CMPE 320 Project 1
DATE: 28 February 2021
TO: Professor LaBerge
FROM: Mick Harrigan
SUBJECT: Histograms, PDFs and PMFs

1 INTRODUCTION

This project illustrates the effects of a histogram in representing a probability density function (pdf or PDF) or a probability mass function (pmf or PMF) for a random variable. In addition, the uses of PMF/PDF for computation of probabilities.

The scope of the project includes many different types of PMF/PDF distributions through a variety of means. These include, uniform distribution, geometric distribution, exponential distribution, and Gaussian distribution.

2 SIMULATION AND DISCUSSION

2.1 PMF for a single fair die

This experiment uses the MATLAB function randi(imax, m, n) to model the events of rolling a fair 6 sided die for N trials. The resulting random variable is the value from the die's face, thus the output for the produced histogram is one of the values 1 through 6. Logically, this should then yield a uniform distribution across each of these values, as they are independent.

In the experiment the amount of trials done (N) was as follows, 120; 1200; 12,000; 120,000. The output histograms are shown below in Figure 1, with 1 through 4 aligning to 120 through 120,000.

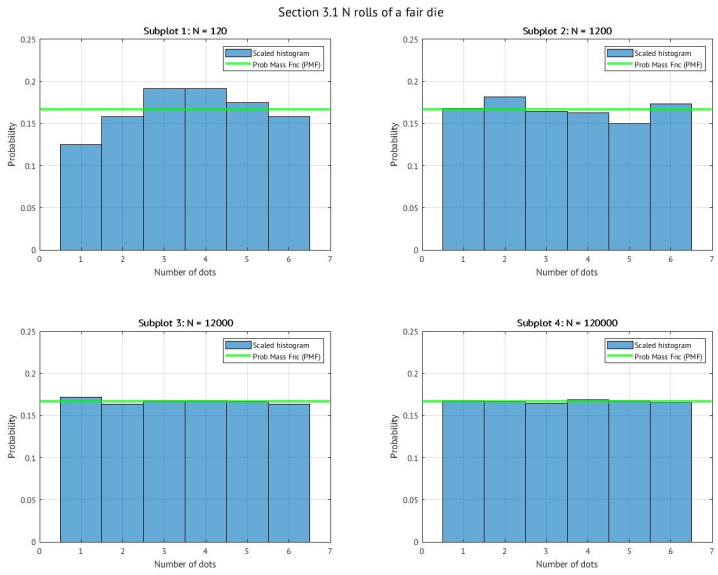

Figure 1: Uniform Distribution Results

N	Mean	Variance
120	3.7917	2.7546
1200	3.4742	3.0452
12000	3.4983	2.9204
120000	3.5007	2.9217
Theoretical	3.5	2.9167

Table 1: Mean and Variance of Section 3.1 Trials

Figure 1 shows the results of the experiment given each of the number of trials used. Each has the theoretical PMF value shown as the expected output. As N increases from plot 1 to 4 the distribution gets closer and closer to that theoretical value.

To calculate the theoretical PMF value of a uniformly distributed data set the following equation is used.

$$f_X(x_k) = Pr[X = X(s_k)] = 1/X \quad \textbf{(1)}$$

Where X is the random variable, which in this case is 6 (the sides of the die) and x is the event outcomes, which is 1 through 6.

The analytical mean of this data set is simply the mean of the available options, which yields 3.5. In addition to this, the analytical variance is 2.9167. For each value of N the calculated mean and variance should approach the value of the theoretical values. Table 1 shows these values for each trial and the phenomenon that is to be expected.

2.2 PMF for binary strings

This experiment aims to generate 100 binary values such that each bit may be either 0 or 1. This is done with different trials with different probabilities of each value being a 1. These values, called p_1, are: 0.5, 0.9, and 0.1. With these binary strings generated, the index of the first 1 is tracked for each trial. This value is then counted for each trial of N (100; 1000; 10,000) trials in a histogram.

The resulting distribution of the created histograms is geometric as the probability of the next "bit" having the first 1 value is geometrically less than that of the current index. Figures 3, 4, and 5 show this with varying trial sizes and with each probability value of 1 appearing. Projected on top of each graph is that of the theoretical PMF function of the given 1 appearance probability.

The equation that defines the theoretical geometric PMF is as follows:

$$f_X(k) = p^k(1-p) \quad \textbf{(2)}$$

Where X is the random variable, the amount of digits until 1 appears, k is the actual index of the first 1 appearing, and $f_X(k)$ is the amount of times in the trial that the first 1 was in that index (k).

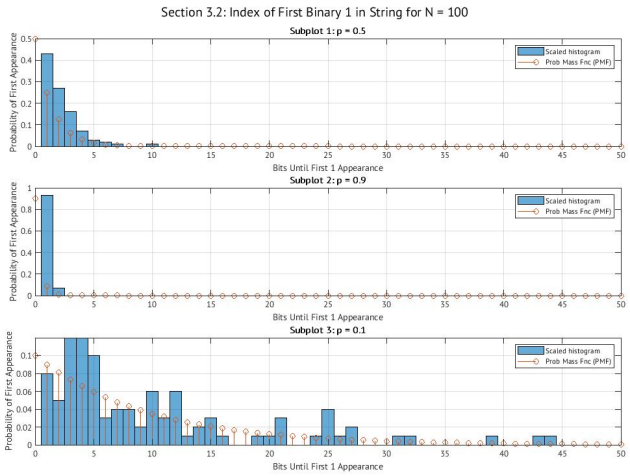

Figure 2: Smallest Trial Results

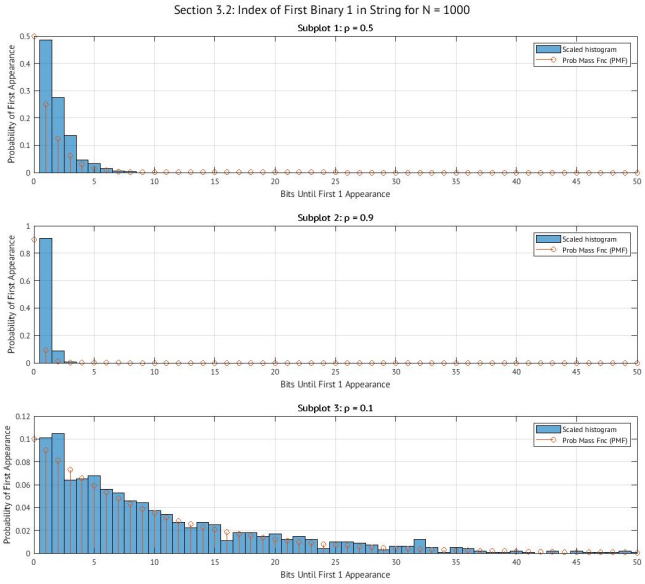

Figure 3: Larger Trial Results

N	p	Sample Mean	Sample Variance	Population Mean	Population Variance
100	0.5	2.17	2.3445	2	2
	0.9	1.07	0.065758	1.1111	0.12346
	0.1	10.1	90.9394	10	90
1000	0.5	1.952	1.6233	2	2
	0.9	1.095	0.096071	1.1111	0.12346
	0.1	10.674	113.8776	10	90
10000	0.5	2.0079	2.0868	2	2
	0.9	1.1166	0.12602	1.111	0.12346
	0.1	10.0403	91.9411	10	90

Table 2: Mean and Variance of Section 3.2 Trials

For each value of N, and each value of p, the population mean and variance were calculated using $\mu = \frac{1}{p}$ and $\sigma^2 = \frac{1-p}{p^2}$ respectively. Table 2 below shows each of these value in each case, as well as the sample mean and variance.

The output histograms trend more and more towards that of the theoretical PMF function as N increases. There are outliers to this trend, but those outliers become less common as N grows as well. On top of this, the sample mean and variance track closely to the population mean and variance, where they generally get closer as N grows as well. In the case of the sample variance of N = 1000 the difference is larger than initially expected, but this makes sense given the imperfect nature of the output histogram.

2.3 PDF for an exponentially distributed random variable.

This experiment uses the MATLAB function randx(n, k, lambda) to generate histograms in an exponential distribution for each amount of trials (N).

The resulting histograms are Figures 5,6,and 7.

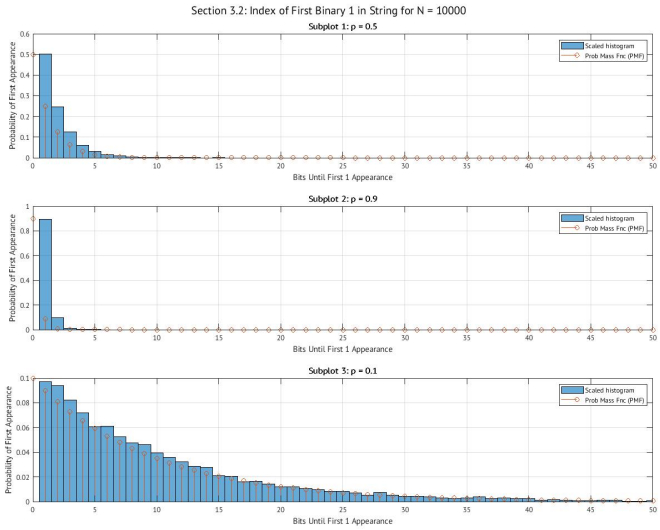

Figure 4: Largest Trial Results

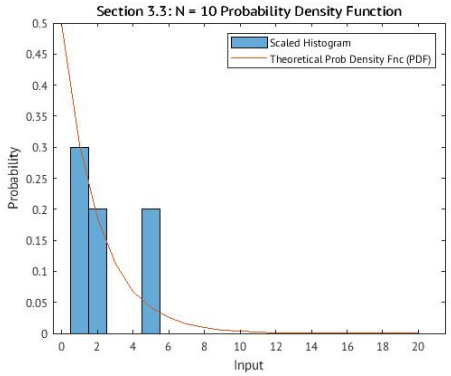

Figure 5: Exponential Curve of N = 10

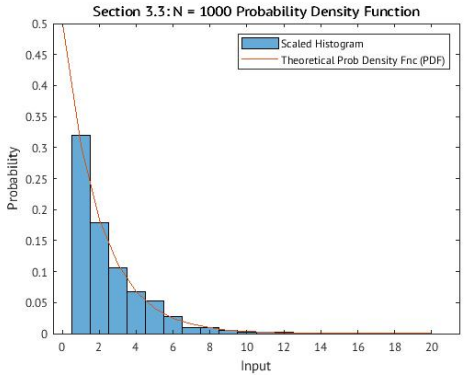

Figure 6: Exponential Curve of N = 1000

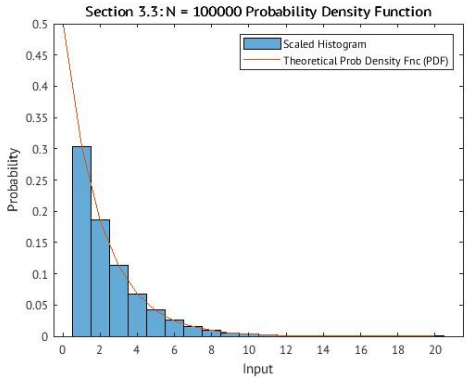

Figure 7: Exponential Curve of N = 10,0000

N	sample mean	sample variance	population mean	population variance
10	1.6542	1.1219	2	4
1000	1.9384	3.3917	2	4
100000	2.0069	3.9972	2	4

Table 3: Mean and Variance of Section 2.3 Trials

The above figures show the outputs of the different sizes of trials when using the given equation for exponential distribution. This is as follows.

$$f(x) = \lambda e^{-\lambda t}, t > 0 \quad \textbf{(3)}$$

The theoretical PDF in each of these cases is defined as the same function from above. With each case the random variables follow suit with this curve, and this is most apparent in the change of the histograms from N = 10 to N = 1000 where the curve becomes almost entirely in line with that of the theoretical output.

The scaling factor is based on a twofold factor. One of those is simply the amount of trials that is had across the entire data set. The other factor involved is more important. This is the factor that decides the dx that makes a PDF represent a specific probability. The width of the bins is this factor and by changing them to be smaller (effectively shrinking dx)it allows the density of the random variable's appearance to be more in line with a specific value's probability. Then, with this said, the scaling factor used to normalize a PDF is $\frac{1}{N*BW}$ where BW is the width of the bins. In terms of the PDF, as the bin width decreases, less can be fit within them and the histogram then shows a less dense section of the data set eventually (as the bin width gets infinitesimally small) getting to the probability of a single value. This scaling is necessary to be used as to have the histogram fit to the theoretical PDF, which is based on being able to integrate from its lowest bound to its highest and have a value of 1.

In Table 3, the sample mean and variance are compared against their analytical mean and variance. As N increases the sample's get closer and closer to the analytical values which is expected as the final output.

2.4 PDF for a unit variance normal or Gaussian distributed random variable.

This experiment utilizes the MATLAB function randn(n, k) to generate Gaussian histograms of mean = 0 and variance = 1.

The resulting figures show the outputs for the given N values.

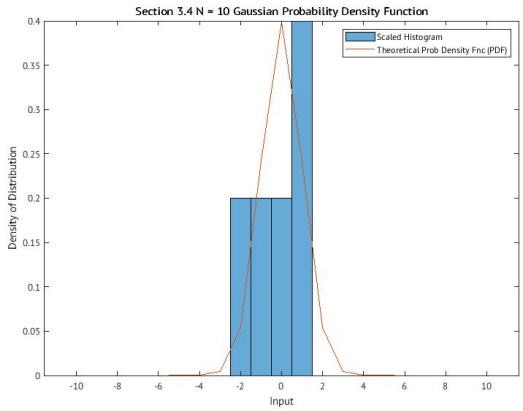

Figure 8: Gaussian Distribution for N = 10

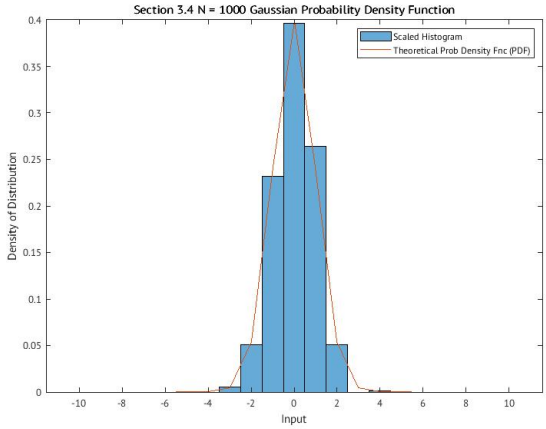

Figure 9: Gaussian Distribution for N = 1000

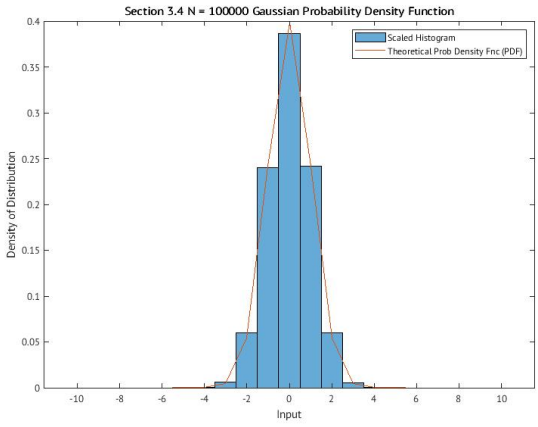

Figure 10: Gaussian Distribution for N = 100000

N	sample mean	sample variance	population mean	population variance
10	-0.17784	1.5322	0	1
1000	0.014481	0.90211	0	1
100000	-9.88E-05	0.99298	0	1

Table 4: Mean and Variance of Section 2.4 Trials

The above figures (8, 9, 10) show the Gaussian distribution formed from the equation as follows.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2}}$$

(4)

The theoretical output is shown on top of the actual results. In the first case (Figure 8) the actual values are very much skewed in a way that wouldn't be instantly seen as Gaussian, but changes quickly in Figure 9 where the theoretical is almost perfectly followed in the actual distribution. This trend follows again in Figure 10, where the same thing applies.

The scaling factor that creates Figures 8 through 10 is shown by $\frac{1}{N*BW}$ where the bin width is represented by BW. In terms of the change from a PMF to a PDF, this is the step where the size of the bins affects the output distribution and is defined by the same dx as in Section 3.2. This scaling is required to force the histogram to fit within a size constraint of 1. If the PDF that defines these outputs is to be integrated, the output of that will be 1.

Above, in Table 4, the mean and variance of the 3 trials of this Gaussian distribution are shown. The error between sample and population shrinks as N gets larger and larger, as expected.

2.5 PDF for a normal or Gaussian distributed random variable.

This experiment shows the outputs of a Gaussian distribution but with the added change of not being the unit Gaussian function. The mean defined for this distribution is -2 and variance defined as 9. The MATLAB function used before, randn(n, k) was modified to fit this change in generation of values.

The outputs for each N value are shown here as the below Figures.

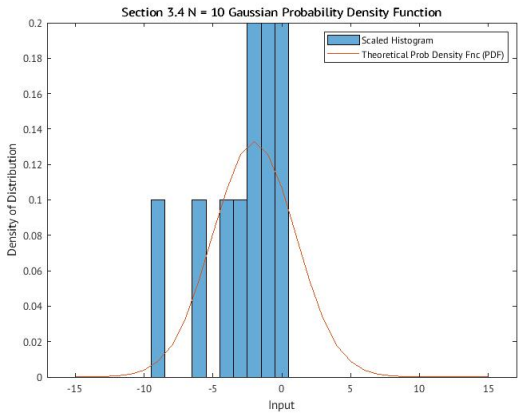

Figure 11: Gaussian Distribution with Mean = -2 and Variance = 9 for N = 10

Figures 11, 12, and 13 illustrate the Gaussian distribution for the given mean and variance. As the amount of trials increases, the outputs more and more commonly follow the theoretical function shown using the below equation.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2}}$$

(5)

The theoretical output is shown on top of the actual output in each of the test trials. When the amount of trials is smaller there greater error and less following the theoretical curve. This is changed when N = 100,000 as it follows the theoretical output as intended. This is expected as when there are more trials the outcome more closely follows that of the theoretical.

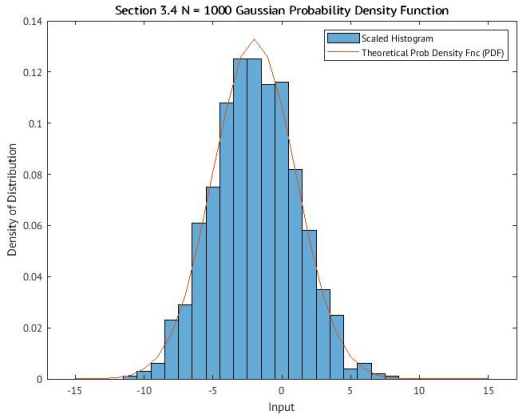

Figure 12: Gaussian Distribution with Mean = -2 and Variance = 9 for N = 1000

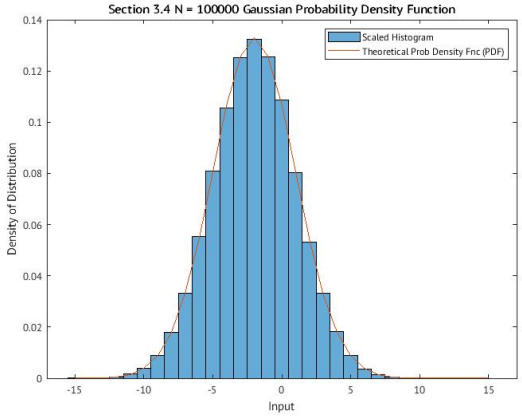

Figure 13: Gaussian Distribution with Mean = -2 and Variance = 9 for N = 100000

N	sample mean	sample variance	population mean	population variance
10	-2.847	8.2799	-2	9
1000	-1.9062	9.1344	-2	9
100000	-2.01E+00	8.9455	-2	9

Table 5: Mean and Variance of Section 2.5 Trials

Raw Probability	Normalized Probability	Numerical Integration
0.67436	0.67436	0.68269

Table 6: Probabilities Calculated via Different Methods

As in Section 3.3 and 3.4, the scaling of this section is defined by $\frac{1}{N*BW}$ where BW is the width of the bins in the histogram. In this case the bins are of width 1. This is tied to the PMF to PDF transformation as the change in the bin widths is effectively changing the dx in the probability of a certain bin. As the bin is widened, the dx is also widened, thus increasing the density of that bin, changing the PDF that the data is meant to display. This then shows that a PDF is a continuous representation of the density of the probability, and by shrinking the dx term, the density more closely follows as the probability itself for a continuous function. In this, and similar, data sets this scaling is required to be able to get the density of probability for a specific range of values. In Figures 11, 12, and 13 this is seen as there is a higher concentration of values within the "center" bins and much less around the edges of the same distribution.

Above in Table 5 the expected outcome is the values of sample mean and variance approaching more closely the values of the population mean and variance. This is seen as the error gets smaller and smaller with each subsequent iteration.

2.6 Computing probabilities from the pdf

This experiment finds the specific probability of the random variable being within -5 and 1. This can be done by summing the bins between -5 and 1, then scaling them using the the total trials, N = 100,000. Then, a similar process is to be done with the normalized histogram, where the normalized values are summed to reach a probability of the range. Finally, if a numerical integral is taken from -5 to 1, the theoretical value can be found. This uses the below equation.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \quad \textbf{(6)}$$

The values obtained from these methods are shown above in Table 6. These results make sense given the method that is used to normalize the histogram. Normalization is dividing by the total number of trials run for each bin to get the density of probability for that specific bin. If these bins were increased in size, the densities would increase alongside that until there is less detail shown and the shape could potentially be lost as well. On the other hand, if they are shrank too much they could lose shape as well as there would be enough bins to fit each and every possible value that would be guessed, with the density of each bin being very low.

3 WHAT WAS LEARNED

The sections of this project showed the differences in what a PMF and PDF are, as well as what the actual distributions are when actually represented with real data. The largest phenomenon that was recorded was the notion that as trials for a specific experiment increase, the distribution is more and more closely resembling that of the theoretical output. This was expected, but with the actual randomness of each variable it was able to be shown very quickly that this phenomenon is true.

This project as a whole helped myself to understand the basics of Statistics and how to begin understanding the more in depth things to come later in this semester. The biggest thing that is learned from this is the actual conformity to a specific writing style and completeness of it.

Most of the critiques that I would have for this project comes down to clarity. The issues that I ended up running into came from vagueness in either the skeleton code, project document having multiple interpretations, or typos and other small issues. Overall, I do not believe that this is a bad project, in fact I do believe that in the process of doing this project I have learned very much in the scope of the class.

The time that was spent on this project from beginning of the skeleton code to the end of this report is somewhere in the neighborhood of 15-16 hours. This is an over-estimate but shows the general amount of work needed to try and iron out the issues that could be found as soon as possible. A breakdown of this time length is about 10 hours on the code, with about 5 hours spent on writing the report. Some of the time on the report could be saved through personal use of different tools to construct it.