MEMO Number CMPE320_S21_PROJ3
DATE: 4/14/2021
TO: EFC LaBerge and CMPE320 Peers
FROM: Nem Negash
SUBJECT: Project 3 Report

1 INTRODUCTION

This project is focused on the Central Limit Theorem which states, for the *sample sum* of *N independent* random variables, $Y = \sum_{k=1}^{n} X_k$,where each random variable, X_k, has a finite mean, m_k, and a finite variance, σ_k^2. The Central Limit Theorem states that in this case, *Y* has a probability density function (pdf) that approaches a Gaussian pdf with mean $m = \sum_{k=1}^{n} m_k$ and a variance $\sigma^2 = \sum_{k=1}^{n} \sigma_k^2$ as *N* grows large (but not infinite!). The three sections of the project involve analytical computations as well as simulations.

2 DISCUSSION

2.1 Sum of Independent, Identically Distributed (iid) Random Variables from *U*(0,1)

Generate the sum of *N* random variables distributed $X_k \sim U(0,1), k=1,2...,N$ for $N=2$, $N=6$, $N=12$. MATLAB users will use the function `rand`.

Generate a large number of such sums, say 100,000 or more, for each value of *N*. Plot a histogram of the results for each *N* , scaling the histogram appropriately to be a probability density function. In each case, compute the mean and standard deviation of the samples and compare it to the theoretical expected value and variance for an infinitely large ensemble of such sums. Theoretical Mean and Variance calculations:

Theoretical expected value(mean) and variance:

$$\text{Mean} = \text{E[x]} = \int_0^1 x \cdot 1 dx = \left.\frac{x^2}{2}\right|_0^1 = \frac{1}{2} - 0 = \frac{1}{2}$$

$$E(y) = \sum E(x)$$

For some N, N = n, $E[y] = \frac{1}{12} * n$

For N = 2, $E[y] = \frac{1}{2} * 2 = 1$

For N = 2, $E[y] = \frac{1}{2} * 6 = 3$

For N = 2, $E[y] = \frac{1}{2} * 12 = 6$

$$\text{Variance} = \sigma^2 = E(x^2) - (E[x])^2$$

$$\text{E}(\text{x}^2) = \int_0^1 \text{x}^2 \text{dx} = \left.\frac{\text{x}^3}{3}\right|_0^1 \frac{1}{3} - 0 = \frac{1}{3}$$

$$\sigma^2 = E(x^2) - (E[x])^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

For some N, N = n, $\sigma^2 = \frac{1}{12} * n$

For N = 2, $\sigma^2 = \frac{1}{12} * 2 = \frac{1}{6}$

$$\text{For N = 2, } \sigma^2 = \frac{1}{12} * 6 = \frac{1}{2}$$
$$\text{For N = 2, } \sigma^2 = \frac{1}{12} * 12 = 1$$

On the same plot as the histogram, plot an appropriately scaled Gaussian curve with the theoretical mean and variance.

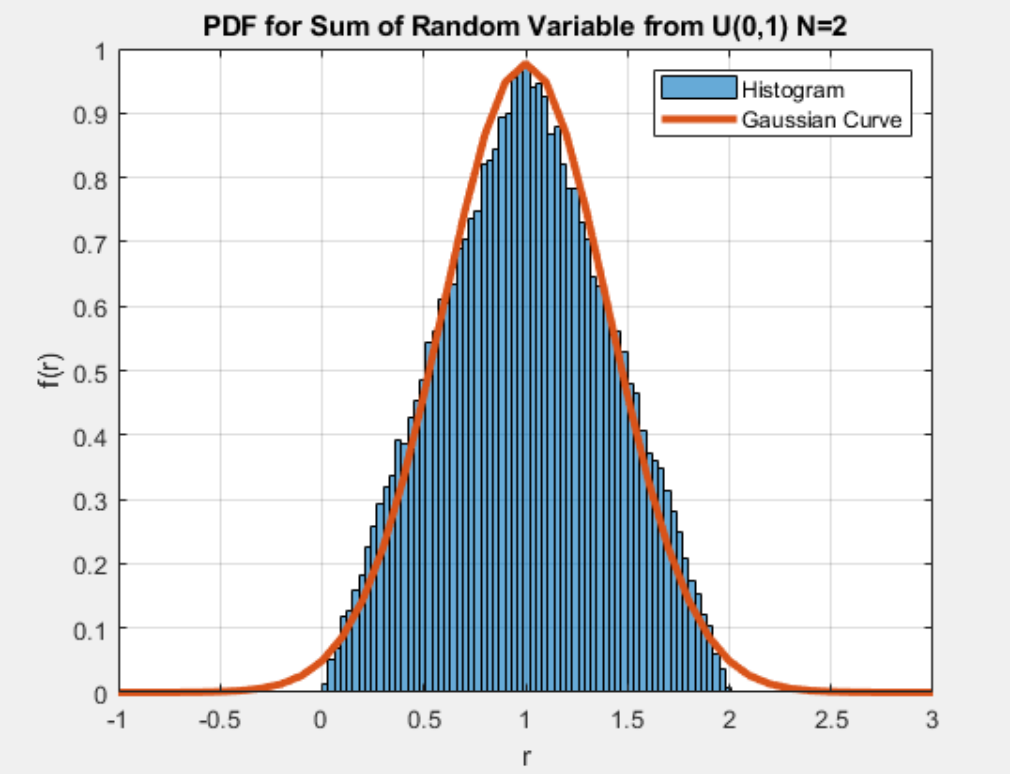

Figure 1: PDF for Sum of Random Variable from U(0,1) N=2

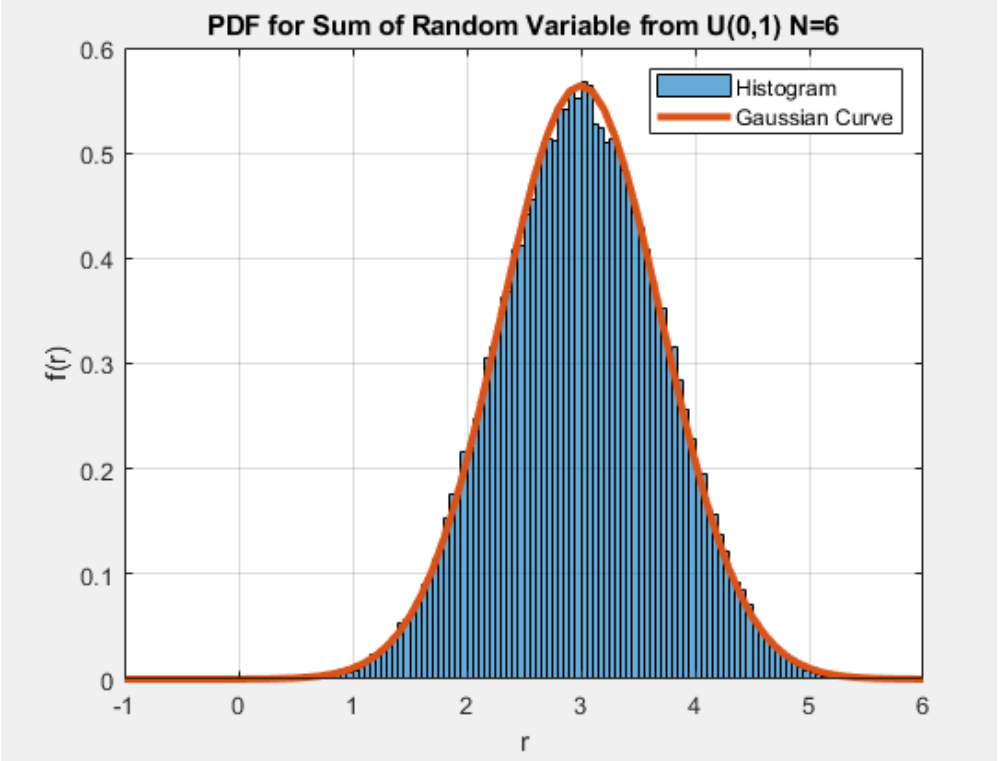

Figure 2: PDF for Sum of Random Variable from U(0,1) N=6

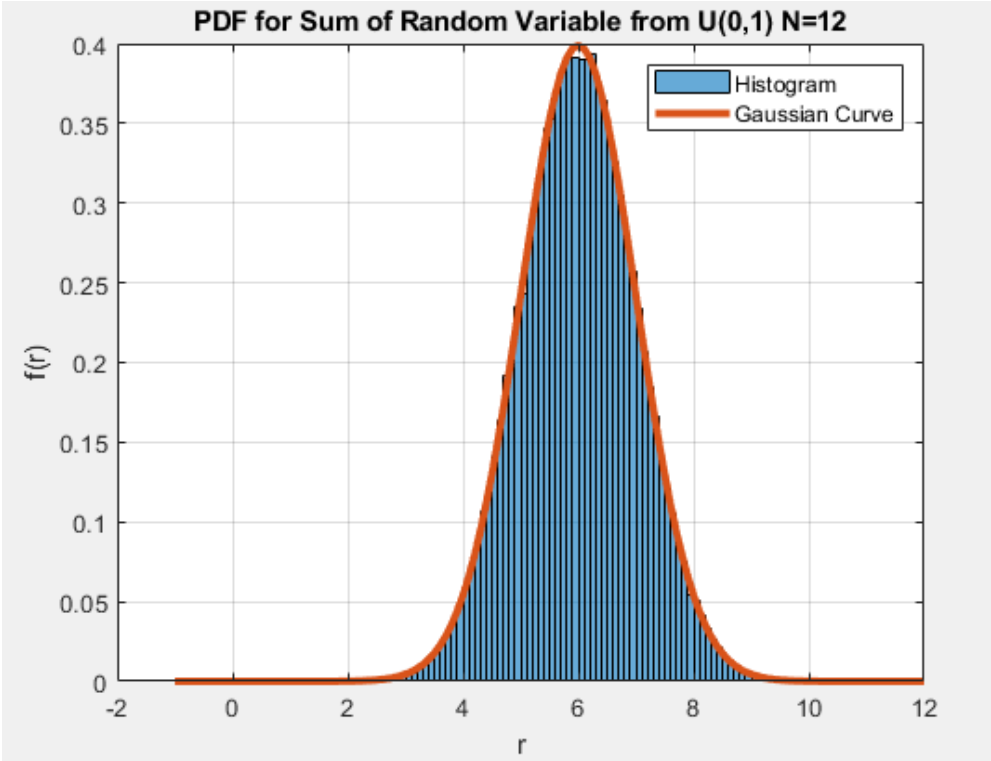

Figure 3: PDF for Sum of Random Variable from U(0,1) N=12

```
Experimental Variance, Mean and Standard Deviation for N = 2:
Mean: 0.999
Variance: 0.166
Standard Deviation: 0.407

Experimental Variance, Mean and Standard Deviation for N = 6:
Mean: 2.997
Variance: 0.498
Standard Deviation: 0.706

Experimental Variance, Mean and Standard Deviation for N = 12:
Mean: 6.000
Variance: 1.001
Standard Deviation: 1.001
```

Figure 4: Experimental Variance, Mean and Standard Deviation for N = 2, 6 ,12

Figure 1, 2, and 3 show the experimental sum of Independent, Identically Distributed Random Variables from U(0,1), using the rand MATLAB function, compared to the theoretical Gaussian curve. The experimental histogram fits the Gaussian curve more as the value for *N* increases. This is due the Central Limit Theorem, which states as the *N* value increases the pdf *Y*, where $Y = \sum_{k=1}^{n} X_k$,where each random variable, X_k, has a finite mean, m_k, and a finite variance, σ_k^2, approaches a Gaussian curve with a mean $m = \sum_{k=1}^{n} m_k$ and a variance $\sigma^2 = \sum_{k=1}^{n} \sigma_k^2$. **Figure 4** shows the experimental variance, mean, and standard deviation. It shows how when *N* increases the experimental mean and variance align with the theoretical values more consistently thus demonstrating the Central Limit Theorem..

2.2 Sum of Independent, Identically Distributed (iid) Discrete Random Variables

Repeat all the sections of 2.1 where the random variables are generated using `randi` simulate the rolling of the *N* fair eight-sided dice, followed by the sum of the values from each roll. Repeat this experiment a large number of times to create the histogram of the sum (I'm *not* interested in the values of the individual rolls!). Repeat the whole process for $N = 2$, $N = 10$, $N = 50$.

For each value of *N*, compute the mean and standard deviation of the samples and compare it to the theoretical expected value and variance for an infinitely large ensemble of such sums.

On the same plot as the histogram, plot an appropriately scaled Gaussian curve with the theoretical mean and variance.

Theoretical expected value(mean) and variance:

$$Mean = E[x] = \sum_{k=0}^{8} k_i * p = 1 * \left(\frac{1}{8}\right) + 2 * \left(\frac{1}{8}\right) + 3 * \left(\frac{1}{8}\right) + 4 * \left(\frac{1}{8}\right) + 5 * \left(\frac{1}{8}\right) + 6 * \left(\frac{1}{8}\right) + 7 * \left(\frac{1}{8}\right) + 8 * \left(\frac{1}{8}\right) = 4.5$$

$$E(y) = \sum E(x)$$

For some N, N = n, $E[y] = 4.5 * n$

For N = 2, $E[y] = 4.5 * 2 = 9$

For N = 2, $E[y] = 4.5 * 10 = 45$
For N = 2, $E[y] = 4.5 * 50 = 225$

$$\text{Variance} = \sigma^2 = E(x^2) - (E[x])^2$$

$$E[x^2] = \sum_{k=0}^{8}(k_i)^2 * p = 1^2 * \left(\frac{1}{8}\right) + 2^2 * \left(\frac{1}{8}\right) + 3^2 * \left(\frac{1}{8}\right) + 4^2 * \left(\frac{1}{8}\right) + 5^2 * \left(\frac{1}{8}\right) + 6 * \left(\frac{1}{8}\right) + 7 * \left(\frac{1}{8}\right) + 8 * \left(\frac{1}{8}\right) = 25.5$$

$$\sigma^2 = E[x^2] - (E[x])^2 = 25.5 - 20.25 = 5.25$$

For some N, N = n, $\sigma^2 = 5.25 * n$
For N = 2, $\sigma^2 = 5.25 * 2 = 10.5$
For N = 2, $\sigma^2 = 5.25 * 10 = 52.5$
For N = 2, $\sigma^2 = 5.25 * 50 = 262.5$

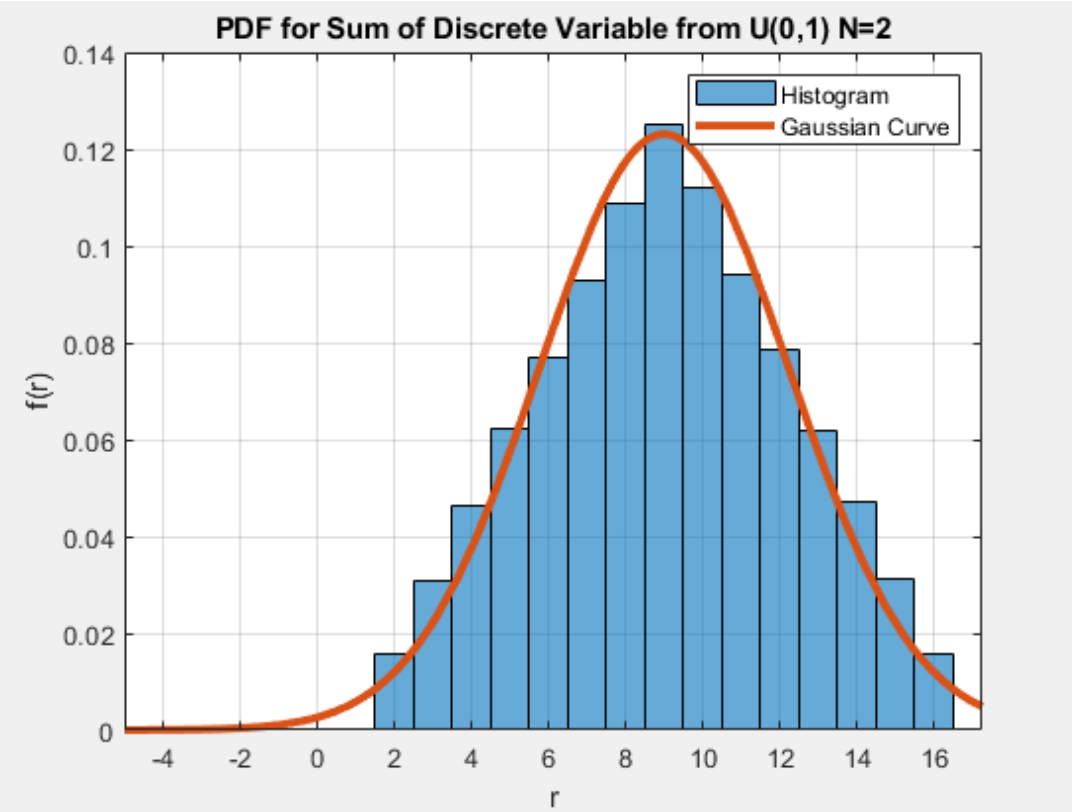

Figure 5: PDF for Sum of Discrete Variable from U(0,1) N=2

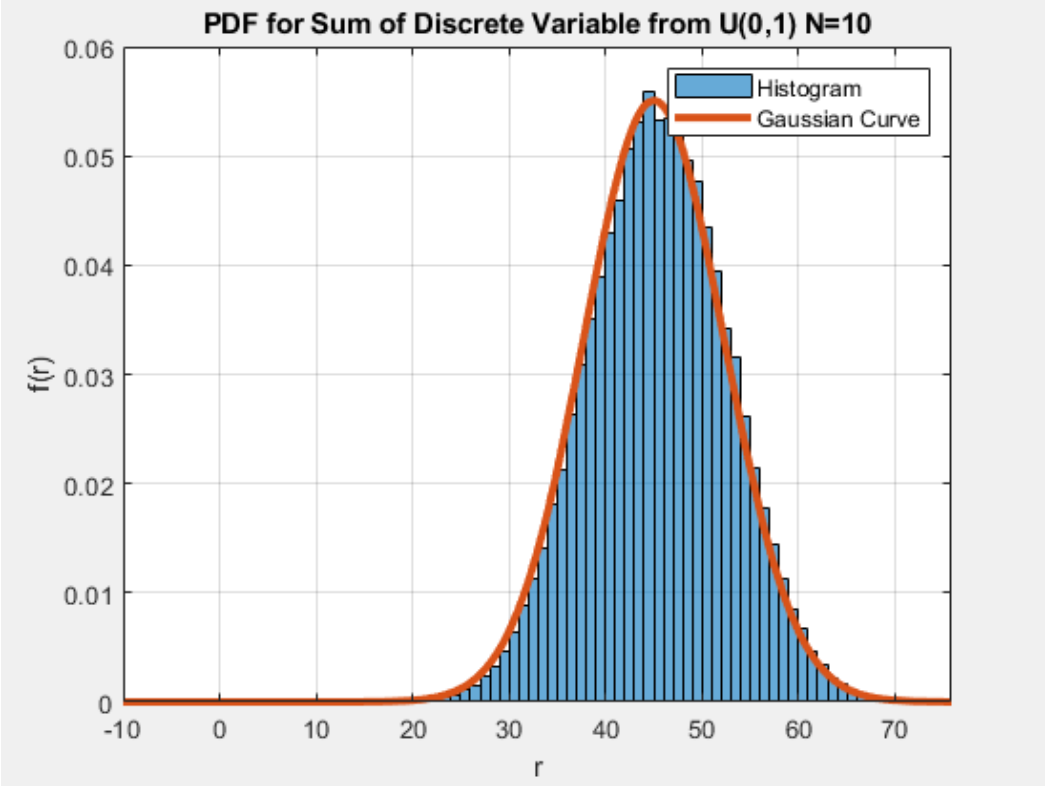

Figure 6: PDF for Sum of Discrete Variable from U(0,1) N=10

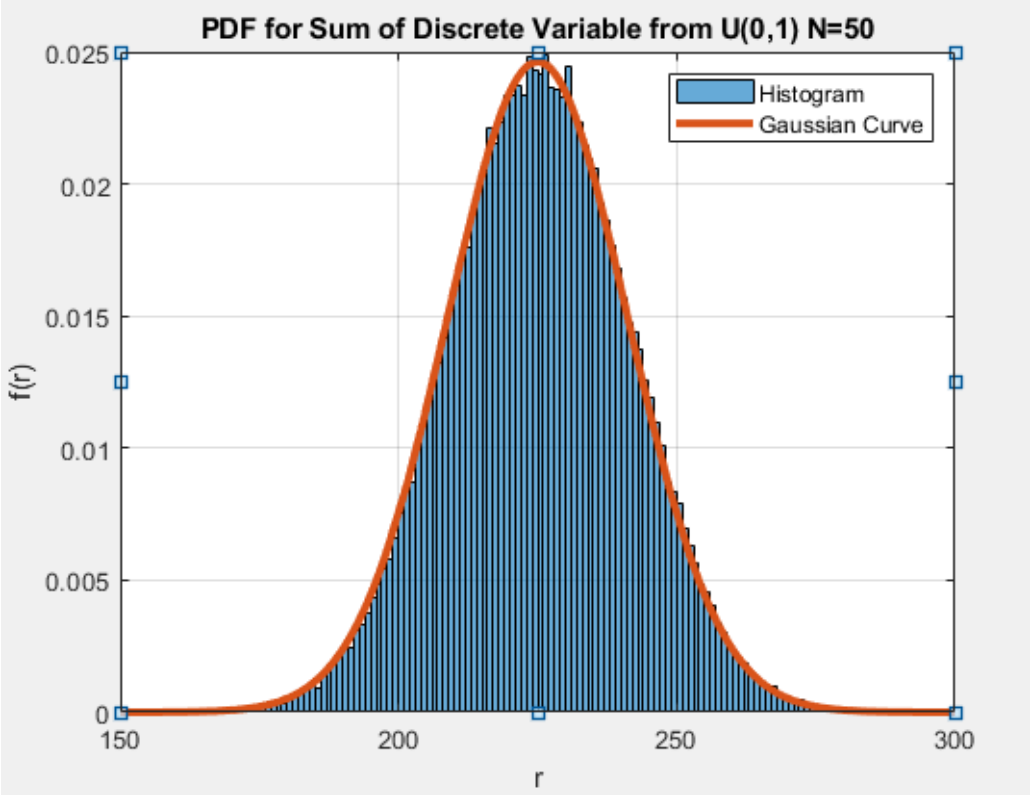

Figure 7: PDF for Sum of Discrete Variable from U(0,1) N=50

```
Experimental Variance, Mean and Standard Deviation for N = 2:
Mean: 9.005
Variance: 10.484
Standard Deviation: 3.238

Experimental Variance, Mean and Standard Deviation for N = 10:
Mean: 45.012
Variance: 52.251
Standard Deviation: 7.228

Experimental Variance, Mean and Standard Deviation for N = 50:
Mean: 225.033
Variance: 262.599
Standard Deviation: 16.205
```

Figure 8: Experimental Variance, Mean and Standard Deviation for N = 2, 10, 50

Figure 5, 6, and 7 show the experimental sum of Independent, Identically Distributed Discrete Variables, using the randi MATLAB function, compared to the theoretical Gaussian curve. The experimental histogram fits the Gaussian curve more as the value for *N* increases. This is due the Central Limit Theorem, which states as the *N* value increases the pdf *Y*, where $Y = \sum_{k=1}^{n} X_k$,where each random variable, X_k, has a finite mean, m_k, and a finite variance, σ_k^2, approaches a Gaussian curve with a mean $m = \sum_{k=1}^{n} m_k$ and a variance $\sigma^2 = \sum_{k=1}^{n} \sigma_k^2$. **Figure 8** shows the experimental variance, mean, and standard deviation. As *N* increases the experimental mean and variance lines up with the theoretical values more consistently showing how the Central Limit Theorem stands.

2.3 Sum of Independent, Identically Distributed (iid) Random Variables from $P_x(x) = 0.5e^{-0.5x}$

Repeat all the sections of 2.1 where the random variables are generated using the function `randx` provided with Project 1. Use $N = 2$, $N = 10$, $N = 100$. Note that this pdf has a sharp discontinuity at $x=0$, but that eventually the histogram *does* approach the Gaussian! The CLT is a *powerful* theorem!

In each case, compute the mean and standard deviation of the samples and compare it to the theoretical expected value and variance for an infinitely large ensemble of such sums.

Theoretical expected value(mean) and variance:

$$\text{Mean} = \text{E[x]} = \int_0^\infty \text{x} * \text{p}_\text{x}(\text{x})\text{dx} = \int_0^\infty \text{x} * 0.5\text{e}^{-0.5\text{x}}\text{dx} = 2$$

$$E(y) = \sum E(x)$$

For some N, N = n, $E[y] = 2 * n$

For N = 2, $E[y] = 2 * 2 = 4$

For N = 2, $E[y] = 2 * 10 = 20$

For N = 2, $E[y] = 2 * 100 = 200$

$$\text{Variance} = \sigma^2 = E(x^2) - (E[x])^2$$

$$E[x^2] = \int_0^\infty x^2 * p_x(x)dx = \int_0^\infty x^2 * 0.5e^{-0.5x}dx = 8$$

$$\sigma^2 = E[x^2] - (E[x])^2 = 8 - 4 = 4$$

For some N, N = n, $\sigma^2 = 4 * n$

For N = 2, $\sigma^2 = 4 * 2 = 8$

For N = 2, $\sigma^2 = 4 * 10 = 40$

For N = 2, $\sigma^2 = 4 * 100 = 400$

On the same plot as the histogram, plot an appropriately scaled Gaussian curve with the theoretical mean and variance.

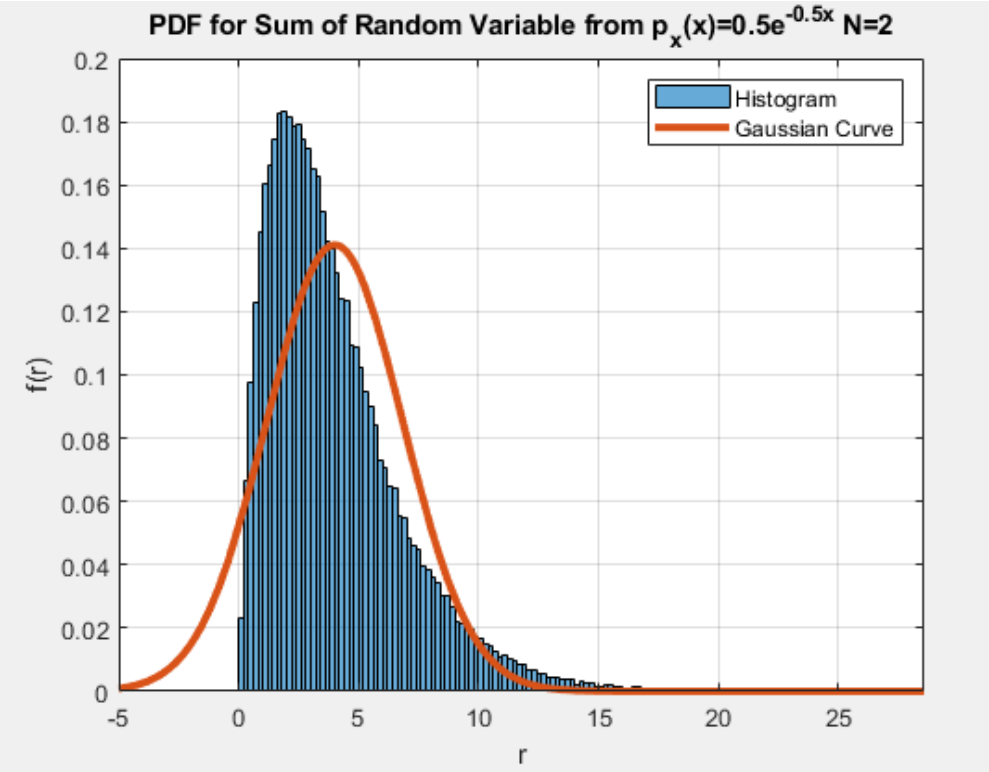

Figure 9: PDF for Sum of Random Variable from $p_x(x)=0.5e^{-0.5x}$ N=2

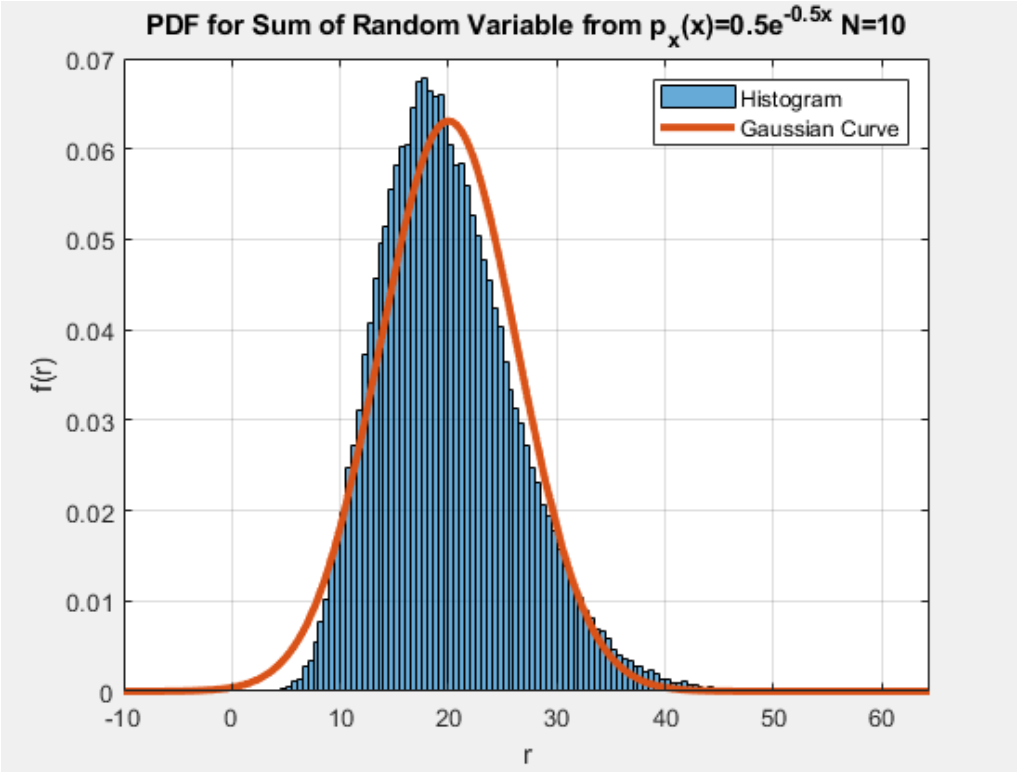

Figure 10: PDF for Sum of Random Variable from $p_x(x)=0.5e^{-0.5x}$ N=10

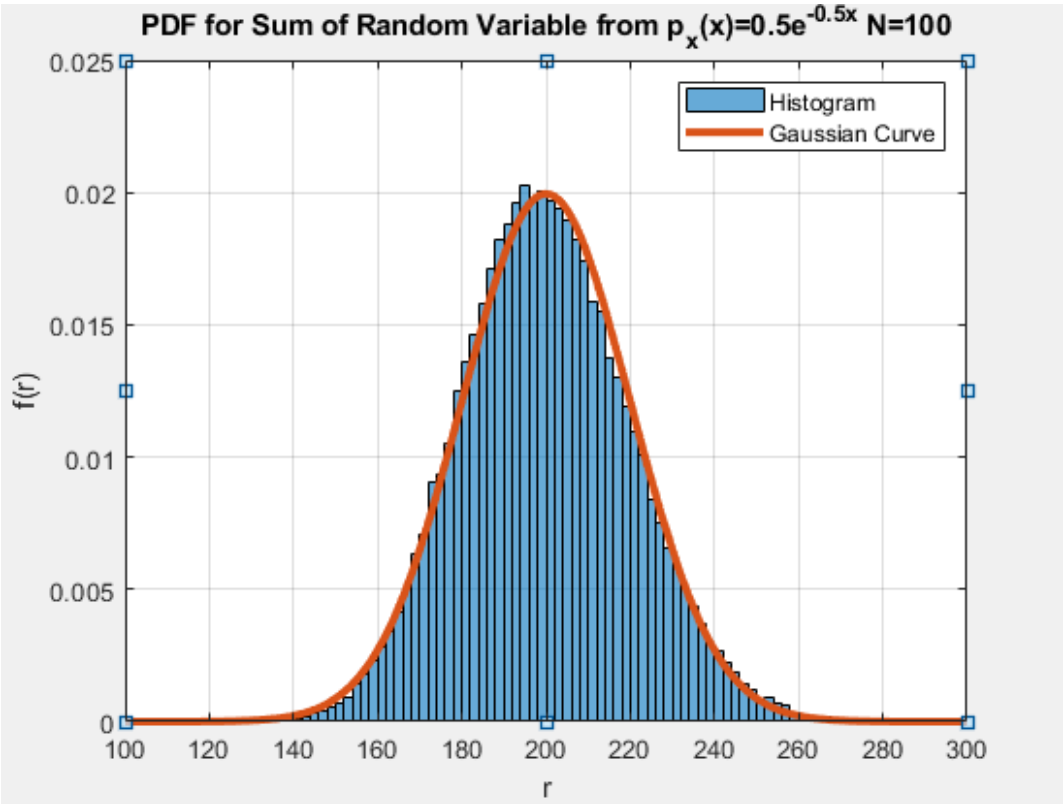

Figure 11: PDF for Sum of Random Variable from $p_x(x)=0.5e^{-0.5x}$ N=100

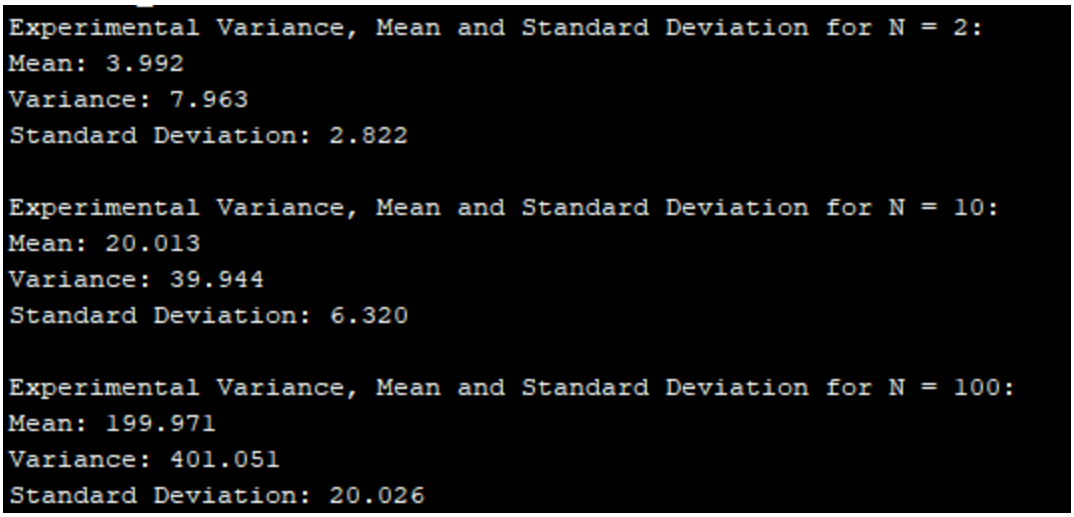

Figure 12: Experimental Variance, Mean and Standard Deviation for N = 2, 10, 100

Figure 9, 10, and 11 show the experimental sum of Independent, Identically Distributed Random Variables from $P_x(x)=0.5e^{-0.5x}$, using the provided randx MATLAB function, compared to the theoretical Gaussian curve. The experimental histogram fits the Gaussian curve more as the value for *N* increases. This is due the Central Limit Theorem, which states as the *N* value increases the pdf *Y*, where $Y = \sum_{k=1}^{n} X_k$,where each random variable, X_k, has a finite mean, m_k, and a finite variance, σ_k^2, approaches a Gaussian curve with a mean $m = \sum_{k=1}^{n} m_k$ and a variance $\sigma^2 = \sum_{k=1}^{n} \sigma^2_k$. **Figure 12** shows the experimental variance, mean, and standard deviation. As *N* increases the experimental mean and variance lines up with the theoretical values more consistently once again proving the Central Limit Theorem.

2.4 Sum of Independent, Identically Distributed (iid) Bernoulli Trials

Let a single Bernoulli trial result in either a one (1) or a zero (0), with Pr[$X = = 0.5$] [1]. Perform *N* independent trials. What is the form of the pmf of the random variable *K*= number of 1's in *N* trials? (I'm looking for a specific name here, go review the standard pmfs.) Because the random variable *K* is a sum of independent random variables, each of which has finite mean *m*= 0.5 and finite variance, s^2 = 0.5, the CLT should hold. Use values of *N*= 4, *N*= 8, and *N* as large as you can without causing a MATLAB overflow. Note that *N* is the number of Bernoulli/Binary random variables in the *sum*. The sum itself produces *one* value of the random variable *K*. You need to do this process many times to generate your histogram.

Each figure should consist of two subplots. On the first, plot the (theoretical) probability density function of the sum of *N* independent Bernoulli trials. Plot the CLT Gaussian approximation on the same plot and compare the results.

On the second, generate and plot the scaled histogram for the sum of a large number of random trials of the sum of *N* iid Bernoulli experiments and compare it to the theoretical. Plot the CLT Gaussian approximation on the same histogram plot and compare the results.

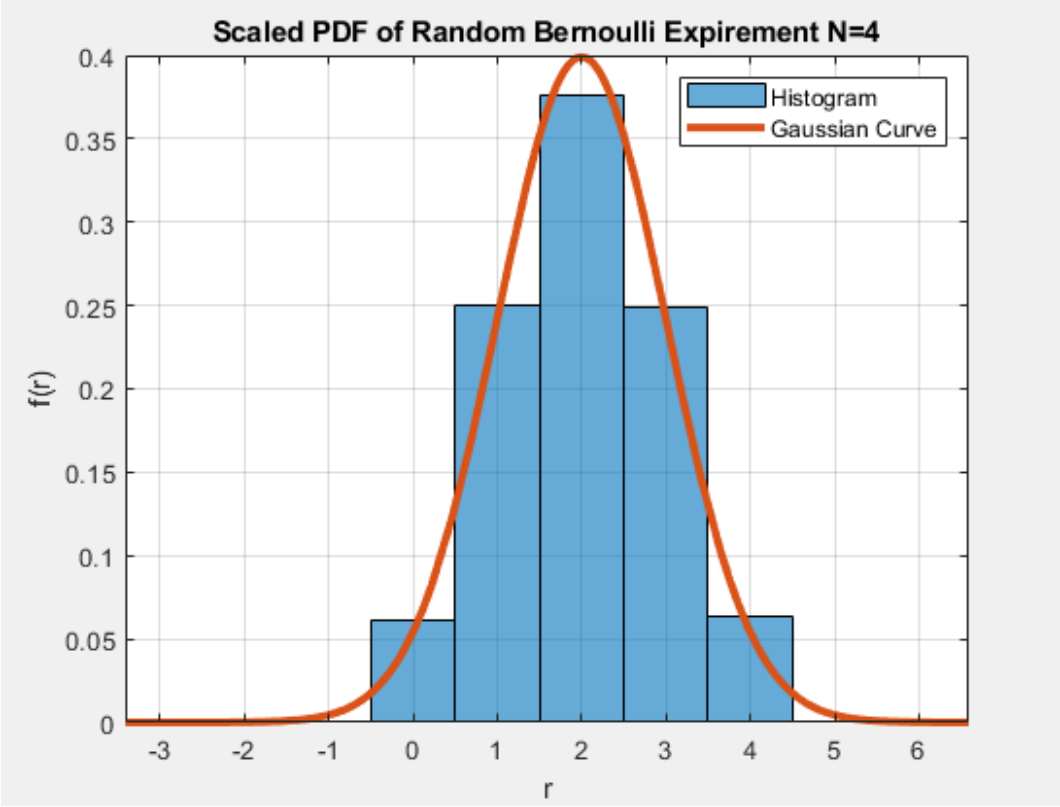

Figure 13: Scaled PDF of Random Bernoulli Experiment N=4

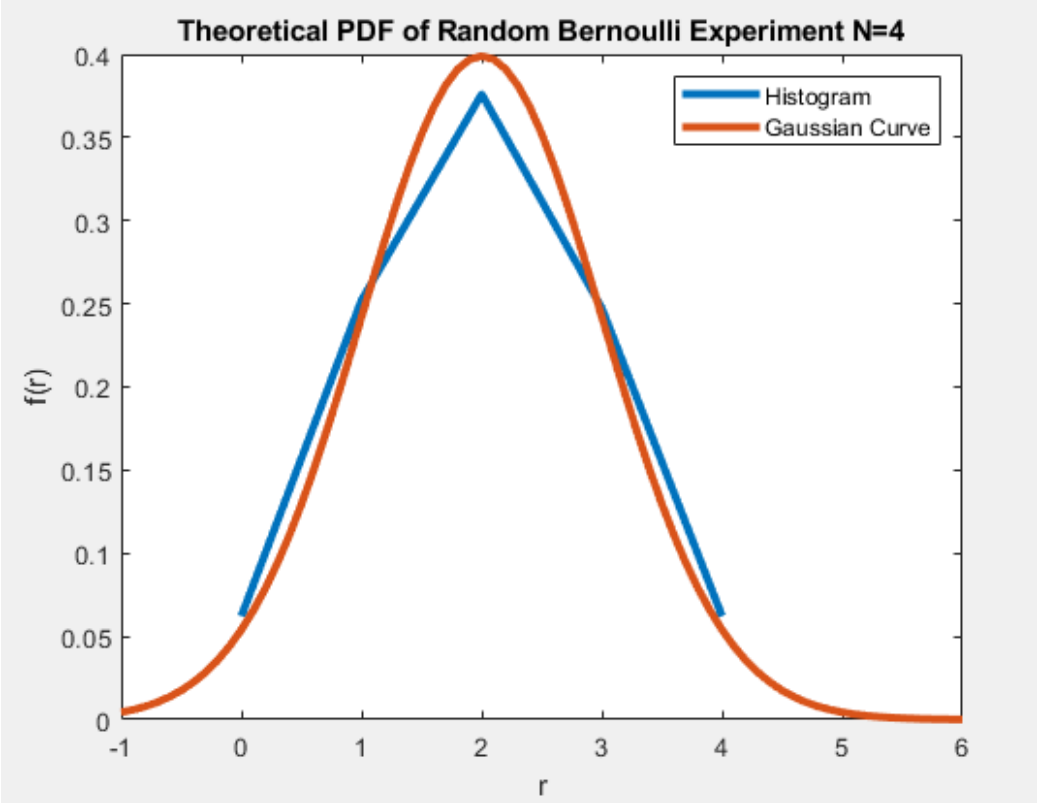

Figure 14: Theoretical PDF of Random Bernoulli Experiment N=4

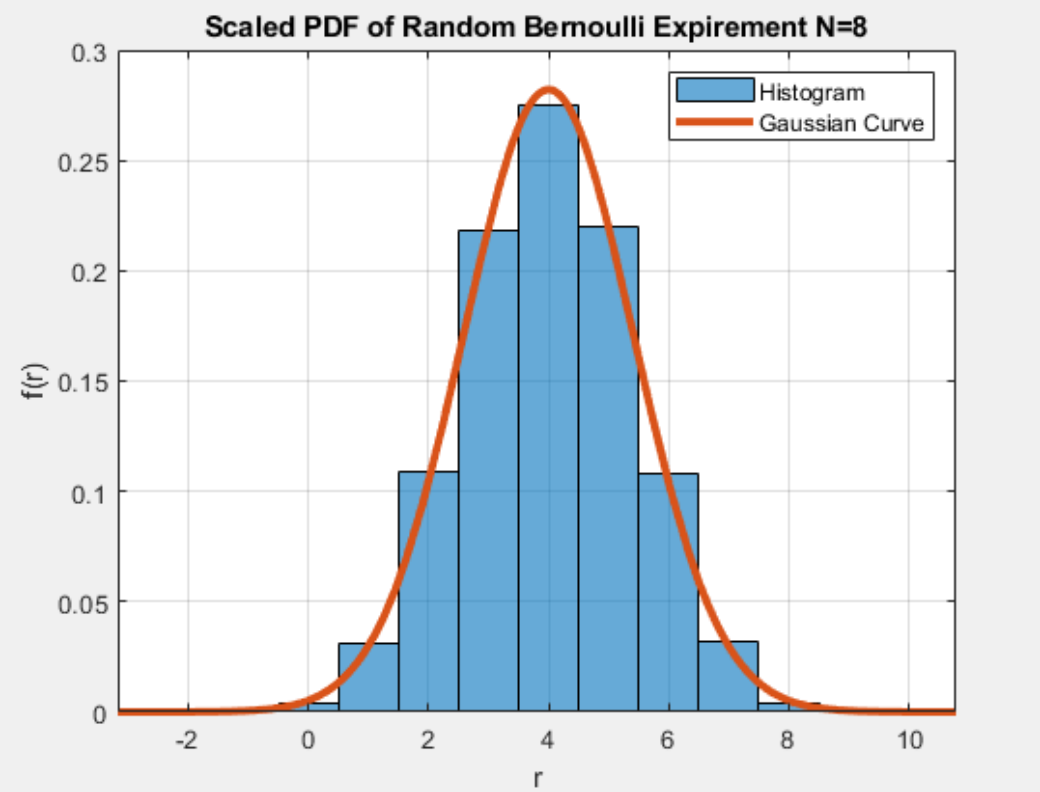

Figure 15: Scaled PDF of Random Bernoulli Experiment N=8

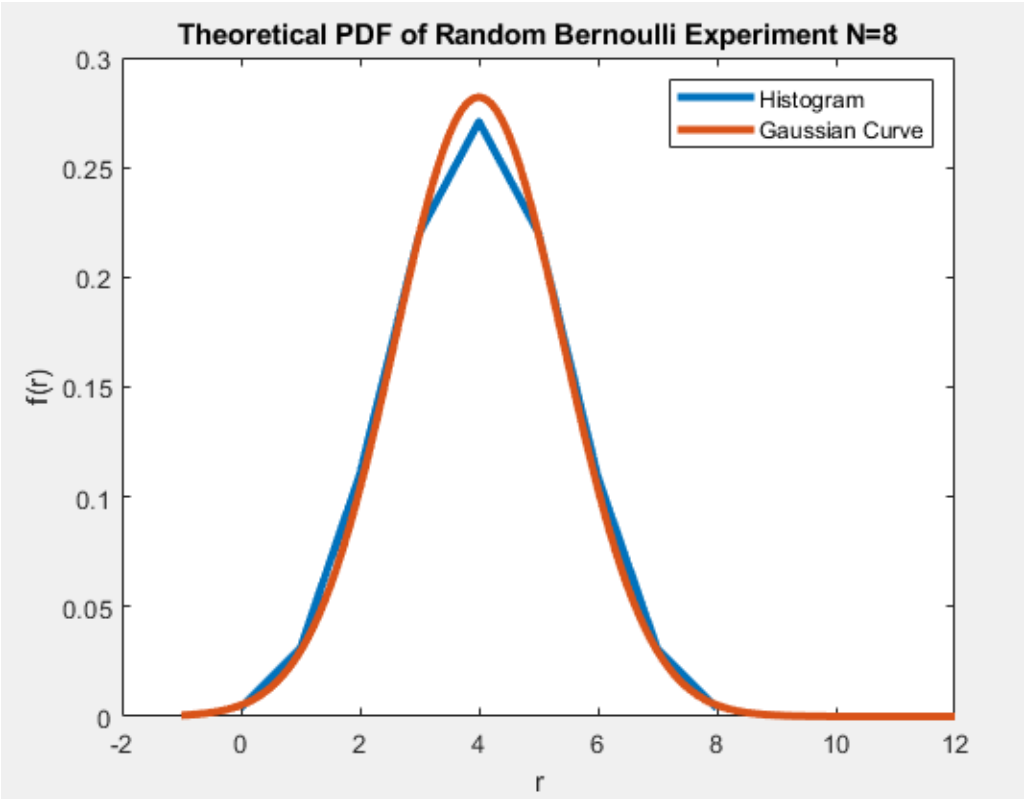

Figure 16: Theoretical PDF of Random Bernoulli Experiment N=8

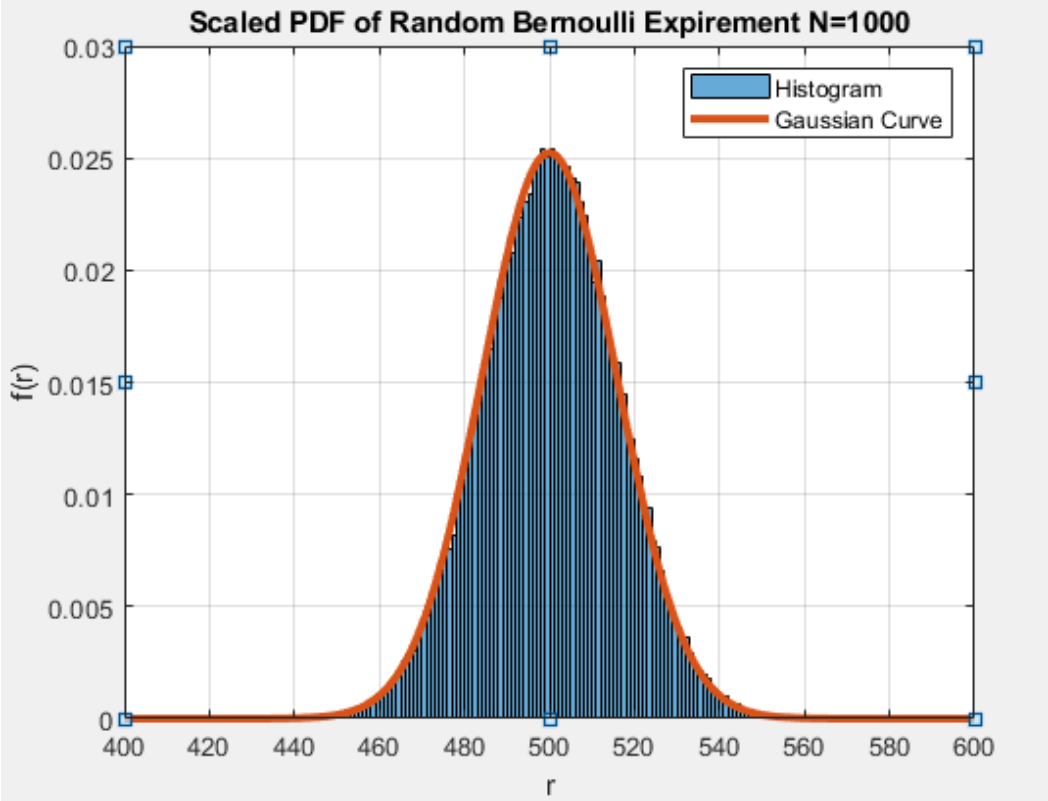

Figure 17: Scaled PDF of Random Bernoulli Experiment N=1000

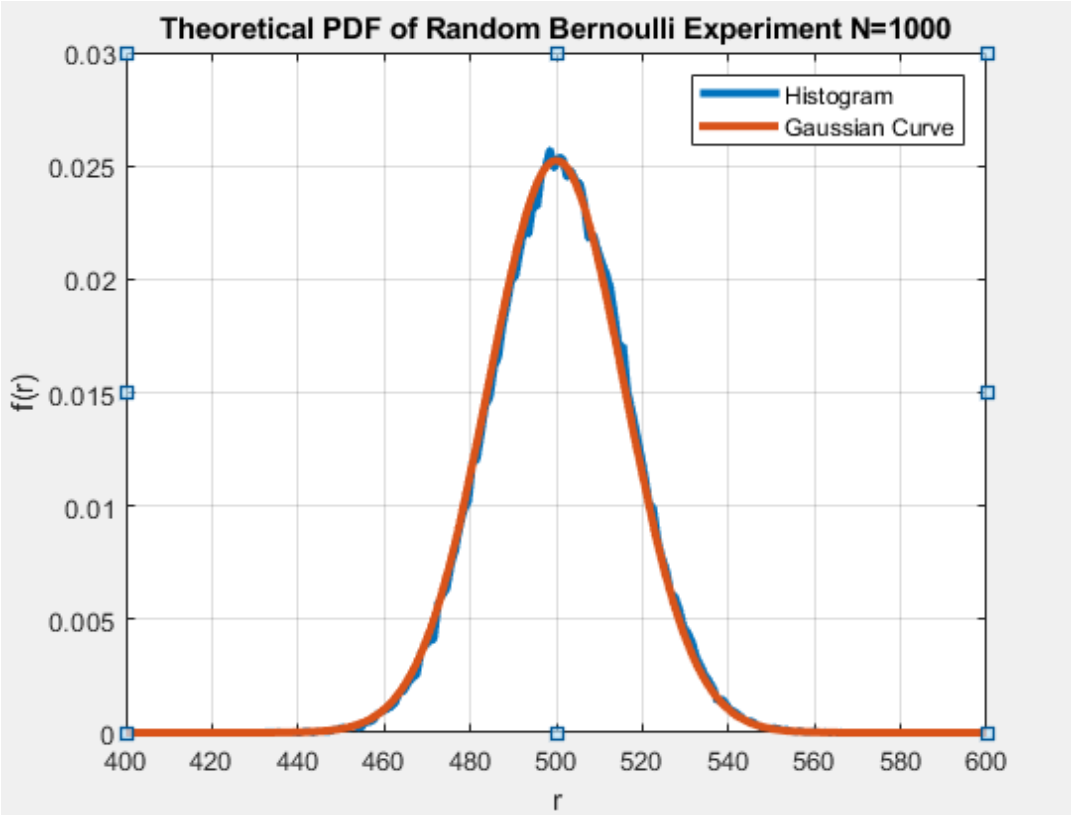

Figure 18: Theoretical PDF of Random Bernoulli Experiment N=1000

Figure 13-18 show the experimental sum of Independent, Identically Distributed Bernoulli trials compared to the theoretical Gaussian curve. The experimental trials are calculated using the rand MATLAB function and taking all values $r \leq 0.5$ to be 1 and the rest as 0. The Binomial distribution histogram that represents the Bernoulli trials fits the Gaussian curve more as the value for N increases. This is due the Central Limit Theorem, which states as the N value increases the pdf Y, where $Y = \sum_{k=1}^{n} X_k$,where each random variable, X_k, has a finite mean, m_k, and a finite variance, σ_k^2, approaches a Gaussian curve with a mean $m = \sum_{k=1}^{n} m_k$ and a variance $\sigma^2 = \sum_{k=1}^{n} \sigma_k^2$.

3 WHAT I LEARNED

This project has really given me a much better understanding of the sum of random and discrete variable PDFs, the sum of Binomial distributed Bernoulli trials and the Central Limit Theorem. Researching further about the Central Limit Theorem has given me context of its application in the real world. This theorem is used in things such as calculating product needs by supermarkets in a specific region to simplify the process from rather using the product sales from every single store in the region, you can use samples from weekly sales in the stores (the population), calculate their mean (the average number of products sold) and build the distribution of the sample means. This project has also expanded my understanding of MATLAB functions by using functions such as syms and the different types of files MATLAB provides such as scripts and live scripts.

3.1 Future suggestions

This project overall was good. It was concise, instructive, helped me learn a lot about MATLAB, and strengthen my understandings of the topics covered in lecture.

3.2 Time spent

Topic	Time Spent
Reading	5 hours
Research	5 hours
Programming	5 hours
Writing	10 hours
Final Preparation	1 hours
Total:	26 hours