MINI PROJECT 2

APMA 3100

By Winston Zhang, Luke Mathe

22. April 2022

TABLE OF CONTENTS, FIGURES AND TABLES

TABLE OF CONTENTS, FIGURES AND TABLES 1

MODEL ANALYSIS 3

Probability Density Function fX(x) 3

Figure 1: Graphical Representation of PDF fX(x) 3

Table 1: Tabular Representation of PDF FX(x) 4

Cumulative Distribution Function FX(x) 4

Figure 2: Graphical Representation of CDF fX(x) 5

Table 2: Tabular Representation of CDF FX(x) 5

Three Circles 6

Table 3: Circle Radii Corresponding to the Specified Probabilities 6

Figure 3: Circles Corresponding to p, with Radius xp 7

Explanation/Analysis 7

EXPERIMENT - LAW OF LARGE NUMBERS 7

Monte-Carlo Simulation Algorithm 7

Table 4: Parameter Definitions for Linear Congruential Random Number Generator 8

Table 5: Values Generated by Linear Congruential Random Number Generator 8

Simulation 8

Table 6: Realizations of X generated by u1, u2, u3 9

Calculation 9

Table 7: First Estimation of Sample Mean for Each Sample Size 10

Visualization 10

Figure 4: Sample size vs. Estimator of Mn 10

Interpretation 10

Recommendation 11

Estimation 11

EXPERIMENT - CENTRAL LIMIT THEOREM 11

Sample Preparation 11

Table 8: First Estimations Generated for Each Sample Size n 12

Analysis 12

Table 9: Moments of Samples for Each Sample Size n 12

Table 10: Probabilities of Standardized Random Variable Zn from Mn 13

Figure 5: Standard Normal CDF with Empirical CDF Points, Intervals 14

Summarization 14

Figure 6: Standard Normal CDF with n = 3 Points 14

Figure 7: Standard Normal CDF with n = 9 Points 15

Figure 8: Standard Normal CDF with n = 27 Points 15

Figure 9: Standard Normal CDF with n = 81 Points 15

Table 10: Estimates vs. Population Values 16
Table 11: Absolute Difference 16
Conclusion **16**
Reflection **17**
HONOR PLEDGE **17**

1. MODEL ANALYSIS

The objective of this section is to gain an understanding of the drop error model. For background, the model being analyzed is the distance X between the intended drop point T of a newspaper via unmanned drone and the actual drop point A. Point T acts as the origin of a Cartesian coordinate plane, with the coordinates of point A being (Y_1, Y_2), where Y_1, Y_2 are assumed to be independent and identically distributed Gaussian random variables with mean 0, variance τ^2 (found to be $\tau = 57$ inches in prior experimental flights). X is then defined as:

$$X = \sqrt{Y_1^2 + Y_2^2},$$

the Rayleigh distribution with scale parameter $a = 1/\tau$. The following subsections are to visualize and analyze the model in question.

1. Probability Density Function $f_X(x)$

The PDF of X is defined as:

$$f_x(x) = a^2 x e^{-\frac{1}{2}a^2 x^2}, \quad x > 0$$

$f_X(x)$ is visualized in graphical form in Figure 1:

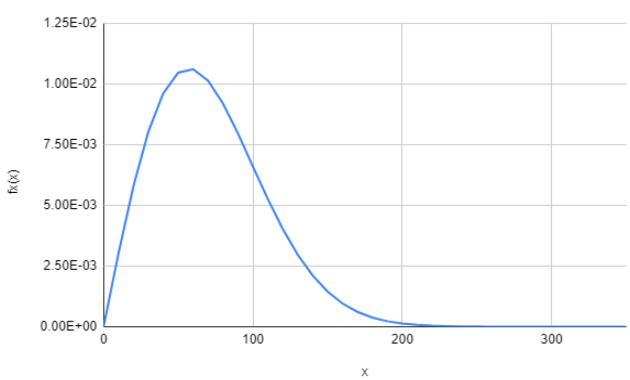

Figure 1: Graphical Representation of PDF $f_X(x)$

This was done by first entering the expression into a Texas Instruments TI-84 calculator to get an idea of the range of values that should be used, which was found to be the interval (0, 300). The calculator plot was then verified creating a Google Sheets spreadsheet and entering the same expression for the specified interval of input values x. X values fed into the formula were autofilled by Google sheets in increments of ten. The first five rows of the spreadsheet are listed in Table 1:

x	**$f_X(x)$**
0	0
10	0.0030
20	0.0058
30	0.0080
40	0.00968
50	0.0105
…	…

Table 1: Tabular Representation of PDF $F_X(x)$

2. Cumulative Distribution Function $F_X(x)$

The CDF of X is defined as:

$$F_x(x) = 1 - e^{-\frac{1}{2}a^2x^2}, \quad x > 0$$

$F_X(x)$ is visualized in graphical form in Figure 2:

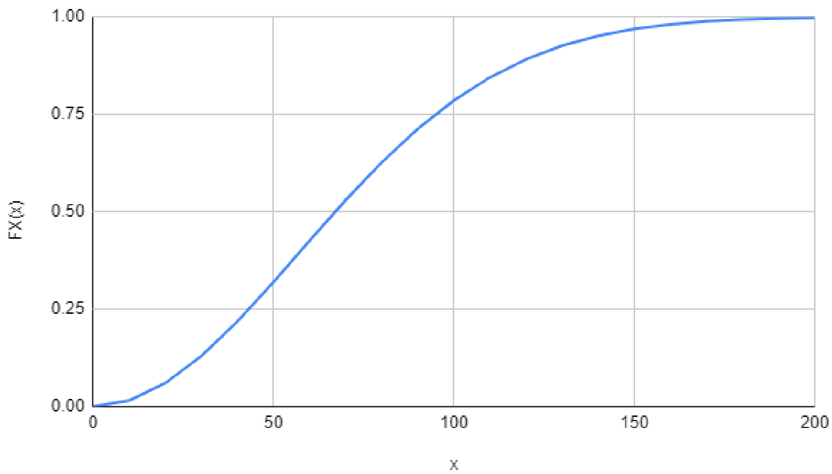

Figure 2: Graphical Representation of CDF $f_X(x)$

Mirroring what was done for the Probability Density Function, a Texas Instruments TI-84 calculator was used to find the interval of input values (0, 300). Then a spreadsheet was used to feed these values into the formula in increments of ten. As such, the first five rows of the spreadsheet are listed in Table 2:

x	$f_X(x)$
0	0
10	0.0153
20	0.05970
30	0.1293
40	0.2183
50	0.3194
…	…

Table 2: Tabular Representation of CDF $F_X(x)$

The moments for X, given scale parameter $\alpha = \frac{1}{57}$, are defined as:

$$\mu_X = \frac{1}{\frac{1}{57}}\sqrt{\frac{\pi}{2}} = 71.4359 \text{ inches,}$$

$$\sigma_X^2 = \frac{4-\pi}{2\frac{1}{57}^2} = 1394.4827 \text{ inches}$$

3. Three Circles

Three circles were plotted, centered at a point T, each having radius x_p such that:

$$P[X \leq x_p] = p, \quad p \in \{0.5, 0.7, 0.9\}$$

The values for radii x_p were determined by deriving an inverse CDF $F_X(x)$:

$$x_p = F_X^{-1}(p) = \sqrt{-6498ln(1-p)}, \quad 0 \leq p \leq 1$$

The values x_p were then determined by using the values of p as inputs to the inverse CDF. Additionally, a second method of determining the radii was used as a sanity check. The spreadsheet originally used to plot the CDF had its input values expanded from increments of 10 to increments of 0.25. Values of x were chosen from the spreadsheet that corresponded to the CDF values closest to those specified. These x_p values are enumerated in Table 3:

$P[X \leq x_p] = F_X(x_p)$	$x = F_X^{-1}(p)$	x, found by spreadsheet
0.5	67.1124	67.25
0.7	88.4501	88.5
0.9	122.3201	122.5

Table 3: Circle Radii Corresponding to the Specified Probabilities

With these radius values determined, the three circles of corresponding radius were plotted in a free online plotting tool called Geogebra. The circles are shown in Figure 3:

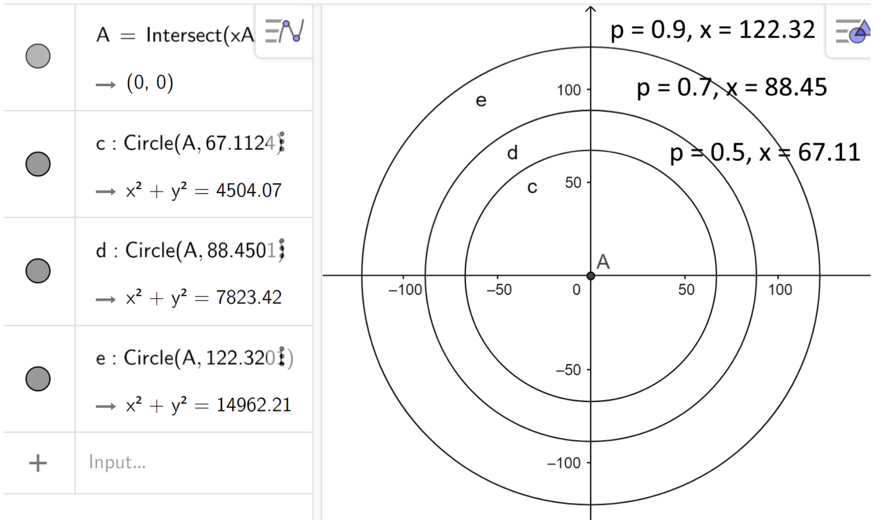

Figure 3: Circles Corresponding to *p*, with Radius x_p

4. **Explanation/Analysis**

   The circles shown in the figure correspond to the likelihood that the newspaper will land in relation to the specified drop point. In other words, there is a 50% chance the newspaper will land within 67.11 inches of the drop point, 70% chance it lands within 88.45 inches, and 90% it lands within 122.32 inches, meaning our drones are 90% accurate within about a ten foot radius around the drop point, making them perfect for leaving newspapers in a large open area such as the front or backyard.

2. EXPERIMENT - LAW OF LARGE NUMBERS

The objective of this section is to empirically demonstrate the convergence of the sample mean M_n to the population μ_X when the sample size gets sufficiently large enough - this property is known as the Law of Large Numbers.

1. **Monte-Carlo Simulation Algorithm**

   The simulation algorithm to generate realizations of random variable X will be the same developed in the telemarketing model used in Mini-Project 1. The same method of generating pseudo-random numbers via linear congruential random number generator was used, albeit several parameters were changed. The random number generator was implemented in Java. The parameters for this random number generator are listed in Table 4:

Starting value (seed)	$x_0 = 1000$
Multiplier a	$a = 24\,693$
Increment c	$c = 3967$
Modulus K	$K = 2^{18}$

Table 4: Parameter Definitions for Linear Congruential Random Number Generator

The random number generator was run to output a few values to ensure correctness, enumerated in Table 5, with expected outputs given by the assignment instructions highlighted in boldface:

Pseudo-Random Number u_i	**Value**
$\boldsymbol{u_1}$	**0.2115**
$\boldsymbol{u_2}$	**0.4113**
$\boldsymbol{u_3}$	**0.8275**
u_{51}	0.1995
u_{52}	0.2001
u_{53}	0.0469

Table 5: Values Generated by Linear Congruential Random Number Generator

2. **Simulation**

The simulation was run in a looping fashion in Java, with a seed value passed in as a parameter. To maintain independence of relizations, this value was incremented by 1 upon every iteration. The algorithm began with a value of 51, as then the linear congruential random number generator would generate the pseudo random number u_{51}, which would then be used to generate a realization of Rayleigh random variable *X*. Thus the first three realizations generated by the algorithm used u_{51}, u_{52}, u_{53}, respectively, were used to calculate the first sample mean of 110 for sample size n = 10. This was done by generating a realization of *X* and adding it to an array, then finding the summation of all array elements and dividing it by the sample size. The realizations of *X* generated by pseudo-random numbers u_1, u_2, u_3 are enumerated in Table 6:

Pseudorandom Number u_i	**Realization of *X* generated**
u_{51}	38.02505
u_{52}	38.08488
u_{53}	17.6706
…	…

Table 6: Realizations of *X* generated by u_1, u_2, u_3

As previously stated, by the iterative nature of the algorithm in repetitively generating realizations of *X*, the *x* values found by inputting u_{51}, u_{52}, u_{53} into the inverse CDF were all used in the first sample mean of sample size n = 10.

3. Calculation

This process of generating 110 sample mean estimates m_n was repeated for every sample size, and all 770 estimates were output to the terminal. The first sample mean generated for each sample size n is enumerated in Table 7:

n	First estimation of m_n
10	66.0899
30	70.3486
50	64.0847
100	73.5822
250	71.6719
500	70.2169
1000	71.1937

Table 7: First Estimation of Sample Mean for Each Sample Size

4. **Visualization**

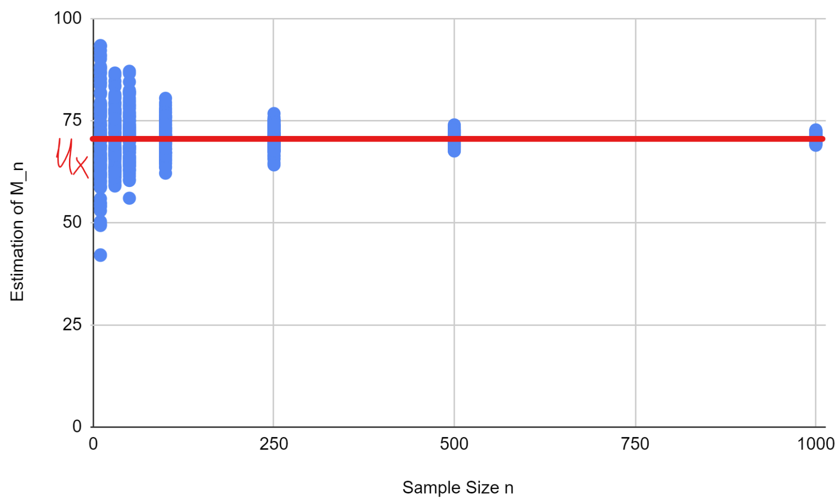

Figure 4: Sample size vs. Estimator of M_n

5. **Interpretation**

The graph demonstrates the strong law of large numbers, as n approaches infinity, the probability of the sequence of sample means M_n converging to the population mean μ_X is equal to one. We can see that as the sample size increases, the groupings of estimators get tighter and tighter, converging on the population mean.

6. Recommendation

We recommend a sample size, n^*, of 250, which would minimize the number of trials needed, while also following the weak law of probability. The sequence of probabilities of the distance between M_n and μ_X being less than 10 inches would approach 1 with a number of trials equal to 250, as seen in the distance of the farthest points from the line in Figure 4 being less than 10 inches.

7. Estimation

The probability, p, is evaluated as:

$$1 - \frac{Var(M_n)}{c^2} = 1 - \frac{(Var(X)/n^*)}{c^2} = 1 - \frac{Var(X)/250}{100}, \text{ where}$$

$$\text{Var(X)} = \frac{4 - \pi}{2\left(\frac{1}{57}\right)^2} = 1394.48.$$

$$\text{Therefore, } p = 1 - \frac{5.57}{100} = 0.9443.$$

3. EXPERIMENT - CENTRAL LIMIT THEOREM

The objective of this section is to empirically demonstrate the convergence $Z_n \rightarrow Z$ in distribution. When an empirical CDF $\hat{F}_n$ of Z_n is constructed for each sample size n, and the sequence of functions of $\hat{F}_n$ is examined, we hope to realize the distance between $\hat{F}_n$ and Φ decreases, as n increases.

1. Sample Preparation

This process of generating 550 sample mean estimates m_n was repeated for every sample size, and all 2200 estimates were output to the terminal. The various samples of estimates m_n for sample size n were used to calculate estimates of the mean and variance of M_n, which are enumerated in Table 8:

n	First estimation of M_n
3	68.2757
9	75.7005
27	58.4087
81	72.8698

Table 8: First Estimations Generated for Each Sample Size n

2. **Analysis**

Means and variances were calculated from each sample of varying size, following the formulas given in the assignment description. The mean and variance corresponding to each sample of size n is listed in Table 9:

Sample size n	Estimation of mean of M_n	Estimation of variances of M_n
3	71.6404	450.4022
9	71.6587	172.6694
27	70.5556	51.5794
81	71.4962	18.9200

Table 9: Moments of Samples for Each Sample Size n

All of the estimates m_n were stored in a Java array, and using the above calculated moments, were transformed into z-scores through use of an iterative loop. The z-scores were also stored in an array data structure, which was used to calculate the probabilities of the events:

$$\widehat{F}_n(z_j) = P[Z_n \leq z_j], \text{ for j = 1, …, 7,}$$

where the set $\{z_1, \ldots, z_7\}$ = {-1.4, -1.0, -0.5, 0, 0.5, 1.0, 1.4}

The probabilities of each event was calculated by counting how many z scores satisfied the inequality and dividing by the sample space. Table 10 lists off the various probabilities for each sample size:

n	P[Z ≤ -1.4]	P[Z ≤ -1.0]	P[Z ≤ -0.5]	P[Z ≤ 0]	P[Z ≤ 0.5]	P[Z ≤ 1.0]	P[Z ≤ 1.4]
3	0.0691	0.1618	0.3291	0.5182	0.6909	0.8436	0.9
9	0.08	0.1455	0.3218	0.5200	0.700	0.8345	0.9145
27	0.07636	0.1727	0.3164	0.4964	0.7000	0.8291	0.9218
81	0.0745	0.1564	0.3164	0.5091	0.6945	0.8291	0.9073

Table 10: Probabilities of Standardized Random Variable Z_n from M_n

The goodness-of-fit of the standard normal CDF Φ to the empirical CDF $\widehat{F}_n$ was evaluated in terms of the maximum absolute difference:

$$MAD_n = max_{1 \le j \le 7} \left| \widehat{F}_n(z_j) - \Phi(z_j) \right|$$

A figure was then created with the points $\{(z_j, \widehat{F}_n(z_j)) : j = 1, \ldots, 7\}$, with the MAD_n as highlighted intervals of probability at each point, plotted over the standard normal CDF Φ over the domain (-2.5, 2.5) in Figure 5:

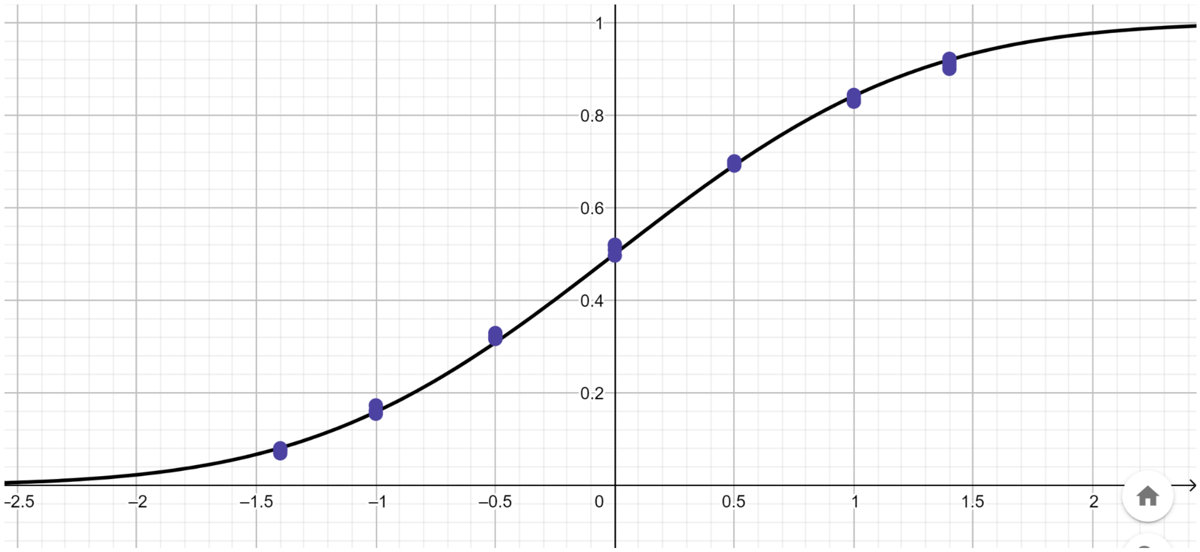

Figure 5: Standard Normal CDF with Empirical CDF Points, Intervals

3. Summarization

The probabilities calculated in the above subsections were plotted against the standard normal CDF to verify goodness of fit. The plots of probability are shown in Figures 6 through 9:

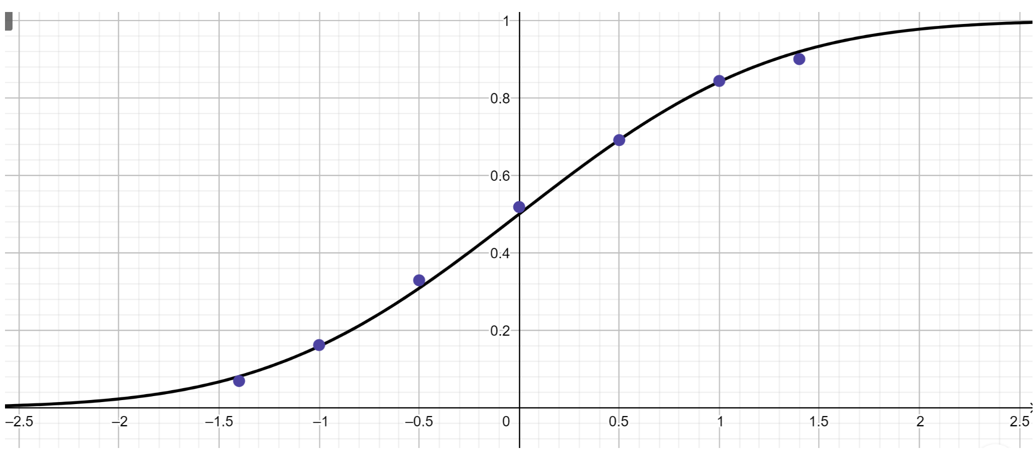

Figure 6: Standard Normal CDF with n = 3 Points

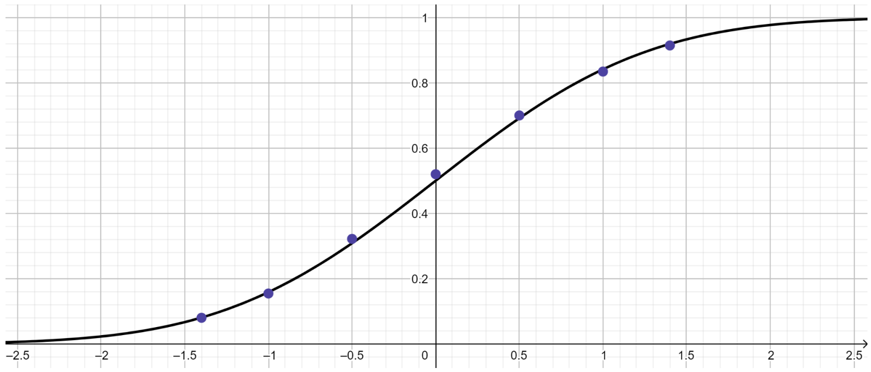

Figure 7: Standard Normal CDF with n = 9 Points

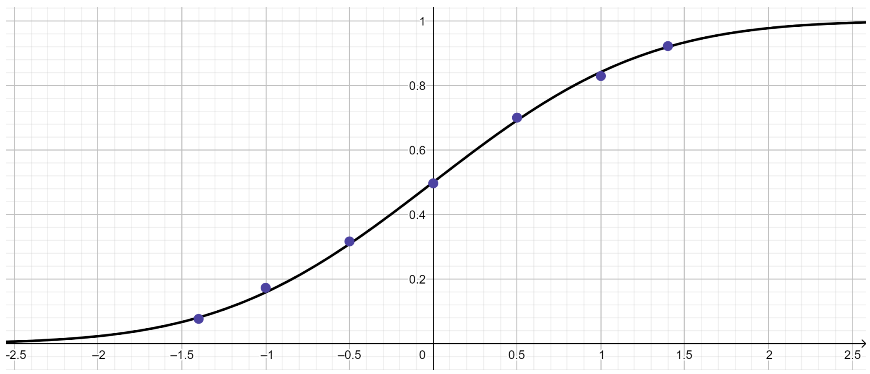

Figure 8: Standard Normal CDF with n = 27 Points

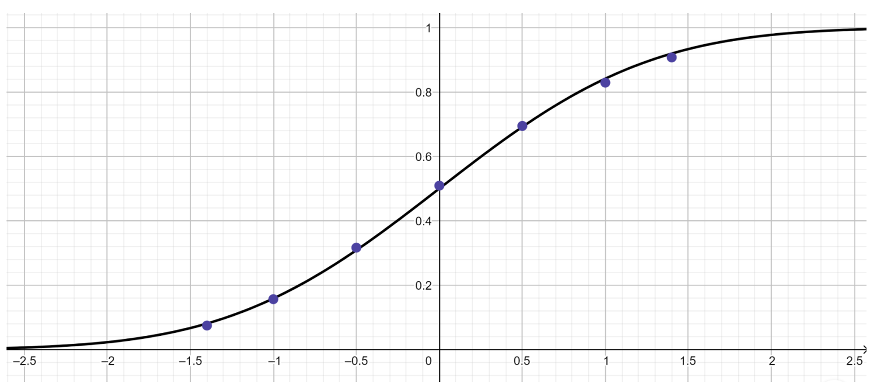

Figure 9: Standard Normal CDF with n = 81 Points

The estimates $\widehat{\mu}_n$, $\widehat{\sigma}_n$ for each sample are listed alongside their population values μ_x, $\frac{\sigma_x}{\sqrt{n}}$ in Table 10 for comparison:

n	$\widehat{\mu}_n$	$\widehat{\sigma}_n$	μ_x	$\frac{\sigma_x}{\sqrt{n}}$
3	71.6404	450.4022	71.4359	21.56
9	71.6587	172.6694	71.4359	12.45
27	70.9197	48.5790	71.4359	7.19
81	71.6811	18.8330	71.4359	4.15

Table 10: Estimates vs. Population Values

The additionally, the absolute difference $\left|\widehat{F}_n(z_j) - \Phi(z_j)\right|$ for every *j* and *n*, and MAD_n for every n are shown in Table 11:

n	J = -1.4	J = -1.0	J = -0.5	J = 0.0	J = 0.5	J = 1.0	J = 1.4	MAD_n
3	0.0117	0.0261	0.0206	0.0182	0.0006	0.0023	0.0192	0.0261
9	0.0008	0.0097	0.01332	0.02	0.0085	0.0066	0.0047	0.02
27	0.0044	0.037	0.0079	0.0036	0.0085	0.0122	0.0026	0.037
81	0.0063	0.0207	0.0079	0.0091	0.0030	0.0122	0.0119	0.0207

Table 11: Absolute Difference

4. Conclusion

All of the points on the empirical CDFs matched well with the Standard Normal CDF curve. This is also visible in the MAD_n numbers, where the highest absolute differences between the empirical and standard CDFs were only a couple hundreths. $\hat{F}_n$ seemingly does not converge towards Φ with an increased sample size, as the MAD_n values do not significantly decrease with an increase in sample size. The MAD_n values all stay within a few hundredths of 0, so though they do not necessarily converge to 0, they are certainly very close to it for all sample sizes. The small simulation did not demonstrate the meaning of the CLT very well, because though n increased from 3 to 81, the empirical CDF did not seem to converge much. This could be due to the empirical CDFs being a good approximation of the standard CDF from the start, so a visibly obvious convergence wouldn't be possible.

5. **Reflection**

   If we were to repeat this experiment, we would recommend increasing the sample size of the high end to perhaps 500 or 1000, in order to really see if there could be a strong convergence that just was not visible due to the relatively small difference between the sample sizes.

HONOR PLEDGE

On our honor as students, we have neither given nor received unauthorized aid on this assignment.

Winston Zhang (wyz5rge) and Luke Mathe (lkm6eka)