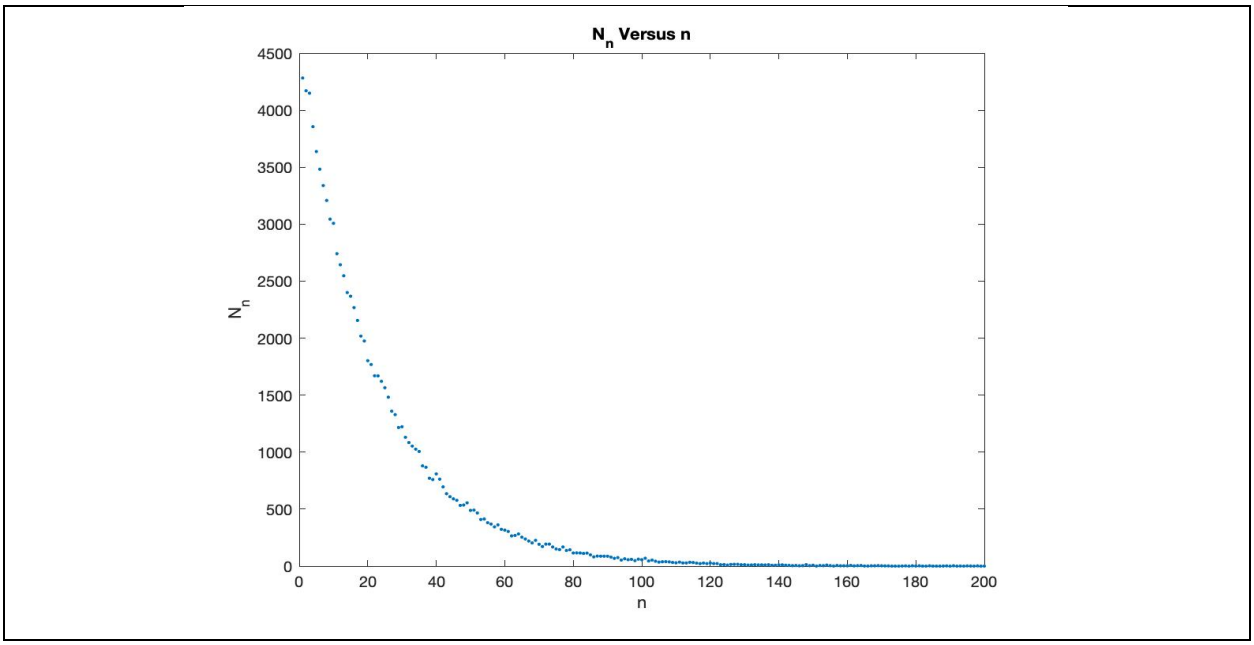


Figure 2: The scatter plot of N_n versus n.

For this experiment, we performed 100,000 times and when the first king and ace pair appear, we recorded it as x_i, where i is the number of permutations. The number of x_i that is equal to some number n were recorded in N_n. A geometric distribution was observed (figure 2) which was what we have expected.

b. Show that X follows a geometric distribution. What is the relation between its parameter and p?

To show that X follows a geometric distribution, the data was scaled by $\frac{1}{total\ number\ of\ trials}$.

Given that we have n-1 unsuccessful attempts followed by 1 successful attempt to succeed in n attempts, geometric distribution is given to be:

$$p_k = p(1-p)^{k-1}$$

To solve for p at k = 1:

$$0.0433 = (1-p)^0 * p = 0.0433.$$

X follows a geometric distribution so the probability $P[X = k|p]$ equals $(1-p)^{k-1} * p$. If $\{x_i\}_{i=1}^m$ are taken as samples of m random variables $\{X_i\}_{i=1}^m$, the probability of taking m independent samples and getting any set of X_is is

$$P[X_1 = x_1, \dots, X_m = x_m | \, p] = (1-p)^{x_1-1} * p * (1-p)^{x_2-1} * p * \dots * (1-p)^{x_m-1} * p \quad (1)$$

$$= p^m (1-p)^{\sum_{i=1}^m x_i - m}, \quad (2)$$

Let (2) be referred to as L(p), in order to find p that maximizes L(p):

$$\ln\big(L(p)\big) = m * \ln(p) * \ln(1-p) * \sum_{i=1}^m x_i - m \quad (3)$$

Take the derivative of (3)

$$\frac{\partial \ln\,(L(p))}{\partial p} = \frac{m}{p} - \sum_{i=1}^m \frac{x_i - m}{1-p} = 0 \quad (4)$$

$$p = \frac{m}{\sum_{i=1}^m x_i} = \hat{p} \quad (5)$$

This $\hat{p}$ is known as the max-likelihood estimator of p and is considered the ratio of valid permutations amongst the entire m permutations. This value is further confirmed in parts (d) and (e) of this experiment.

(c) Assume that X follows a geometric distribution with parameter p.

a. What is the probability P(X = k | p)?

$$P(X = k \mid p) = p(1-p)^{k-1}, \quad k = 1, 2, 3, \dots$$

b. We can also see $\{x_i\}_{i=1}^m$ as samples of m random variables $\{X_i\}_{i=1}^m$ respectively, where $X_i \sim^{idd}$ Geometric(p).
What is the probability that one takes m independent samples of X and gets the x_i 's you observed?

$$P(X_1 = x_1, \dots, X_m = x_m / p) \quad (1)$$

$$\prod_{i=1}^m P(X_i = x_i \, / p) \quad (2)$$

$$= \prod_{i=1}^m p(1-p)^{x_i - 1} \quad (3)$$

Thus,

$$p^m (1-p)^{\sum_{i=1}^m (x_i - 1)} \quad (4)$$

c. What is the value of p that can maximize $P(X_1 = x_1, \dots, X_m = m_m \mid p)$? Show that this value is $\hat{p} = \frac{1}{\bar{x}} = \frac{m}{\sum_{i=1}^m x_i}, where\ \bar{\bar{x}} \equiv \frac{\sum_{i=1}^m x_i}{m}$

$$\log\,(P(X_1 = x_1, \dots, X_m = x_m \, / p)) \quad (1)$$

$$= \log\left(p^m(1-p)^{\sum_{i=1}^{m}(x_i-1)}\right) \qquad (2)$$

$$= m * log(p) + \sum_{i=1}^{m}(x_i - 1)\log(1-p) \quad (3)$$

Now,

$$\frac{d}{dp}(\log(P(X_1 = x_1, \dots, X_m = x_m\ /p)) = 0 \quad (4)$$

$$= \frac{d}{dp}(m * log(p) + \sum_{i=1}^{m}(x_i - 1)\log(1-p)) = 0 \quad (5)$$

$$= \frac{m}{p} - \frac{\sum_{i=1}^{m}(x_i-1)}{1-p} = 0 \qquad (6)$$

$$= \frac{m}{p} - \frac{\sum_{i=1}^{m} x_i - \sum_{i=1}^{m} 1}{1-p} = 0 \quad (7)$$

$$= \frac{m}{p} - \frac{m\bar{x}-m}{1-p} \qquad (8)$$

$$m - mp\bar{x} = 0 \quad (9)$$

$$p = \frac{1}{\bar{x}} \quad (10)$$

So,

$$\log\big(P(X_1 = x_1, \dots, X_m = x_m\ /\ p)\big)\ is\ \max\ at\ \hat{p} = \frac{1}{\bar{x}} \quad (11)$$

(d) $\hat{p}$ is called the max-likelihood estimator of p.

a. Calculate $\hat{p}$ for m = 100, 000

$\hat{p}$ = 0.044 was calculated. See code for further elaboration.

b. Compare the variance of $\hat{p}$ for different m:
Calculate $\hat{p}$ for 100 times for m = 30, 100, 300, 1000, 3000, respectively. (1) Report the sample variance of the ˆp's for each m. (2) Use the boxplot function to plot a boxplot where each m value corresponds to a box.

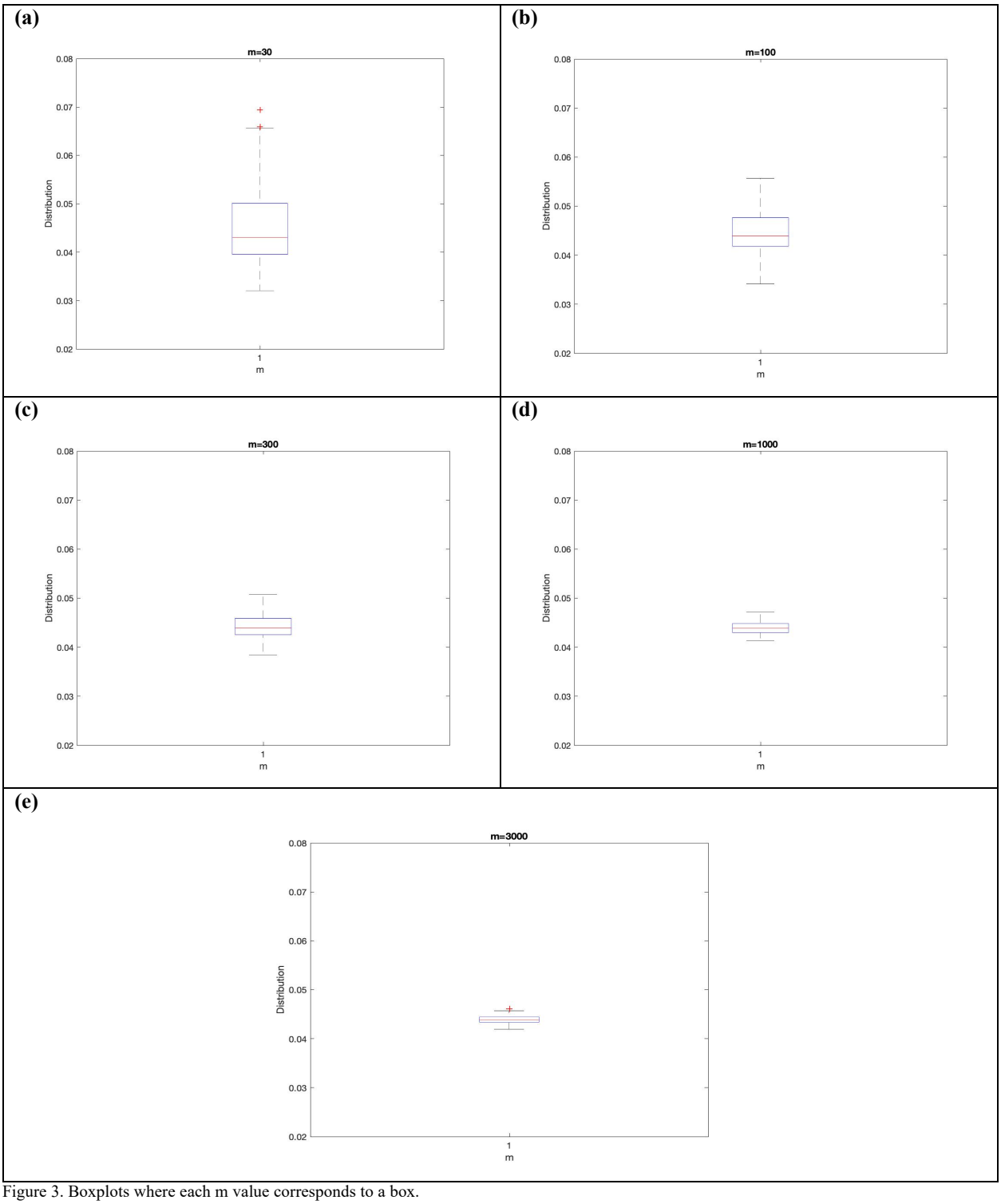


Figure 3. Boxplots where each m value corresponds to a box.

From Figure 3, we can see that boxes gets narrower (shorter) as m increases. Y-axis is adjusted for clearer illustration of this effect. This result makes sense as you would expect as m increases, the "distribution" would decrease.

(e) One can also make use of the empirical distribution N to estimate p.
Let m = 100, 000, use polyfit to fit a line for log (N) vs n. Use the slope to estimate p.

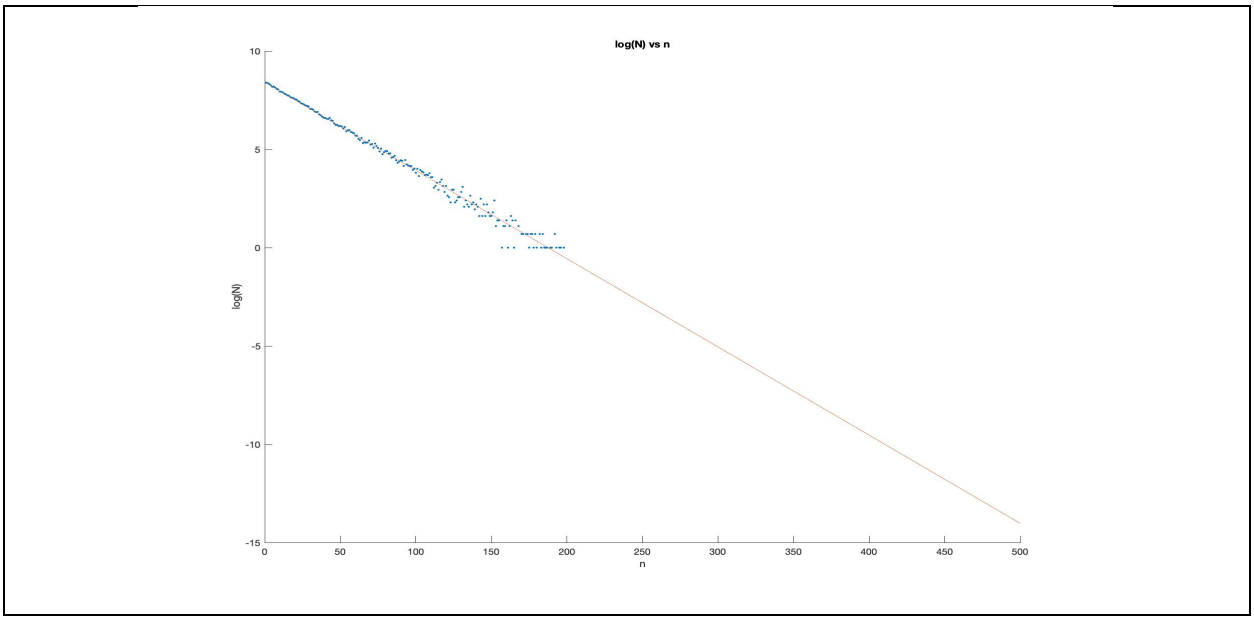


Figure 4. Graph of using empirical distribution N to estimate p. The slope of this graph is what will be used to estimate p.

Further analyzing this graph using MATLAB, a slope of -0.0438 was obtained for this line. This value is the -p value of other sections of this experiment.

(f) One may find that it is quite intuitive to first estimate E[X] using sample mean $\bar{X} = \frac{\sum_{i=1}^{m} x_i}{m}$ then use the relation E[X] $= \frac{1}{p}$ to estimate p. Is it also valid if we use $\overline{(\frac{1}{X})} = \frac{\sum_{i=1}^{m} \frac{1}{x_i}}{m}$ to estimate p?

a.

$$Given\ E(x) = \frac{1}{p} \qquad (1)$$

$$E(\bar{X}) = E(\frac{1}{n}\sum_1^n X_i) \qquad (2)$$

$$= \frac{1}{n}\sum_1^n \frac{1}{p} = \frac{1}{p} = E(x) \qquad (3)$$

$$E\left(\overline{\frac{1}{x}}\right) = E(\frac{1}{n}\sum_1^n \frac{1}{X_i}) \qquad (4)$$

Therefore,

$$\frac{1}{n}\sum_1^n E\left(\frac{1}{X_i}\right) = \frac{1}{n}\sum_1^n E\left(\frac{1}{X}\right) = E(\frac{1}{X}) \quad (5)$$

b.

Using Markov's inequality, which states that: $P[x \geq a] \leq \frac{E[x]}{a}$. We substitute in $\frac{1}{X}\ for\ x$:

$$P[\frac{1}{X} \geq a] \leq \frac{E\left[\frac{1}{X}\right]}{a} \ (1)$$

Let $a = 1$:

$$P[\tfrac{1}{X} \geq a] \leq E\left[\frac{1}{X}\right] \qquad (2)$$

$$P[X \leq 1] = E\left[\frac{1}{X}\right] \qquad (3)$$

We know that $P[X \leq 1] = p$, therefore:

$$E\left[\frac{1}{X}\right] \geq p \qquad (4)$$

c.

Given that in a geometric distribution E[X] can be expressed as:

$$E(X) = \sum_{x=1}^{\infty} x(p(1-p)^{x-1}) \qquad (1)$$

So,

$$E\left(\frac{1}{X}\right) = \sum_{x=1}^{\infty} \frac{p(1-p)^{x-1}}{x} \qquad (2)$$

Let

$$p(1-p) = t \qquad (3)$$

$$\int_0^a \sum_{x=1}^{\infty} t^{x-1} dt \qquad (4)$$

$$= \sum_{x=1}^{\infty} \int_0^a t^{x-1} dt \qquad (5)$$

$$= \sum_{x=1}^{\infty} \frac{a^x}{x} \qquad (6)$$

Integrate:

$$\int_0^a \frac{1}{1-t} dt = -\ln(1-a) \quad (7)$$

So using (6) and (7):

$$\sum_{x=1}^{\infty} \frac{a^{x-1}}{x} = \frac{1}{a} \sum_{x=1}^{\infty} \frac{a^x}{x} = \frac{-\ln(1-a)}{a}$$

Let a = 1-p:

$$E\left(\frac{1}{x}\right) = \frac{-p(\ln p)}{1-p} \neq p$$

Experiment 2

Introduction:

The purpose of this experiment is to simulate a binary transmission system. In this experiment, a binary transmission system sends a '0' bit using a -2 voltage signal and a '1' using a +2 voltage signal. The signal received is corrupted by a noise, N, that has a Laplacian distribution with parameter α. It is assumed that '0' and '1' bits are equally likely.

Definitions:

N: variable denoting noise that corrupts the signal that has a Laplacian distribution.

α: parameter of Laplacian distribution.

'0' bit: denoted by a -2 voltage signal.

'1' bit: denoted by a +2 voltage signal.

(a) Write a MATLAB program to generate Laplacian distribution with parameter α = 2, 0.5 respectively. Use the transformation method (refer to Section 4.9 in the textbook) based on t = 10, 000 samples of the unit-interval uniform random variable. Plot the resulting empirical (Laplacian) cdf and pdf in each case.

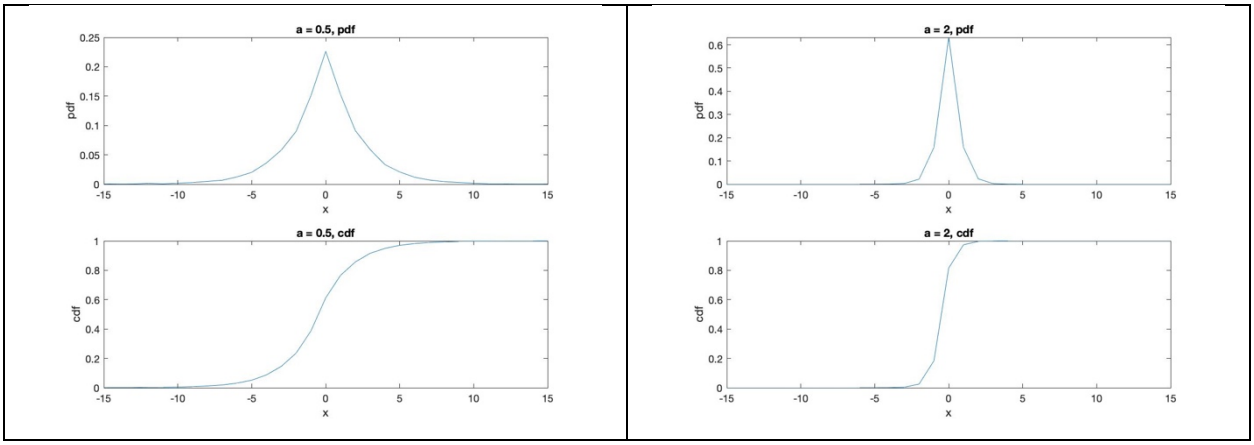


Figure 5. Empirical (Laplacian) cdf and pdf with parameter α = 2, 0.5 respectively.

Figure 5 shows that as alpha increases, the CDF gets a steeper slope and the PDF gets narrower. These results make sense given what we have learned in this course about Laplacian distribution and how changes in alpha affects PDF and CDF.

(b)) Assume the received signal Y is given by Y = X + N. Suppose that the receiver decides a "0" was sent if Y < 0, and a "1" was sent if Y ≥ 0. Write a MATLAB program to simulate the transmission of 10,000 bits under the SNR = 0dB for this channel and compute the empirical error probability

Trial	Empirical Error Probability (%)
1	2.74
2	2.93
3	3.02
4	2.84
5	2.91
6	2.86
7	3.01
8	2.96
9	3.00
10	2.88

```
%Part B
X3=zeros(1,t);
error=0;
for i=1:1:t
    if rand1(i)>0.5
        X3(i)=-log(2-2*rand1(i))/a3;
    else
        X3(i)=log(2*rand1(i))/a3;
    end
    if i<=5000
        if X3(i) <-2
            error=error+1;
        end
    else
        if X3(i) >+2
            error=error+1;
        end
    end
end

fprintf('Emprical Error Probability = %i%%\n', error/t*100);
```

Calculating the mean of the empirical error probability across 10 trials yields a value of 2.91%.

(c) Derive analytically the expression for the error probability under SNR = 0dB. How does the analysis compare to your simulation from part (b)?

$$f_N(n) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|n|}$$

$$f_Y(y \mid X = -2) = f_N(y + 2) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|y+2|}$$

$$f_Y(y \mid X = +2) = f_N(y - 2) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|y-2|}$$

$$f_{error} = \frac{f(error \mid X = -2)}{2} + \frac{f(error \mid X = +2)}{2} = \frac{1}{2} e^{-2\sqrt{2}} = 0.0296 = 2.96\%$$

Simulated value is in an acceptable percent error range (1.68%) of the theoretical value.

Experiment 3

Introduction:

The purpose of this experiment was to simulate and explore the central limit theorem. In this case, X_1, X_2,… are a sequence of iid random variables with a finite mean, μ and finite variance σ^2. The value S_n is the sum of the first n random variables in the sequence. In this simulation, 10,000 samples were used for all parts.

Definitions:

X_1, X_2,…: is a sequence of iid random variables.

μ: finite mean of the sequence.

σ^2: finite variance of the sequence.

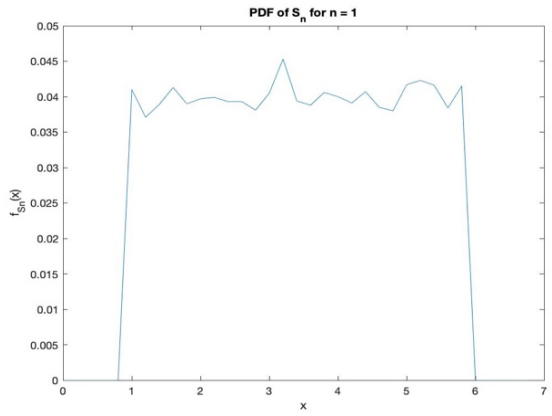
S_n is the sum of the first n random variables in the sequence defined as:

$$S_n = X_1 + X_2 + \ldots + X_n$$

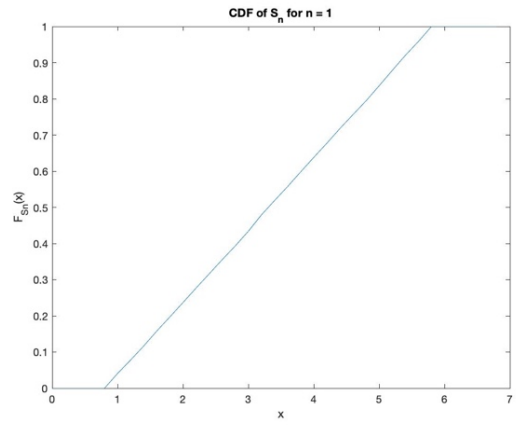
t: the 10,000 samples taken during the experiment.

(a) Let X_i be a continuous uniform random variable taking values in the interval (1, 6). Write a MATLAB program to plot the empirical pdf and cdf of S_n. Consider n = 1, 5, 15, 50 and compare your results.

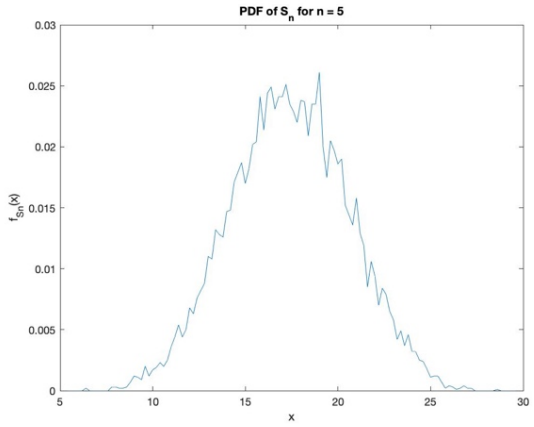
(a)



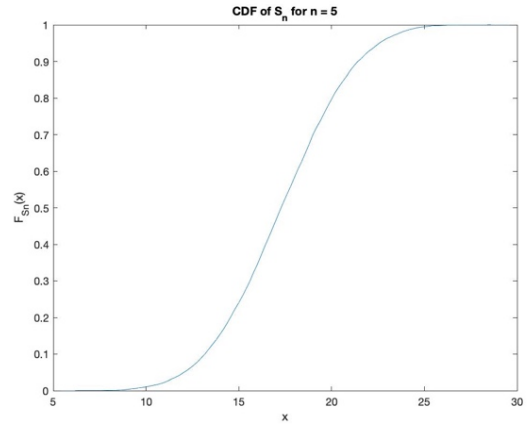
(b)



(c)



(d)



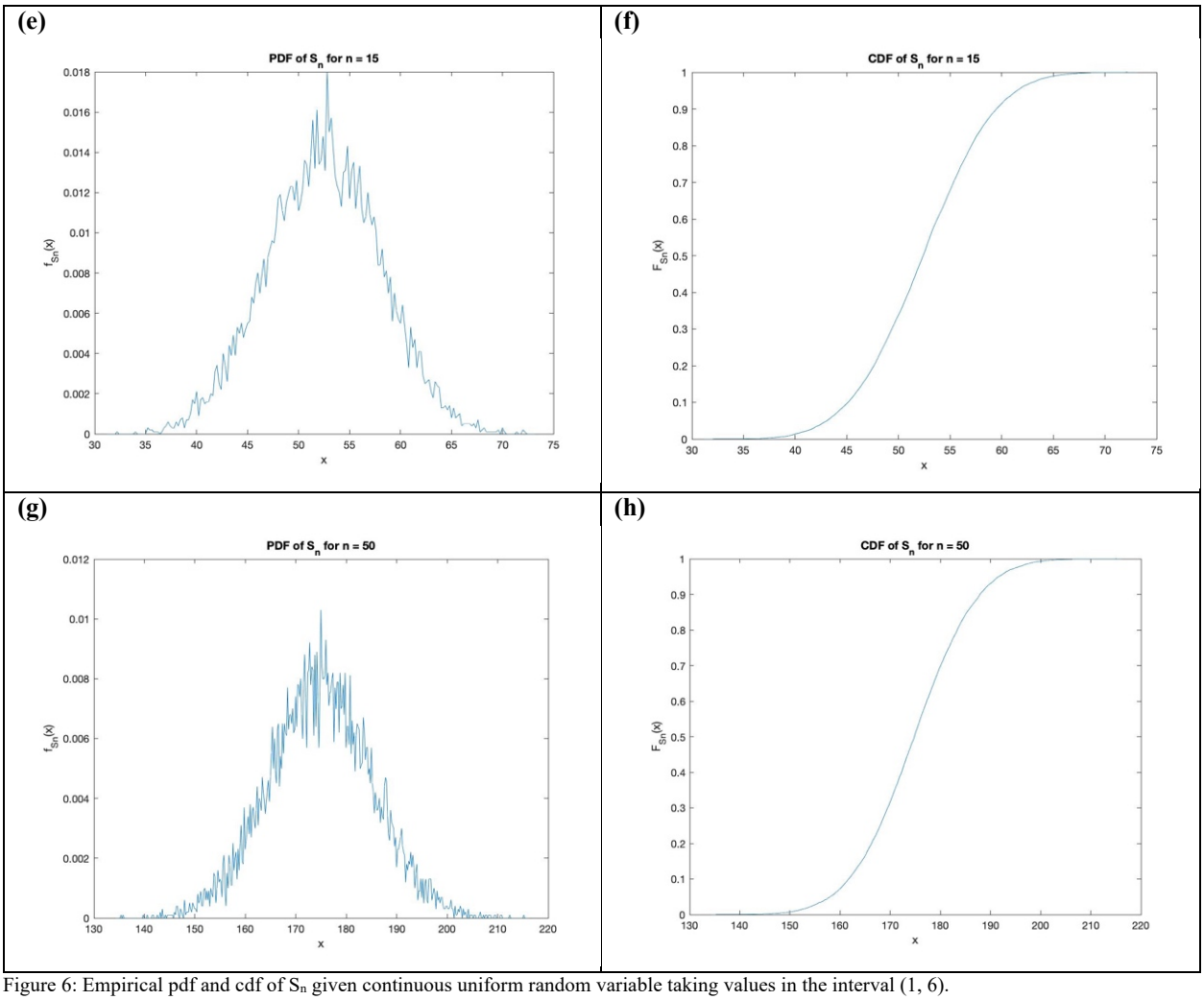


Figure 6: Empirical pdf and cdf of S_n given continuous uniform random variable taking values in the interval (1, 6).

From this data was are able to see that as n increases the PDF becomes for normal and CDF gets steeper. This step is further explained in parts c. and d. in this experiment.

(b) Calculate analytically the mean and the variance of X_i and of S_n in part (a)

Mean of X_i:

$$\frac{\sum_{i=1}^{N} X_i}{N}$$

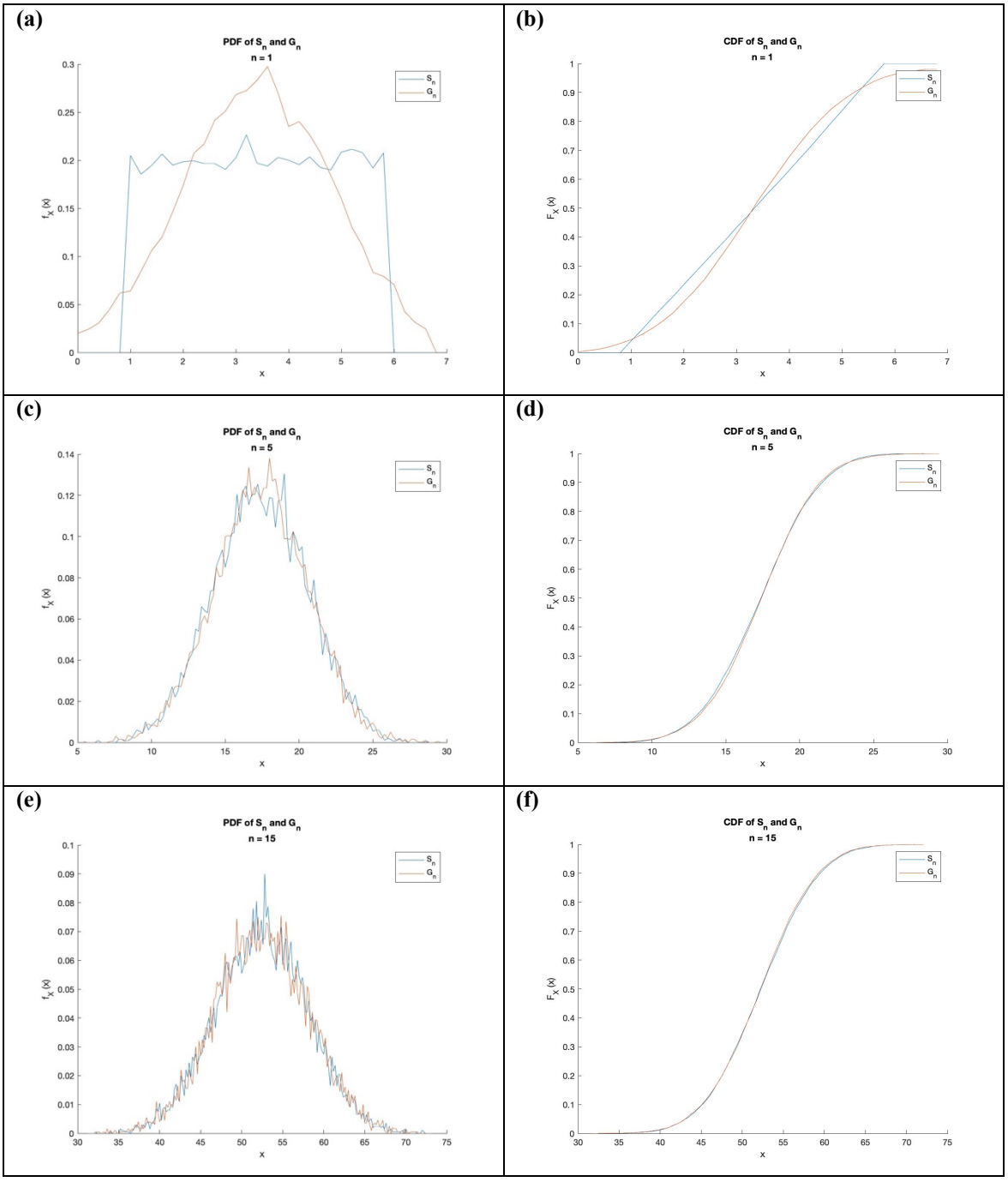
Variance of X_i:

$$E((X-\mu)^2)$$

N	Mean	Variance
1	3.50	2.08
5	17.50	10.40
15	52.50	30.84
50	175.00	103.90

For random variables S_n, the mean is n times mean of each X_i, variance is n times variance of each variance.

(c) Write a MATLAB program to draw the pdf and cdf curves of the Gaussian distribution with the same mean and variance as S_n. Superimpose this plot with the plots from part a.



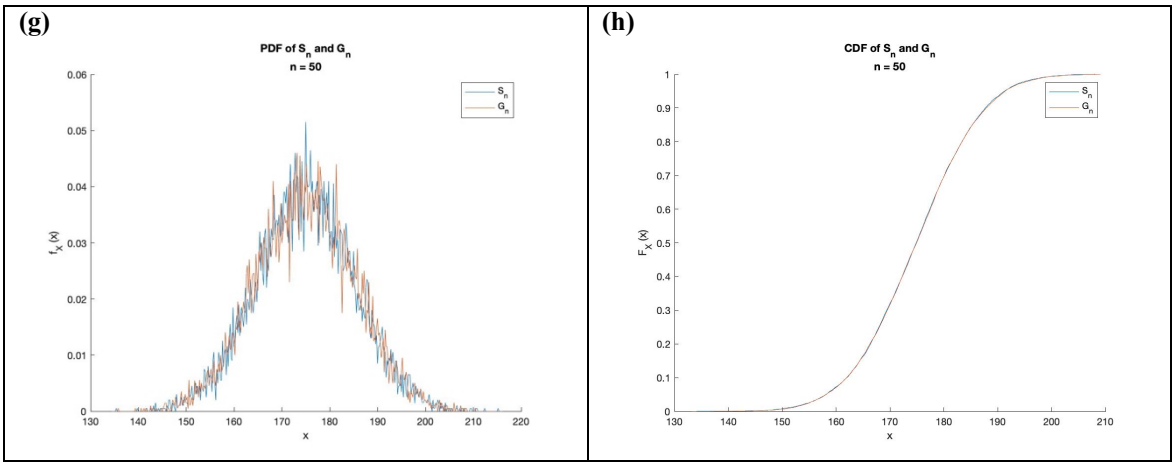


Figure 7: Pdf and cdf curves of the Gaussian distribution with the same mean and variance as S_n superimposed with graphs for part a. We are able to see that as n increases the PDF becomes more Gaussian.

Our results make sense due to the Central Limit Theorem. To verify that the Central Limit Theorem holds for this case, the PDF of the sum of the uniformly distributed variable should follow the normal distribution PDF as n increases. This is further shown in the last part of this experiment.

(d) Similarly, let $Y_1, Y_2 \ldots$ be another sequence of iid random variables following the exponential distribution with $\lambda = 1$, and let S_n' be the sum of the first n random variables. Based on your observation and understanding, pick a suitable n, write a MATLAB program to plot the empirical pdf and cdf of S_n', and superimpose the corresponding Gaussian distributions.

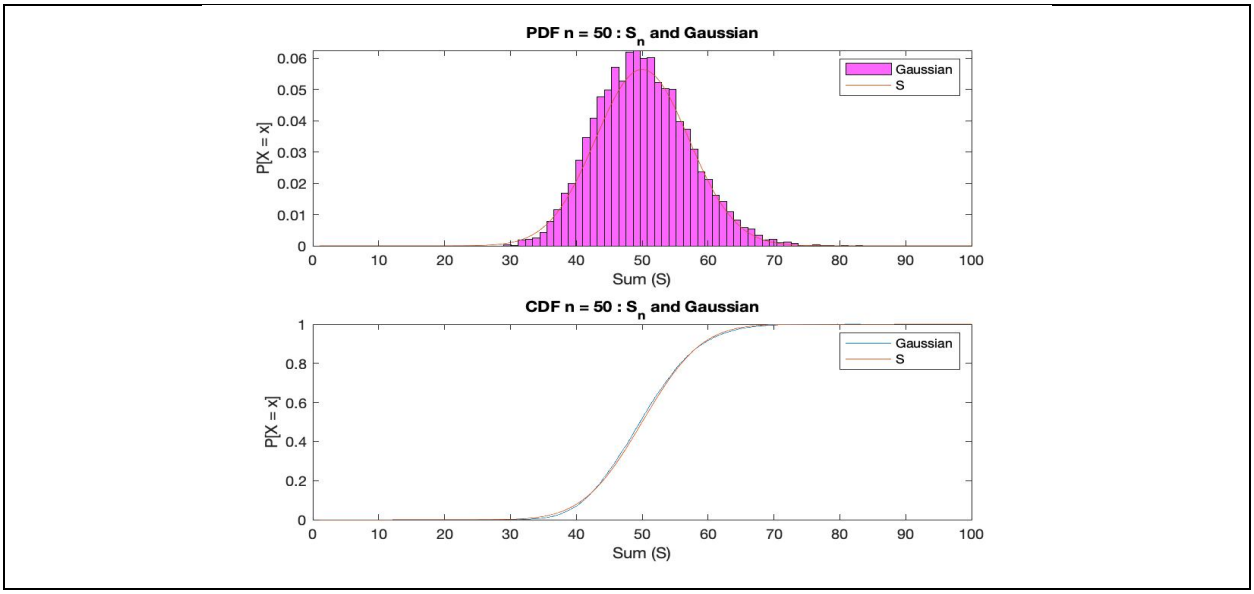


Figure 8: Plot the empirical pdf and cdf of S_n', superimposed with the corresponding Gaussian distributions. The value of n=50 was chosen because from part c it shows that PDF's and CDF's to match the Gaussian shape more closely when n was equal to 50.

n = 50 was chosen to best show that as n increases, the continuous uniform random resembles that of a normal distribution. From figure 8, we can see that our line closely outlines that of the Gaussian distribution, thus verifying our experiment.

Appendix

List of Figures:

MATLAB Files: