MEMO Number CMPE320-S21-PROJ1
DATE: 02/27/2021
TO: EFC LaBerge and CMPE320 classmates
FROM: Nem Negash
SUBJECT: Project 1 Report

1   INTRODUCTION

This project was focused on using histograms to simulate a probability density function (pdf or PDF) or a probability mass function (pmf or PMF) for a random variable and using the PMFs and PDFs to calculate probabilities. The six different parts of the project help expand and educate on the use of MATLAB functions to produce in enticing plots that are informative about the trials conducted.

2   SIMULATION AND DISCUSSION

2.1   PMF for a single fair die

The purpose of this section is, using the built-in MATLAB function randi (imax, m, n), model the probabilities each role for the role of a fair six-sided die. There were a set of four trial groups of which were, 120; 1,200; 12,000;120,000 trials. The plots for each set of trials are as shown:
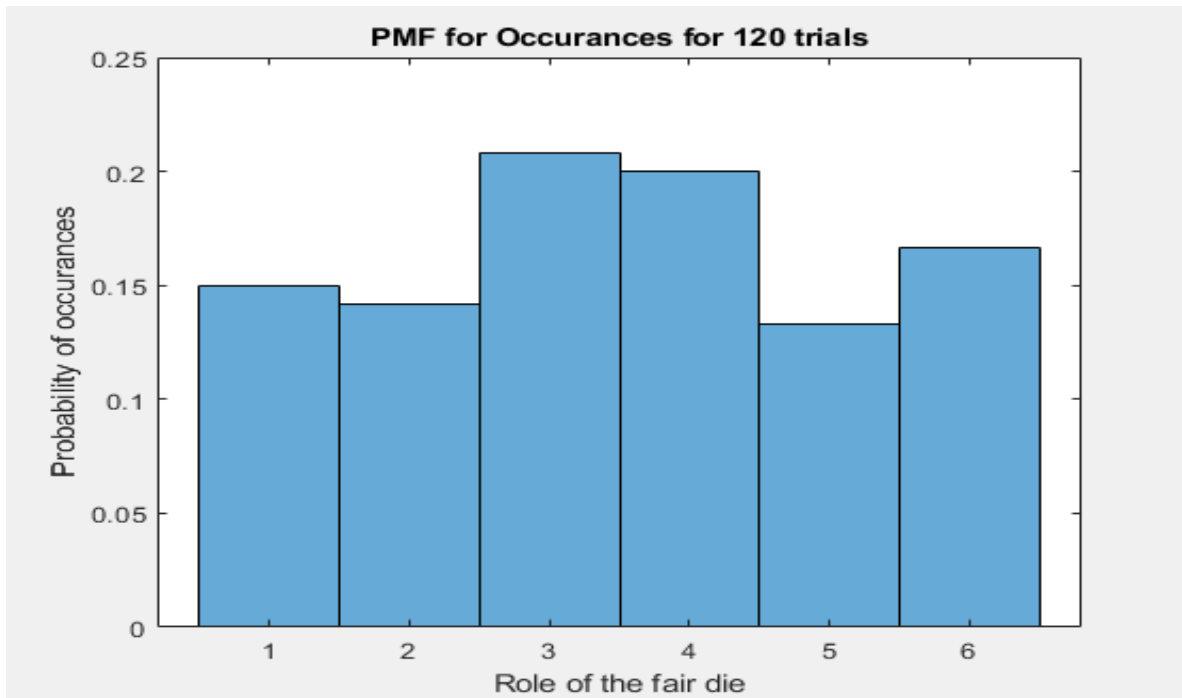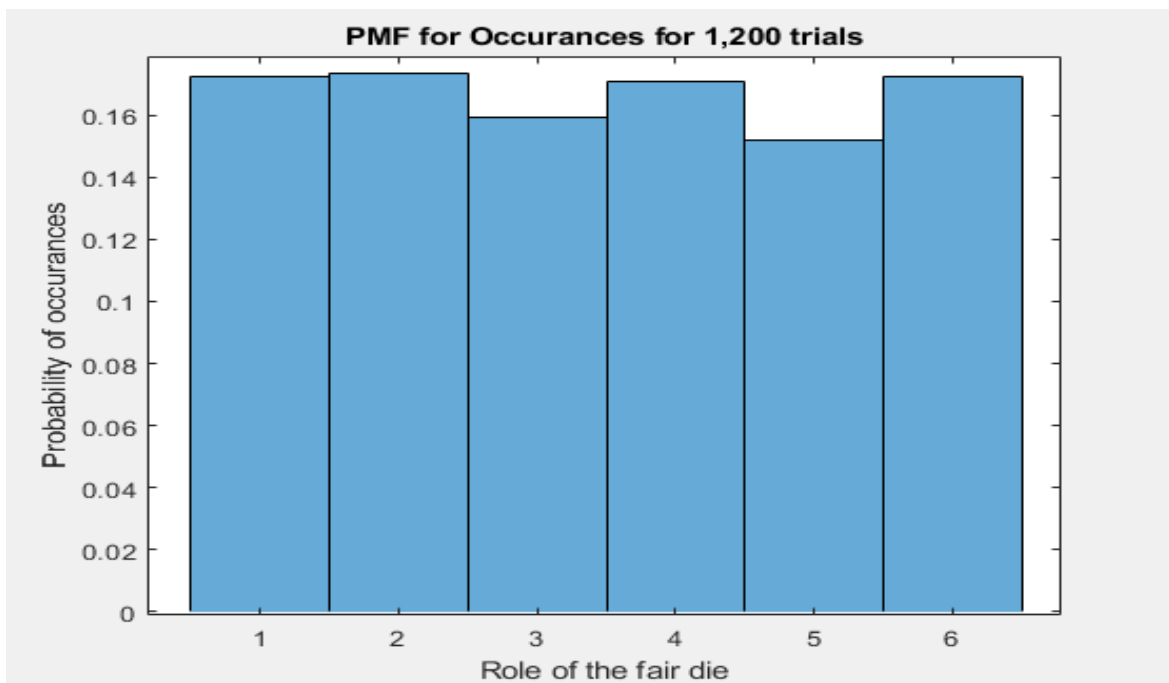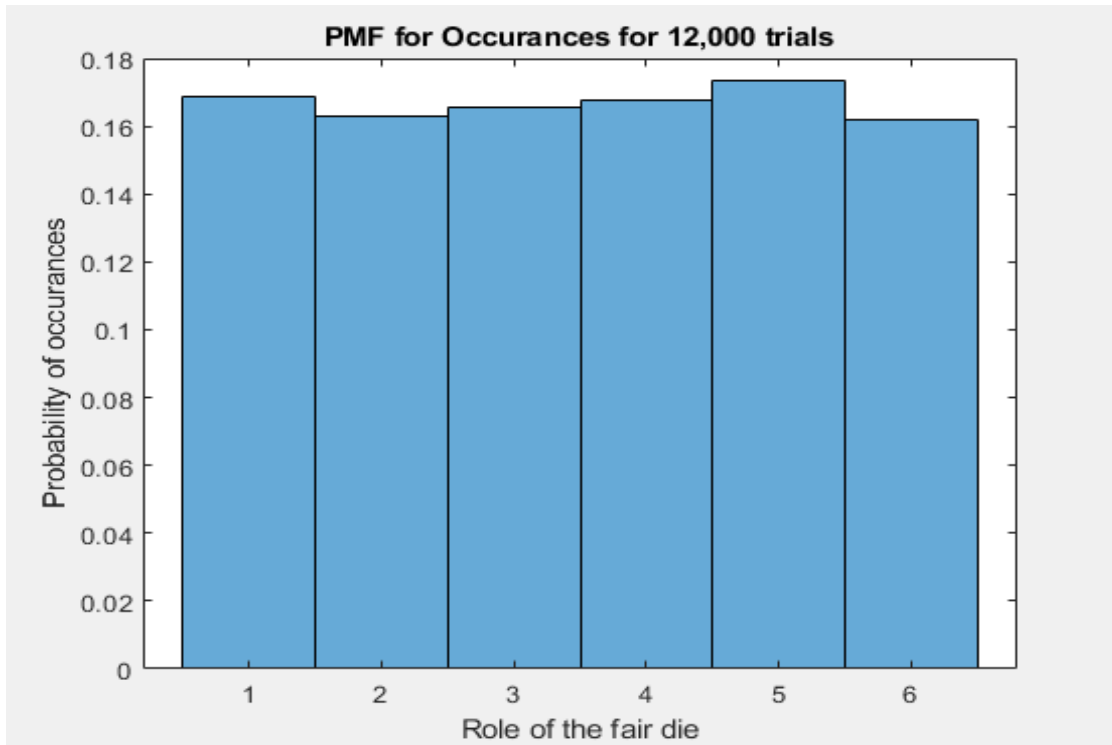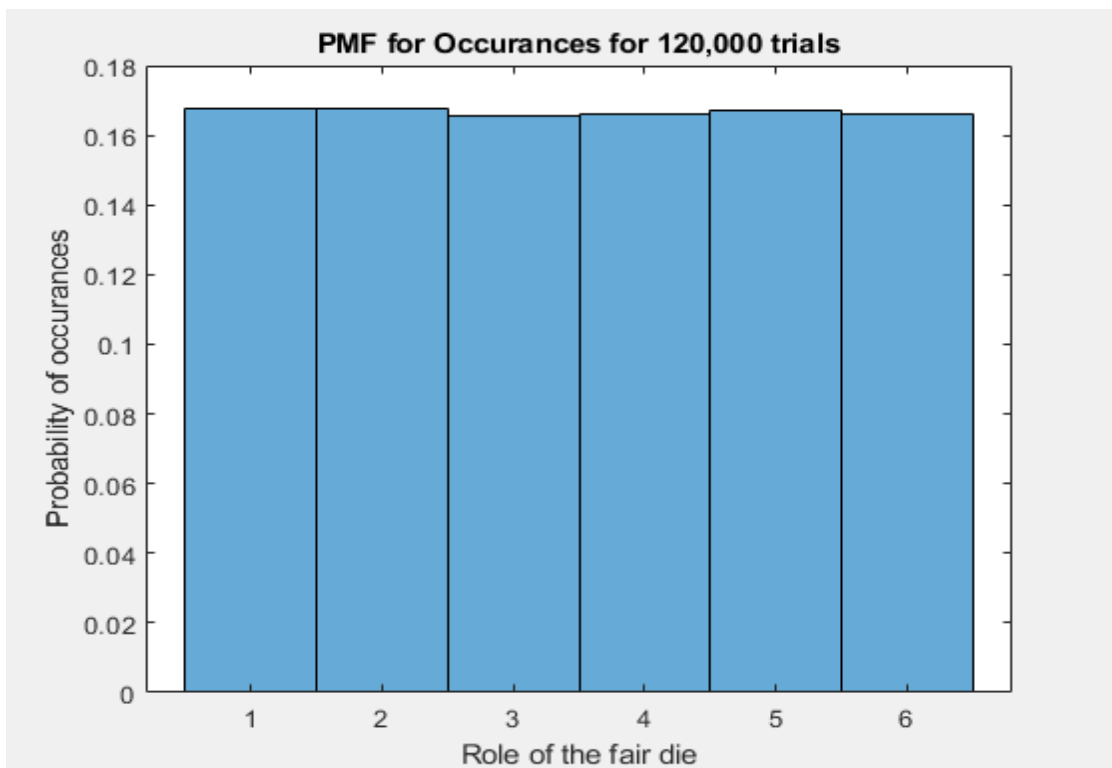
Figure 1: PMF for 120 trials

Figure 2:PMF for 1,200 trials

Figure 3: PMF for 12,000 trials

Figure 4: PMF for 120,000 trials

As shown by the histograms the probability for each role becomes more and more equal to one another as more trials are conducted. This shows that a larger sample size leads to a closer result to the analytical values at which each role would have a probability of $p = 0.167$.

The sample mean and sample variance for the four set of trials is as follows:

```
The sample mean for 120 trials: 3.275
The sample variance for 120 trials: 2.957

The sample mean for 1,200 trials: 3.519
The sample variance for 1,200 trials: 2.769

The sample mean for 12,000 trials: 3.499
The sample variance for 12,000 trials: 2.912

The sample mean for 120,000 trials: 3.506
The sample variance for 120,000 trials: 2.919
```

Figure 5: Sample Mean and Variance for set of trials

The analytical values for the mean and variance are 3.5 and 2.9167. The sample mean and variance slightly differ from the analytical values as shown in **Figure 5**. The sample mean and variance become more consistently aligned with the analytical values as more trials are conducted.

2.2   PMF for binary strings

The goal of this section is to model a series of m strings, $m = [20;200;2,000;200,000]$, of ones and zeros with each a length $n = 100$. Using the built-in MATLAB function rand(m,n), for each set of trials set he probability of getting a one, $p_1 = \Pr\{1\} = 0.5, 0.1, 0.9$. For each value of p_1 find the first occurrence of a 1 and set the index to be the random variable. The Plots for each value of p_1 and four values of m are as follows:
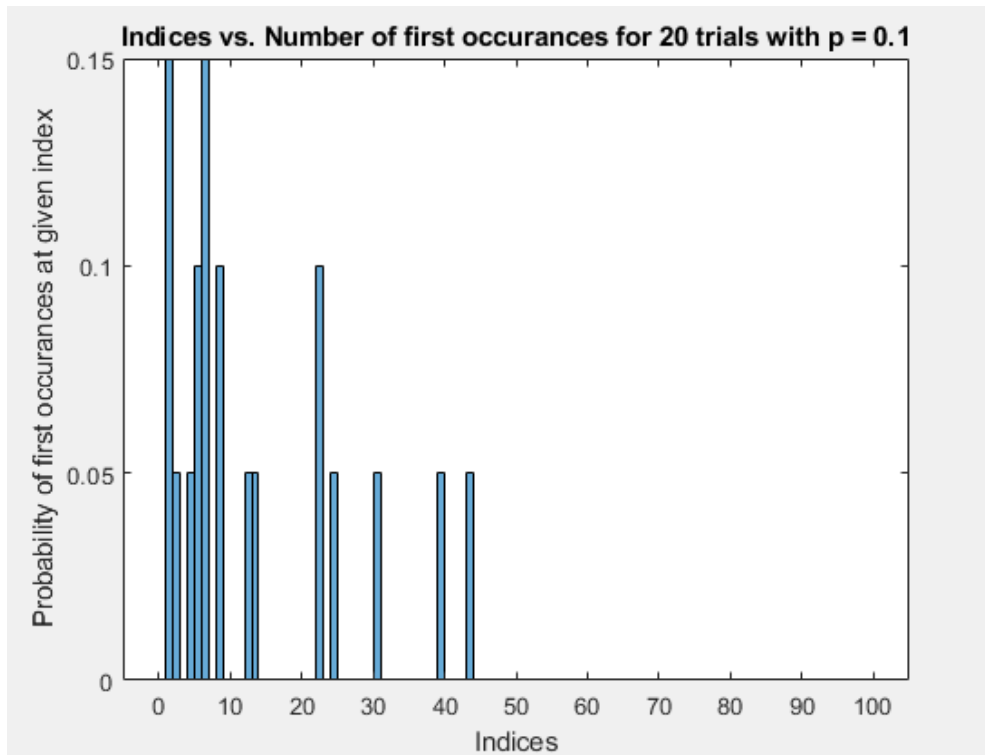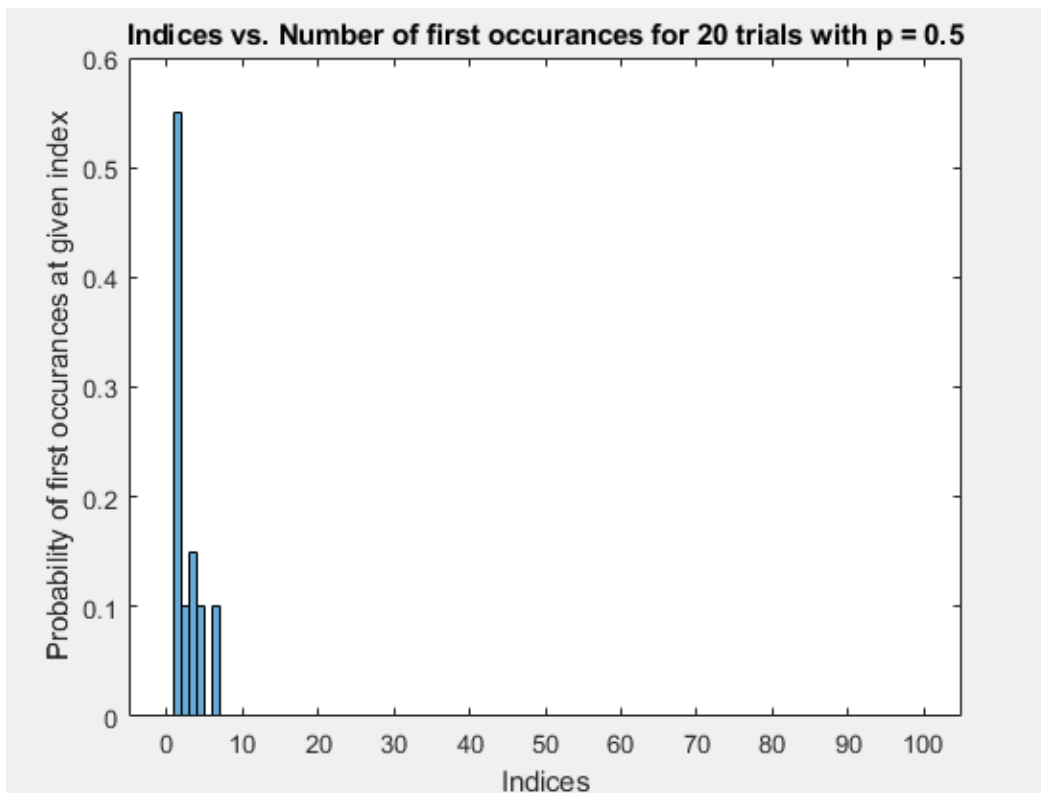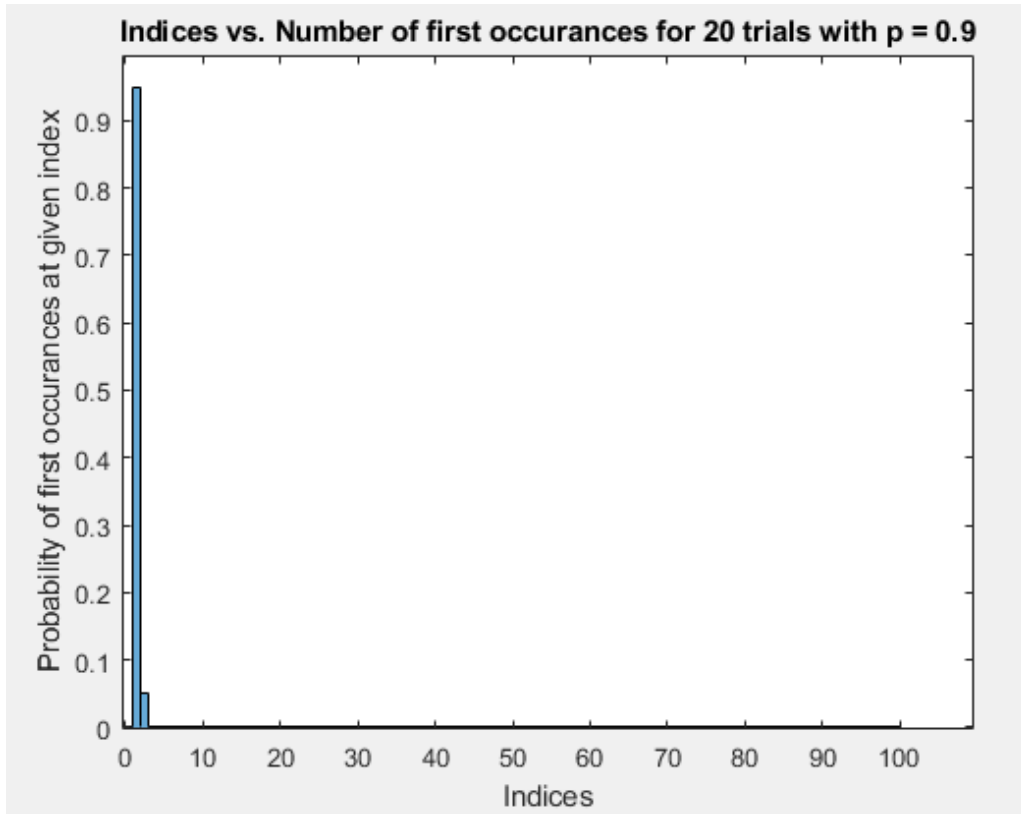
Figure 6: PMF for p1 = 0.1 at 20 trials
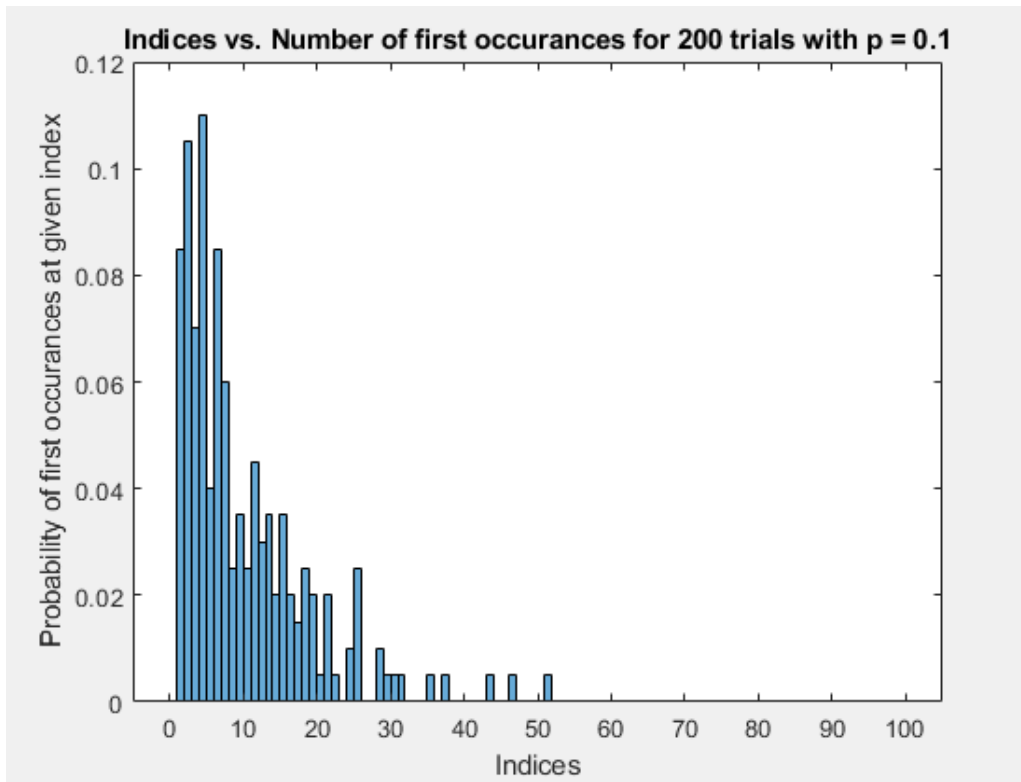

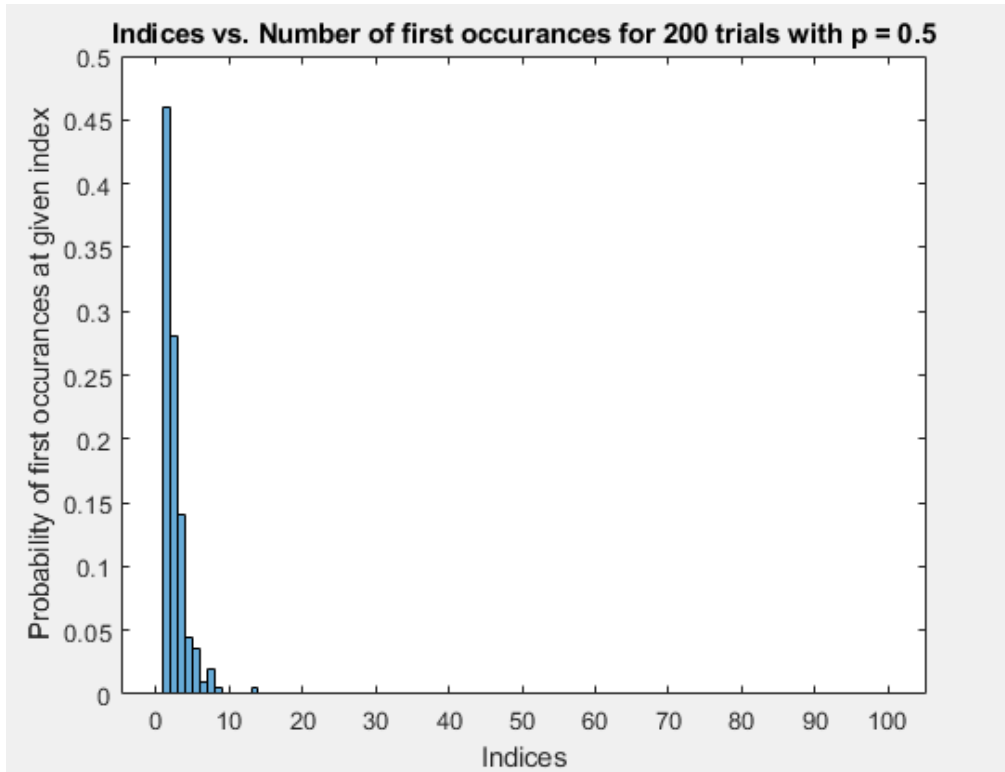**Figure 7: PMF for p1 = 0.5 at 20 trials**
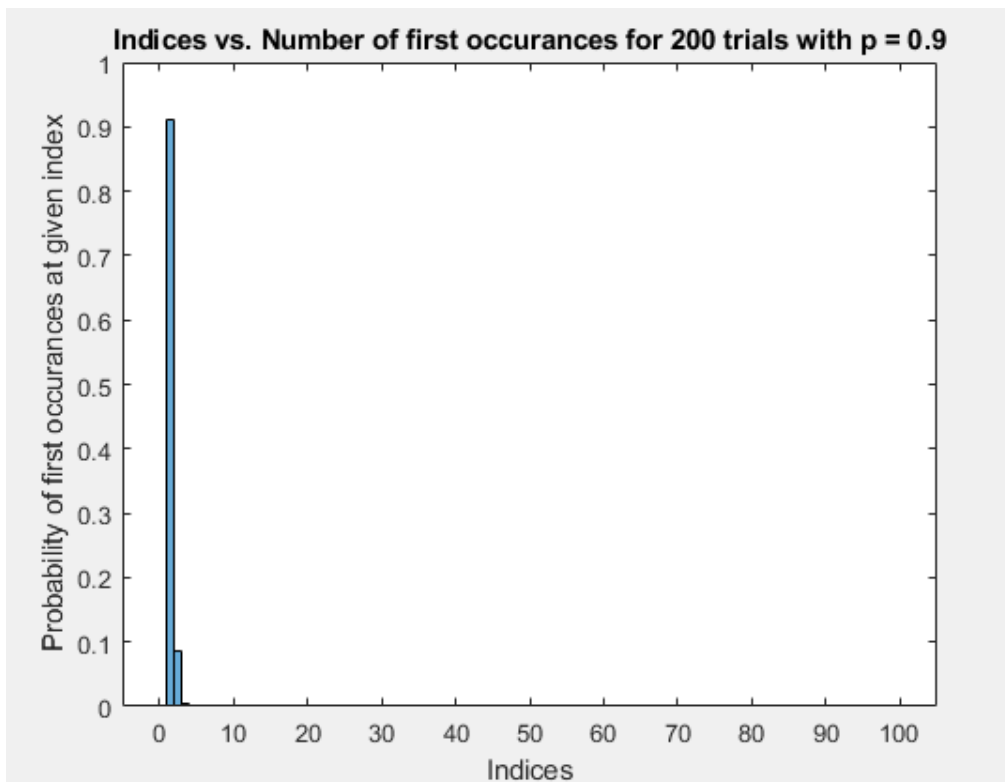
Figure 8: PMF for p1 = 0.9 at 20 trials

Figure 9: PMF for p1 = 0.1at 200 trials

Figure 10: PMF for p1 = 0.5 at 200 trials
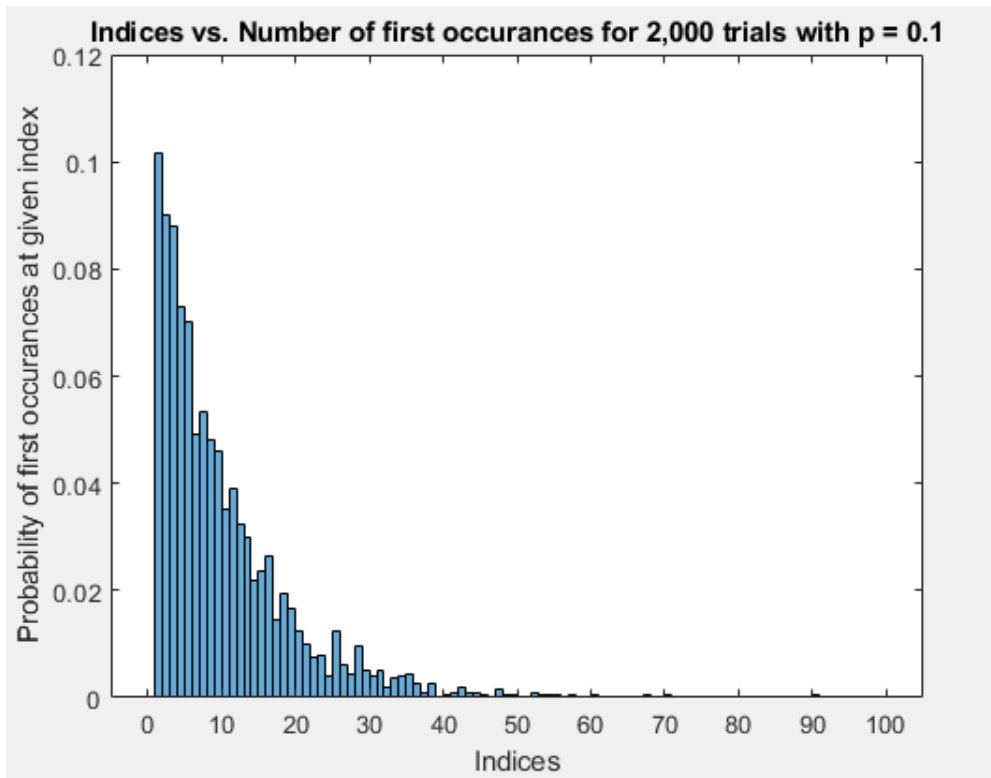

Figure 11: PMF for p1 = 0.9 at 200 trials
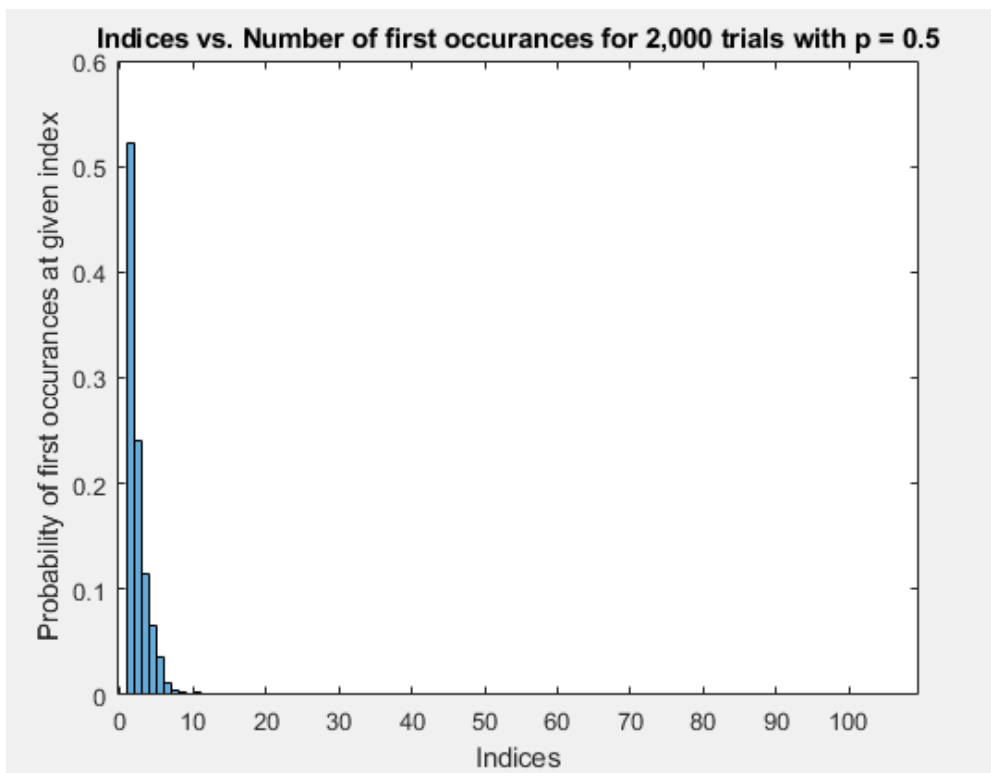
Figure 12: PMF for p1 = 0.1 at 2,000 trials

Figure 13: PMF for p1 = 0.5 at 2,000 trials
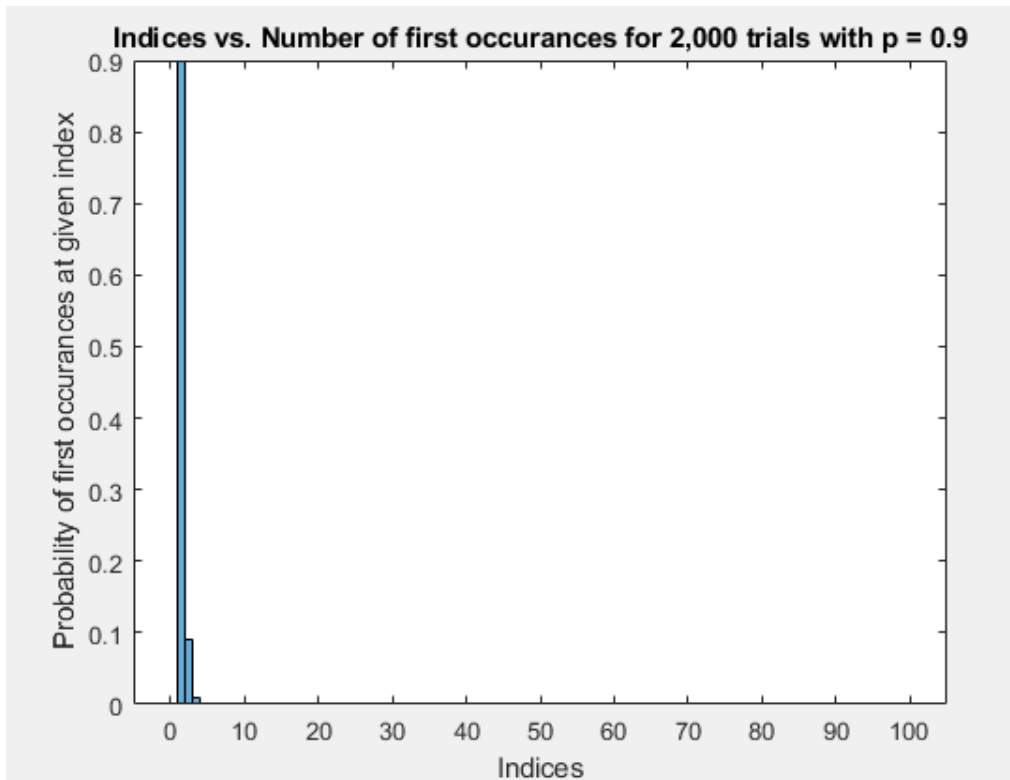
Figure 14: PMF for p1 = 0.9 at 2,000 trials

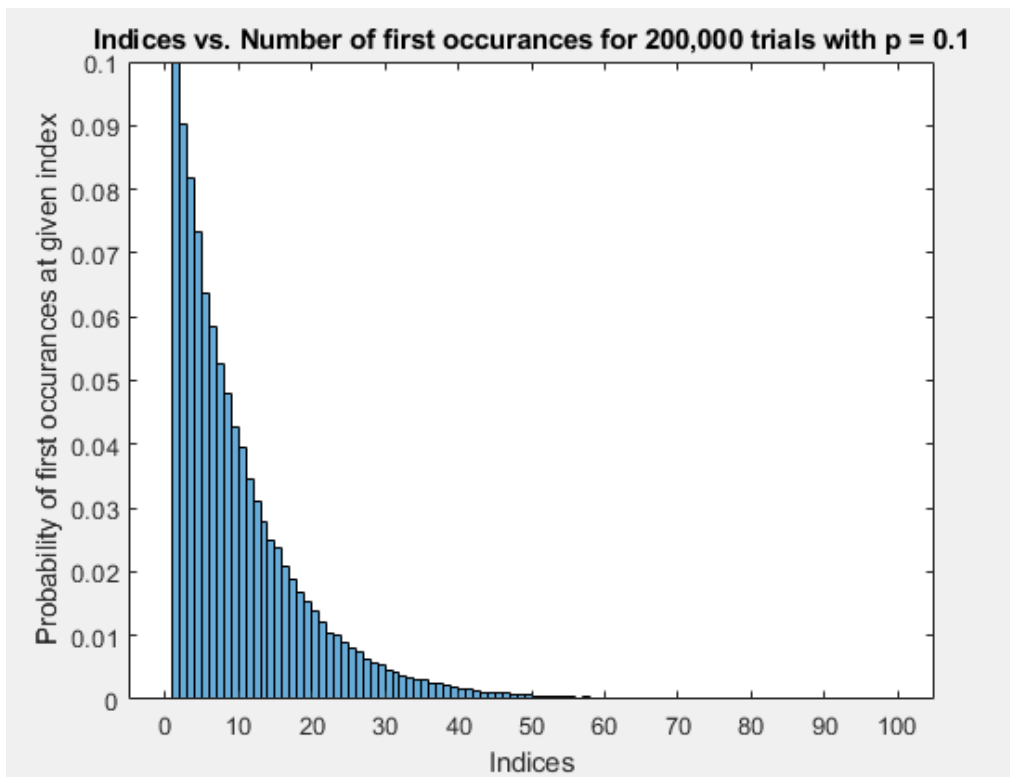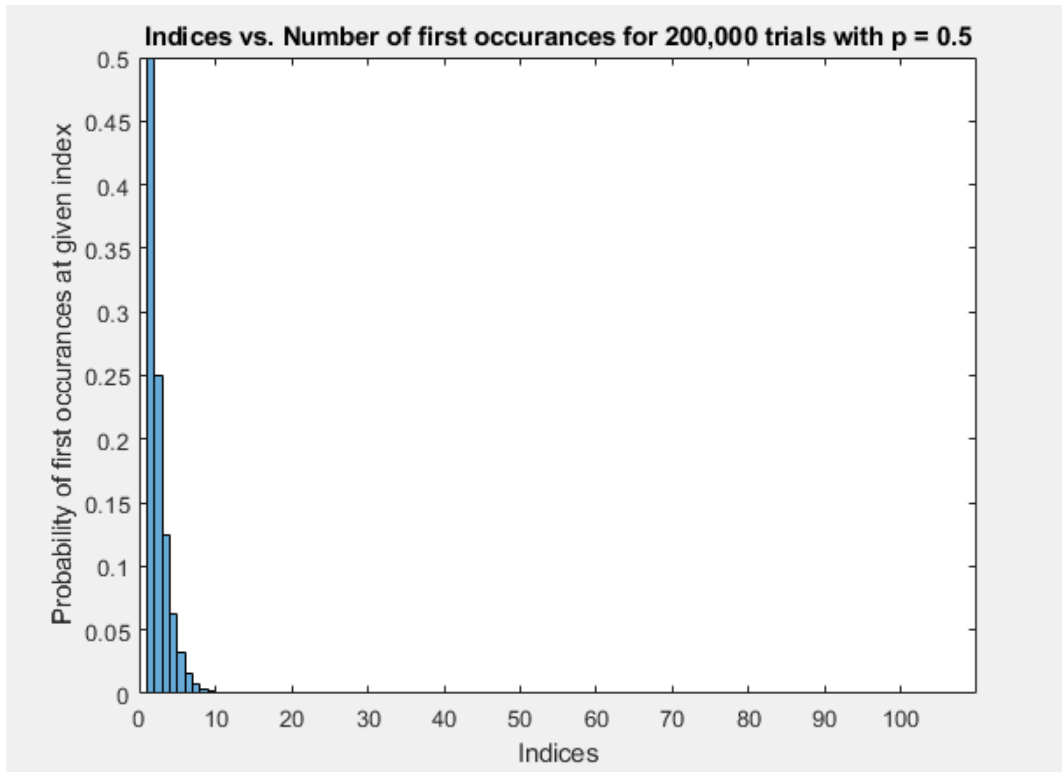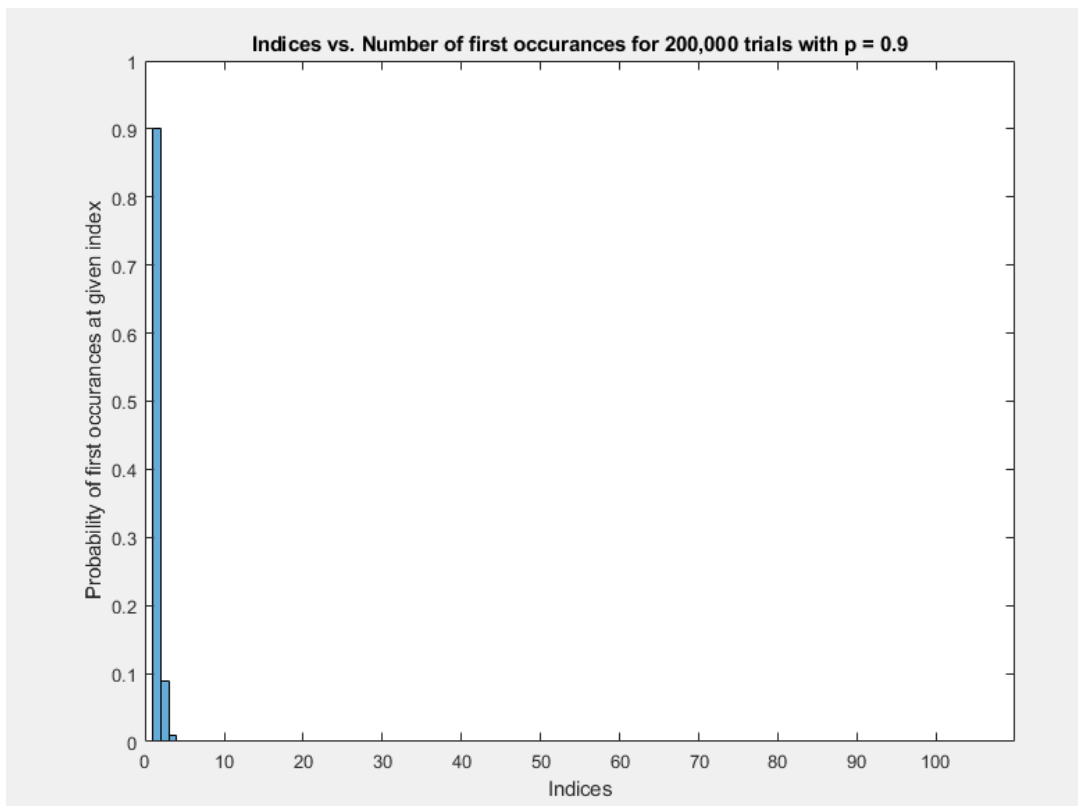**Figure 15: PMF for p1 = 0.1 at 200,000 trials**

Figure 16: PMF for p1 = 0.5 at 200,000 trials

Figure 17: PMF for p1 = 0.9 at 200,000 trials

Page: 10

The histograms each show what the PDF are for the given number of trials and the probability value for p_1. The plots take a more consistent shape as more trials are conducted. Plots that represent more trials also align with the analytical values. The analytical values for the plot are $P\{i\ (index)\} = p_1{}^i$ and this is clearly shown in the plots with 200,000 trials for each value of p_1.

```
Mean for 20 trials @ p = 0.5 = 1.5500
Variance for 20 trials @ p = 0.5 = 0.6816
Mean for 20 trials @ p = 0.9 = 1.3000
Variance for 20 trials @ p = 0.9 = 0.3263
Mean for 20 trials @ p = 0.1 = 8.8000
Variance for 20 trials @ p = 0.1 = 75.0105

Mean for 200 trials @ p = 0.5 = 2.0250
Variance for 200 trials @ p = 0.5 = 2.2557
Mean for 200 trials @ p = 0.9 = 1.0850
Variance for 200 trials @ p = 0.9 = 0.0882
Mean for 200 trials @ p = 0.1 = 8.6700
Variance for 200 trials @ p = 0.1 = 73.3981

Mean for 2,000 trials @ p = 0.5 = 2.0560
Variance for 2,000 trials @ p = 0.5 = 2.1860
Mean for 2,000 trials @ p = 0.9 = 1.1060
Variance for 2,000 trials @ p = 0.9 = 0.1118
Mean for 2,000 trials @ p = 0.1 = 10.2230
Variance for 2,000 trials @ p = 0.1 = 89.6601

Mean for 200,000 trials @ p = 0.5 = 2.0000
Variance for 200,000 trials @ p = 0.5 = 2.1100
Mean for 200,000 trials @ p = 0.9 = 1.1115
Variance for 200,000 trials @ p = 0.9 = 0.1236
Mean for 200,000 trials @ p = 0.1 = 9.9987
Variance for 200,000 trials @ p = 0.1 = 89.9654
```

Figure 18: Sample Mean and Variance for each set of trials and p1 values

The analytical values for the mean are $\mu = \frac{1}{p_1}$, $\mu\{p_1 = 0.5, 0.1, 0.9\} = \{2, 10, 1.111\}$ and the variance is $\sigma^2 = \frac{1-p_1}{p_1^2}$, $\sigma^2\{p_1 = 0.5, 0.1, 0.9\} = \{2, 90, 0.1234\}$. The sample mean and variance became consistently more aligned with the analytical values as more trials were conducted as seen in **Figure 18.**

2.3   PDF for an exponentially distributed random variable

The purpose of this section is to use the provided MATLAB function randx(n, k, lambda) to generate histograms for 10, 1000, and 100,00 independent trials for the function $f(x) = 0.5e^{-0.5x}$, $x > 0$. The plot for each set of trials are as follows:

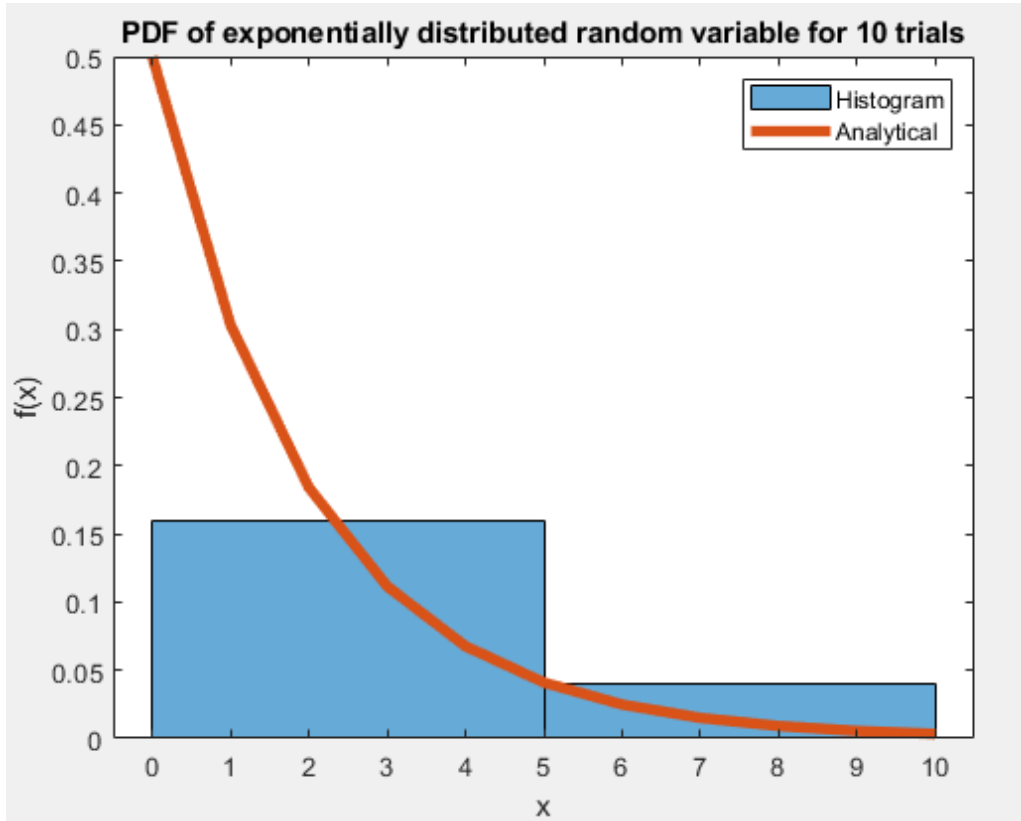
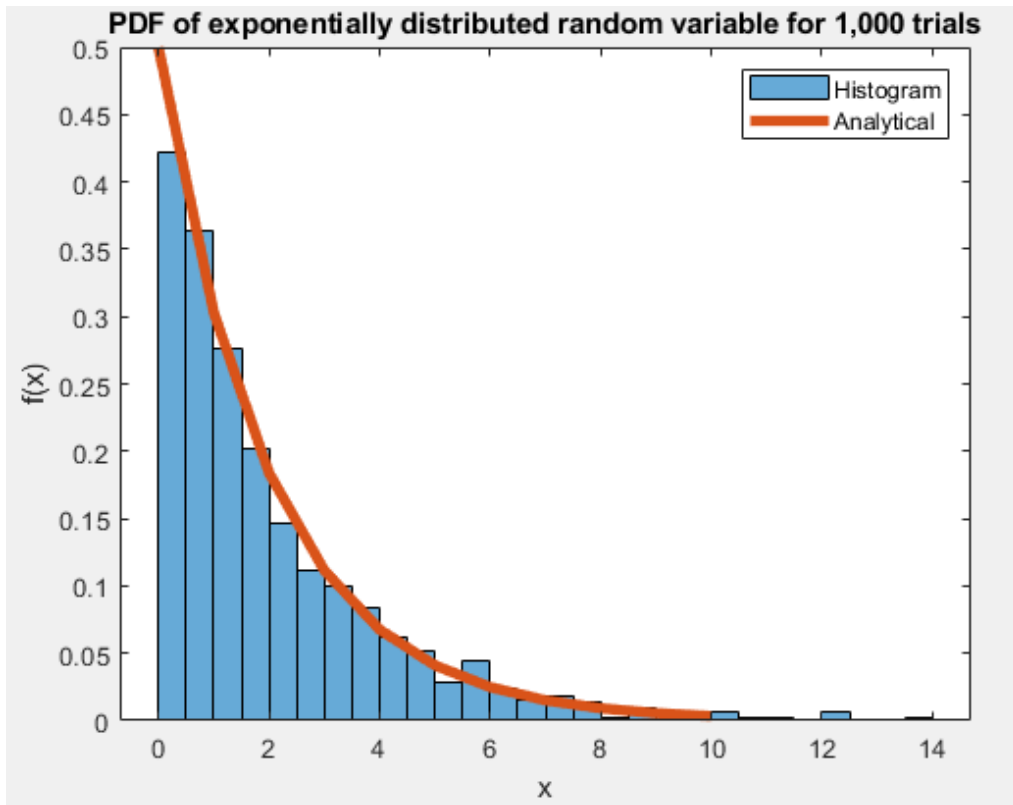**Figure 19: PDF of exponentially distributed random variable for 10 trials**

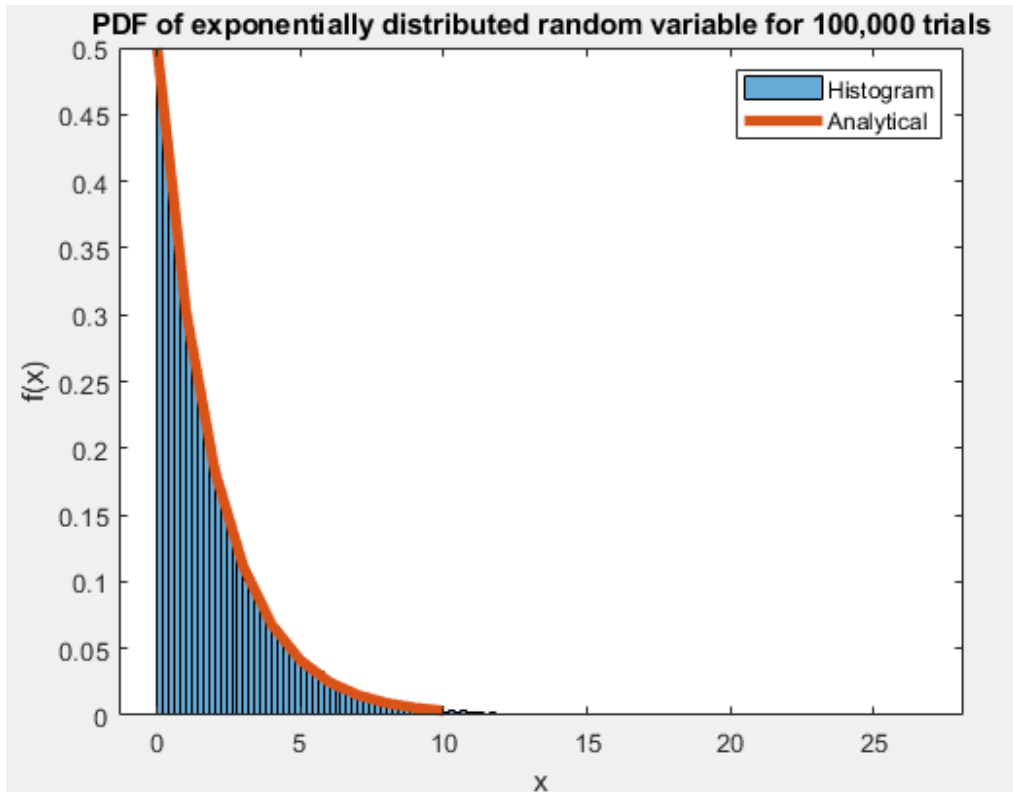**Figure 20: PDF of exponentially distributed random variable for 1,000 trials**

Figure 21: PDF of exponentially distributed random variable for 100,000 trials

These plots show the doubly scaled histograms. The scaled histograms did not match the given function of the PDF because the histogram was using the total occurrences of the values rather than the probability. The second scaling was necessary to turn the histogram into a PDF. This was done using the histogram function with 'Normalization' and 'pdf' as parameters. The sample mean and variance is as follows:

```
Mean for 10 trials = 3.5344
Variance for 10 trials = 8.5700
Mean for 1,000 trial s= 1.9068
Variance for 1,000 trials = 3.5241
Mean for 100,000 trials = 2.0008
Variance for 100,000 trials = 4.0297
```

Figure 22: Sample Mean and Variance for each set of trials

The analytical value for the mean is $\mu = \frac{1}{\lambda} = \frac{1}{0.5} = 2$ and the variance is $\sigma^2 = \frac{1}{\lambda^2} = \frac{1}{0.5^2} = 4$. The sample mean and variance become more consistent with the analytical values as more trials are conducted as shown by **Figure 21**.

2.4   PDF for normal or Gaussian distributed random variable

The goal of this section is to use the built-in MATLAB function $\mathrm{randn}(n, k)$ to generate histograms for 10, 1000, and 100,000 independent trials of a zero mean, unit variance, Gaussian or Normal random variable. The plot for the set of trials are as follows:
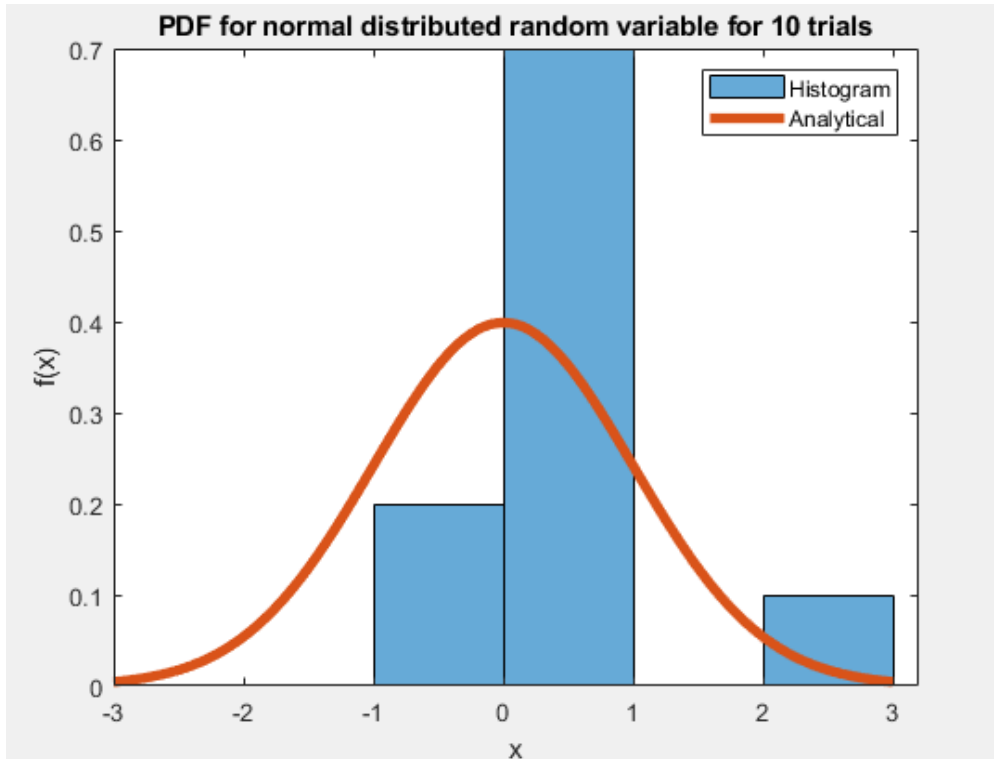
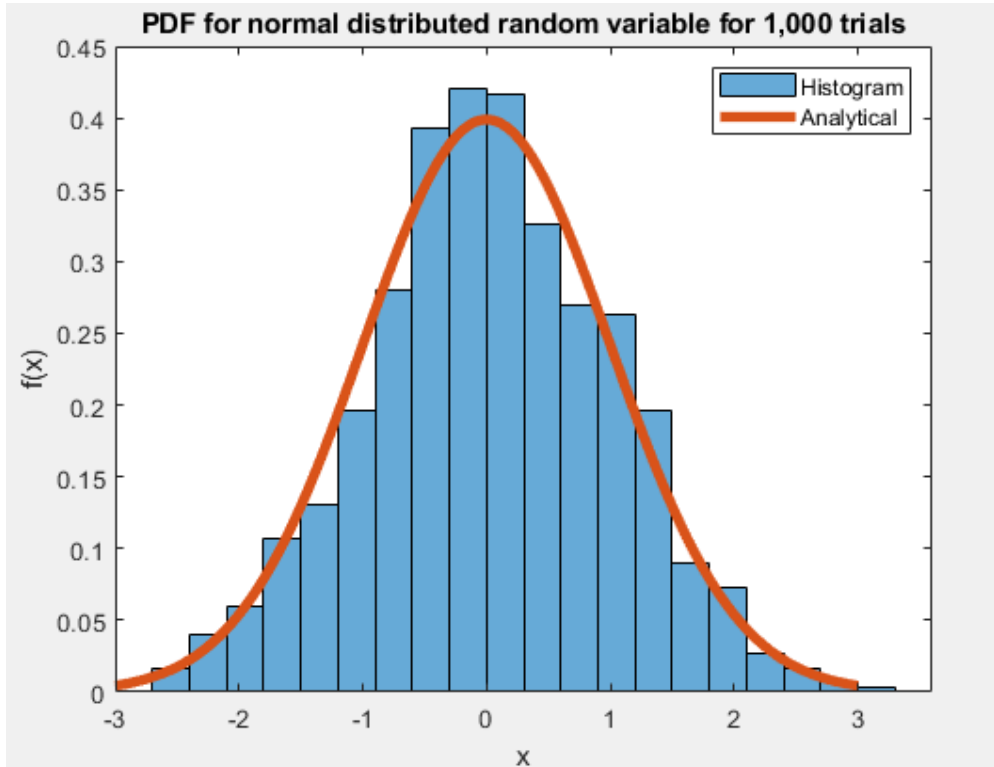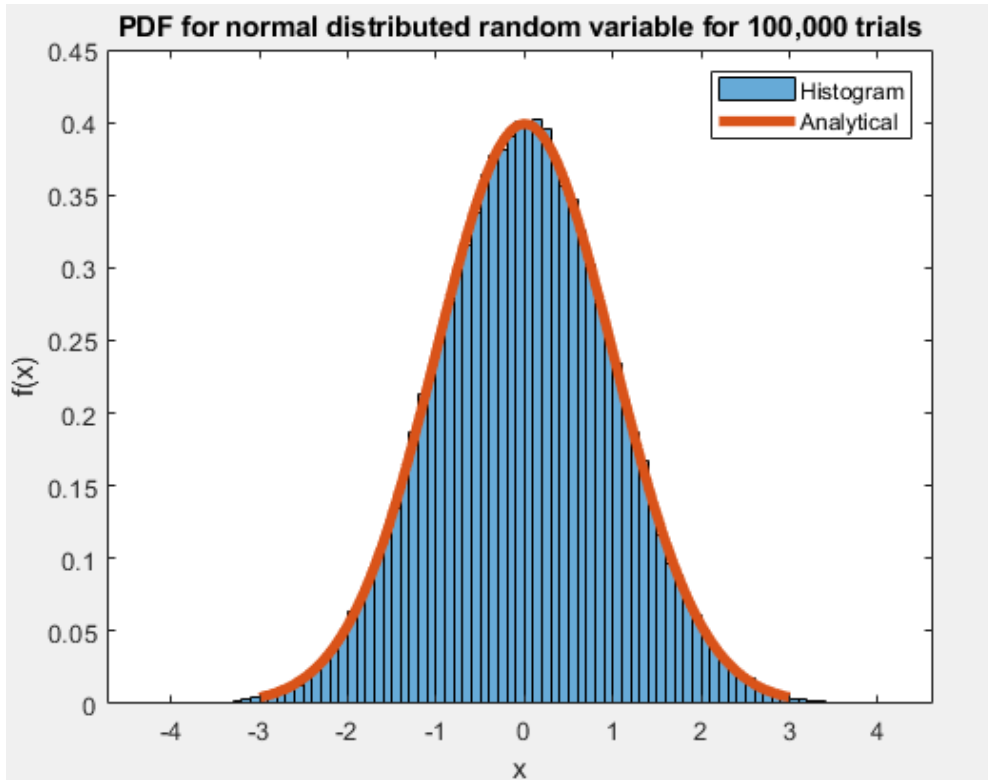**Figure 23: PDF for normal distributed random variable for 10 trials**

Figure 24: PDF for normal distributed random variable for 1,000 trials

Figure 25: PDF for normal distributed random variable for 100,000 trials

The scaled histogram plots the total occurrences of each value in their respected bin rather than the actual probability which is what the *f(x)* is plotting. To make this adjustment, additional parameters, 'Normalization' and 'pdf', had to be added to the histogram function call which changed the plot into a PDF. The sample mean and variance for the set of trials are as follows:

```
Mean for 10 trials= 0.1480
Variance for 10 trials = 1.0715
Mean for 1,000 trials= 0.0103
Variance for 1,000 trials= 1.0641
Mean for 100,000 trials= 0.0024
Variance for 100,000 trials= 0.9966
```

Figure 26: Sample Mean and Variance for each set of trials

The analytical values for the mean and variance are 0 and 1. The sample mean, and variance slightly differ from the analytical values as shown in **Figure 24**. The sample mean and variance become more consistently aligned with the analytical values as more trials are conducted.

2.5   PDF for a normal or Gaussian distributed random variable

This section is similar to section 2.4, but the sample mean, and variance differ where $m = 1$ and $\sigma^2 = 4$. To account for these changes the data from the function randn(n, k) have to be manipulated. The plot for the set of trials are as follows:
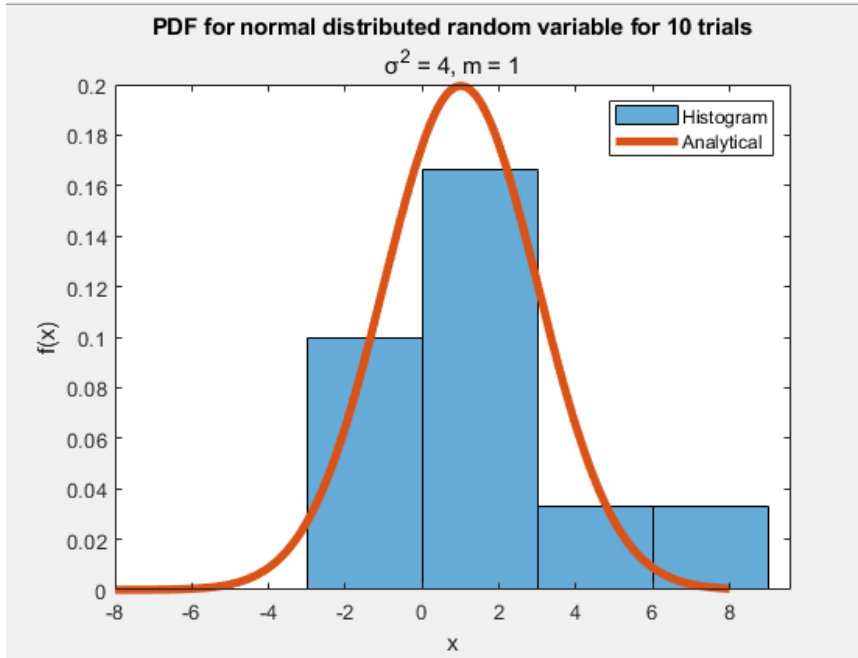
Figure 27: PDF for normal distributed random variable for 10 trials, σ^2 = 4, m = 1
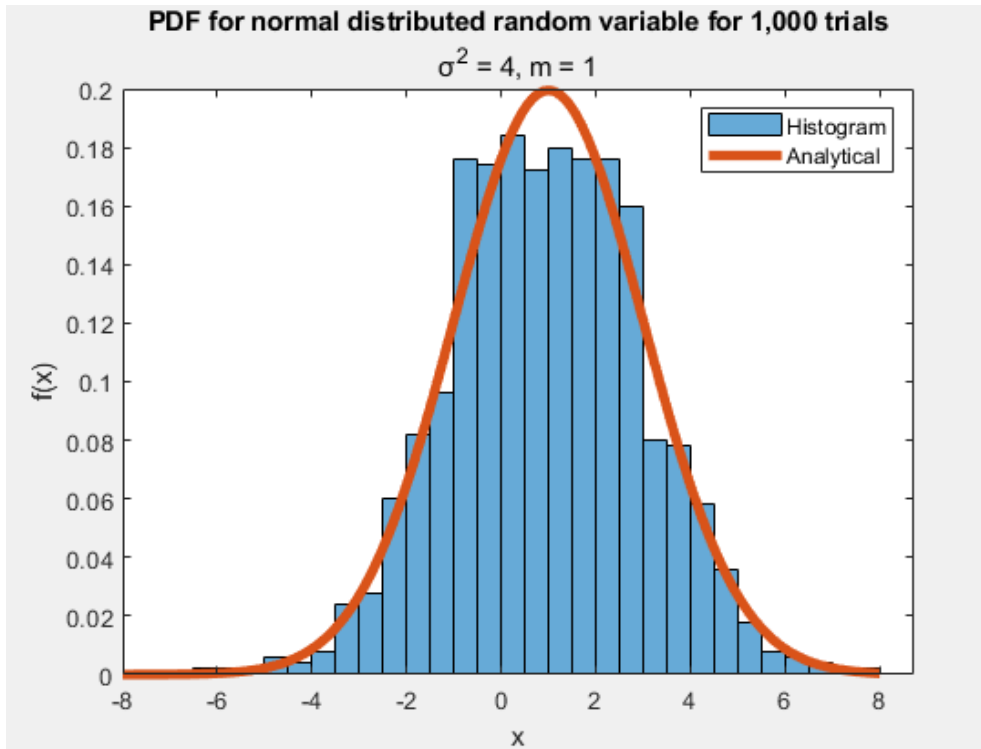
Figure 28: PDF for normal distributed random variable for 1,000 trials, σ^2 = 4, m = 1
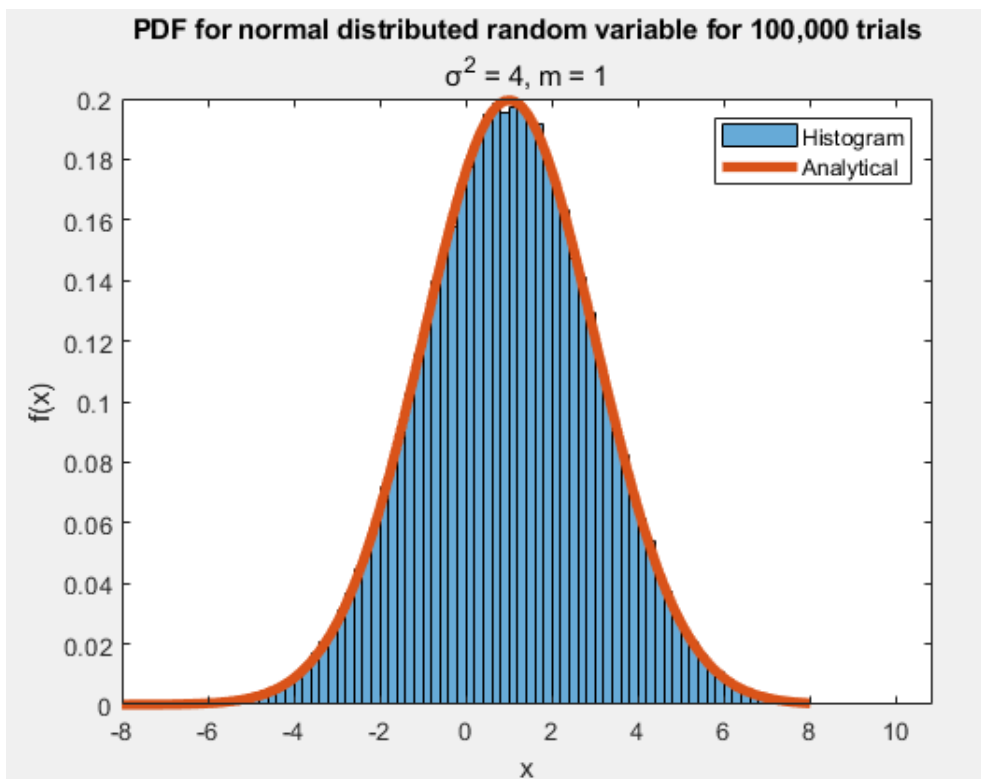
Figure 29: PDF for normal distributed random variable for 100,000 trials, σ^2 = 4, m = 1

The original scaled histograms did not follow the PDF function as it wasn't calculating probability but rather the occurrences. This was fixed using the parameters, 'Normalization' and 'pdf', in the histogram function. This changed the unscaled histograms to doubly scaled histograms and into a PDF. The sample mean and variance are as follows:

```
Mean for 10 trials = 1.4053
Variance for 10 trials = 1.7811
Mean for 1,000 trials = 1.0090
Variance for 1,000 trials = 4.2299
Mean for 100,000 trials = 1.0025
Variance for 100,000 trials = 4.0050
```

Figure 30: Sample Mean and Variance for set of trials

The analytical values for the mean and variance are 1 and 4. The sample mean, and variance slightly differ from the analytical values as shown in **Figure 28**. The sample mean and variance become more consistently aligned with the analytical values as more trials are conducted.

2.6   Computing probabilities from the pdf

The purpose of this section is to first compute the probability that the random variable falls between 1.0 and 3.0 by dividing by the total number of trials, using the unscaled histogram from the section 2.5. Then using the doubly scaled histogram, which shows the pdf, compute the sample probability that the random variable falls between 1.0 and 3.0. Finally, numerically integrate the true probability from the function $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(x-m)^2}{2\sigma^2}}$, $m = 1$, $\sigma^2 = 4$ to find the probability $\Pr\{1.0 \leq X < 3.0\}$. The calculated probabilities are as follows:

```
Unscaled Sample probability: 0.2408
Scaled Sample probability: 0.2408
True PDF probability: 0.3413
>> part2_6
Unscaled Sample probability: 0.2418
Scaled Sample probability: 0.2418
True PDF probability: 0.3413
>> part2_6
Unscaled Sample probability: 0.2419
Scaled Sample probability: 0.2419
True PDF probability: 0.3413
>> part2_6
Unscaled Sample probability: 0.2415
Scaled Sample probability: 0.2415
True PDF probability: 0.3413
>> part2_6
Unscaled Sample probability: 0.2418
Scaled Sample probability: 0.2418
True PDF probability: 0.3413
```

Figure 31: 5 trials with a histogram bin width of 2

```
Unscaled Sample probability: 0.3418
Scaled Sample probability: 0.3418
True PDF probability: 0.3413
>> part2_6
Unscaled Sample probability: 0.3412
Scaled Sample probability: 0.3412
True PDF probability: 0.3413
>> part2_6
Unscaled Sample probability: 0.3411
Scaled Sample probability: 0.3411
True PDF probability: 0.3413
>> part2_6
Unscaled Sample probability: 0.3412
Scaled Sample probability: 0.3412
True PDF probability: 0.3413
>> part2_6
Unscaled Sample probability: 0.3414
Scaled Sample probability: 0.3414
True PDF probability: 0.3413
```

Figure 32: 5 trials with a histogram bin width of 0.1

In **Figure 29**, the data shows that the bin width affects the accuracy of the calculated
sample probability as the values aren't relatively similar to the true PDF probability.

Figure 30 shows that with a bin width of 0.1 the calculated sample probability is almost identical to the true PDF probability.

3   WHAT I LEARNED

This project has really given me a much better understanding of PMFs and PDFs and how they are implemented using different number of trials. For PDFs, I also have learned how to simulate the analytical functions using sets of trials. This project has also advanced my understanding of MATLAB functions, scripts, and the different way of producing histograms and graphs.

3.1   Future suggestions and Time spent

The project overall was a little repetitive such as section 2.4 and 2.5 were nearly identical just a change of the mean and variance values. This made the project take a little longer specially typing up the report. For this project I spent a total of 30 hours, including reading, research, programming, writing, and final preparation.