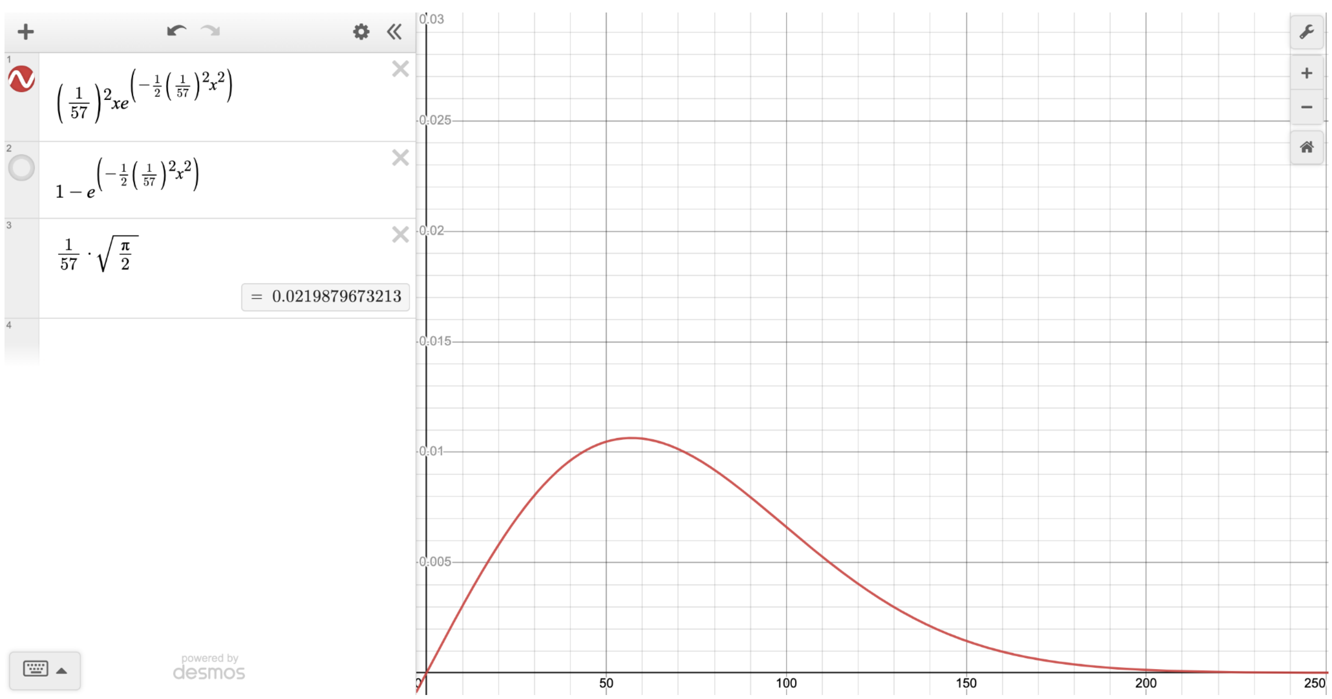
Vihar Shah
APMA 3100 Probability | Hellings

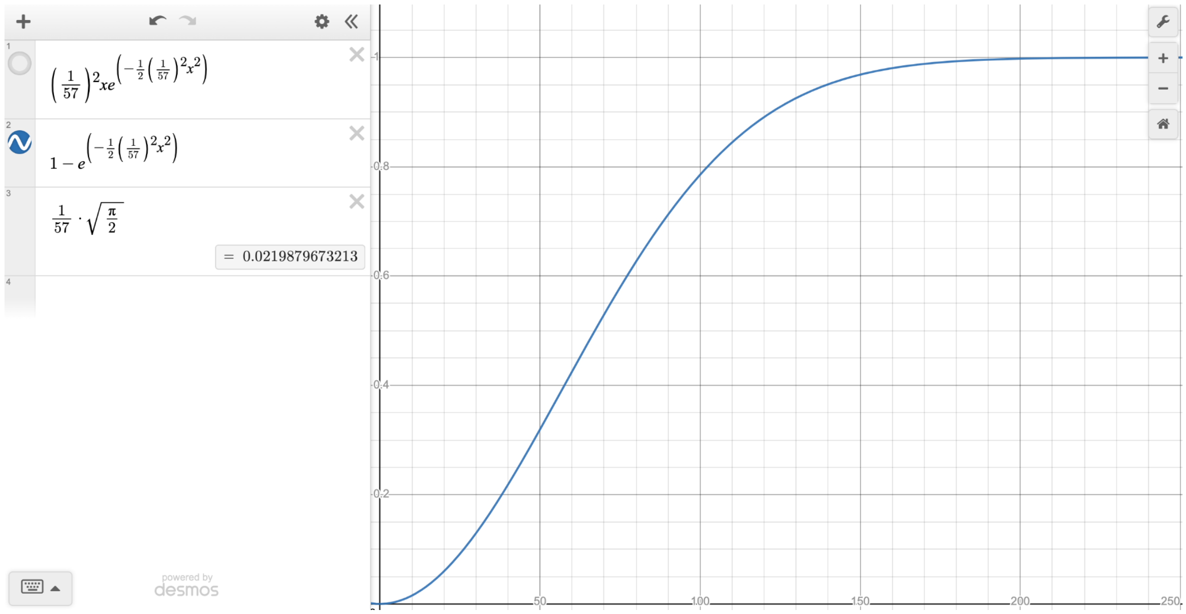
The properties of sample means, the law of large numbers, the central limit theorem, and a simulation of a drone-newspaper drop service are examined in this paper. The examination will start with analyzing the error in the placement of a newspaper to its intended dropzone as a rayleigh distribution and move into working with sample means, the law of large numbers, and normalization to glean more information on working with the same data set.

To begin, due to a variety of factors, the drone can miss the exact drop spot. This distance error can be described as the variable X which is defined by a given Rayleigh distribution with α = 1/τ. The horizontal and vertical distances of error are independent of one another and are both given to be described by Gaussian distributions (0,τ^2). Finally, the last given piece of information is that τ = 57 inches, obtained by conducting experimental flights. This experimental data can help obtain quantifiable models for the previously identified distribution functions. This leads to drop distance error with Rayleigh distribution(α = 1/57) and horizontal/vertical errors with Gaussian distribution(0,57^2).

In order to better understand the situation at hand and understand the simulation conducted, a basic model analysis was conducted. The probability distribution function (f_x) and cumulative distribution function (F_x) can be seen graphed below beside their respective Rayleigh functions.



Graph of f_X



Graph of F_X

These two visualizations describe the instantaneous and cumulative probabilities of error when the drone drops the newspaper directed at the intended drop point. Having their equations at hand can also help identify the expected value (denoted as *m* below)and variance of error for this exact situation. The variance can also be used to obtain the standard deviation (denoted as *s* below) by performing the square root of itself (Var[X] = σ_X^2). The two equations for expected value and standard deviation in a Rayleigh distribution are shown below along with the values obtained by computing them.

$$m = 57 \cdot \sqrt{\frac{\pi}{2}}$$

$$m = 71.438905827$$

$$s = \sqrt{\frac{(4-\pi)}{2\left(\frac{1}{57}\right)^2}}$$

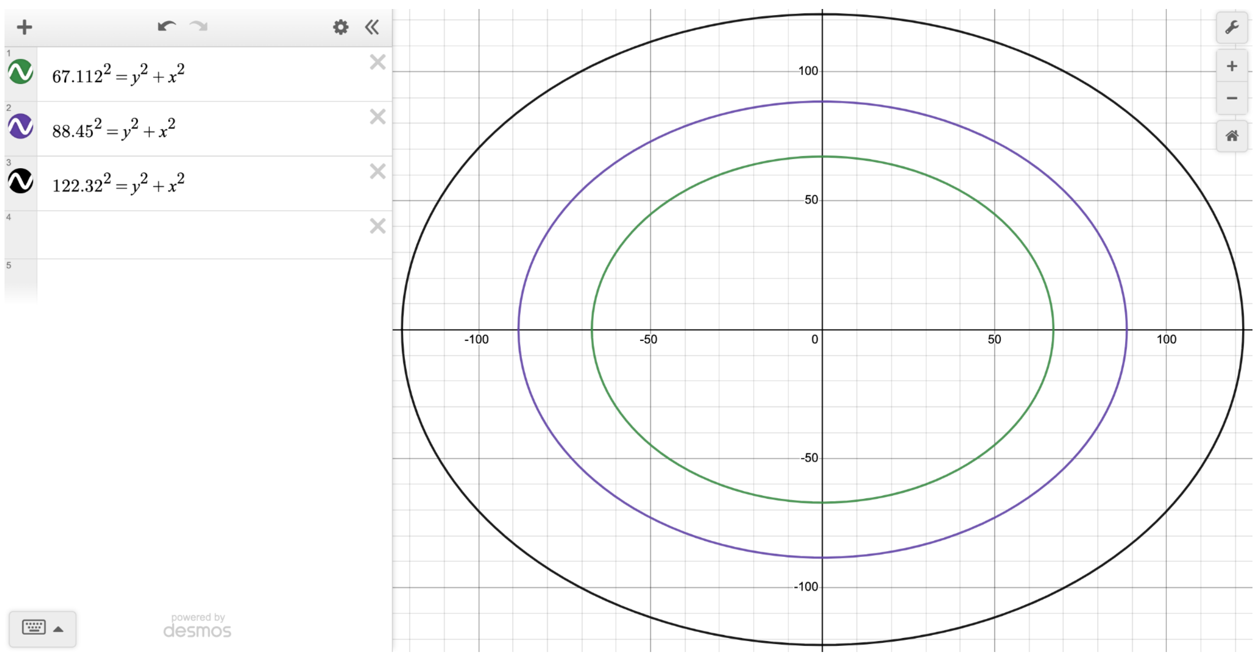
$$s = 37.342773521$$

Mean (top) and Standard Deviation (bottom)

These two values are especially useful in this simulation because they can quantify the randomness/probability of error and help test new strategies. Moving forward with this simulation, it will be increasingly important to compare how the Law of Large Numbers and

Central Limit Theorem match up to the mean and standard deviation produced with the real distributions rather than using sample means which are groups of collected samples.

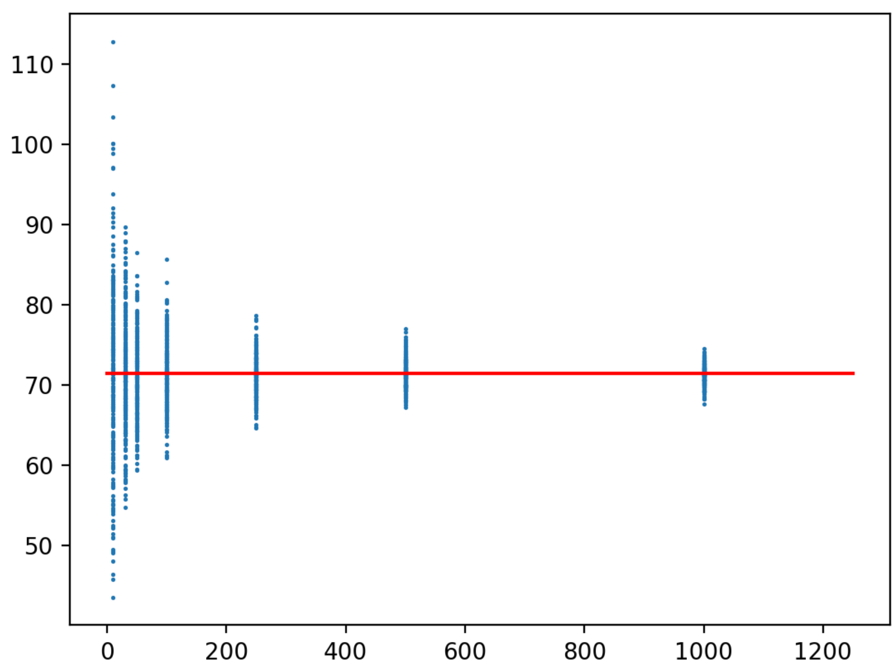
Before moving on to those concepts, it is useful to take into consideration the horizontal and vertical error distributions given. They can be visualized to help find the probability of the distance between the drop point and a point in a radius around it. The distance formula yields that the distance square is equivalent to the square root of the sum of the horizontal distance squared and the vertical distance squared or r^2 = sqrt(x^2 + y^2). This formula can be rearranged and centered around the drop point to find the probabilities that the newspaper falls within a certain radius of the drop point. This is modeled below by finding the distances where the newspaper will fall into 50%, 70%, and 90% of the time. These distances are applied to this formula and now show the drop circles where the newspaper is decently likely (50%), likely (70%), and most likely (90%) to land. The visualization helps understand the non-linear relationship between the probabilities of dropping within a certain distance of the drop point while obtaining an accurate visual aid.



Probability of Falling Within Given Distance of Drop Point
50% (green), 70% (purple), 90% (black)

By using real data from more experimental flights or real flights it is possible to test how well the given distributions fit this scenario. This can be done using a large number of trials and applying them to the law of large numbers (stating that the larger a population grows the closer the mean grows to the real average of the population) and the central limit theorem (which can be used to normalize a large set of data and make claims about).

First, this simulation works with the Law of Large Numbers by generating a function that simulates a sample of any given (n) number of values. By taking the mean of this sample, the average of that population is obtained and as discussed, the Law of Large Numbers states that the larger the population the closer the mean will be to the real average. In order to test this, the simulation creates 250 sample means of data where the number of samples taken varies (10, 30, 50, 100, 250, 500, and 1000). This yields 250 results of the sample mean at each population so that sample means with the same population can be compared to one another and those with different populations can be compared. The code created produced this data and put it into a scatter plot of (250 * 7) values. The x-axis represents the number of samples used and the y-axis represents the values of the sample mean.



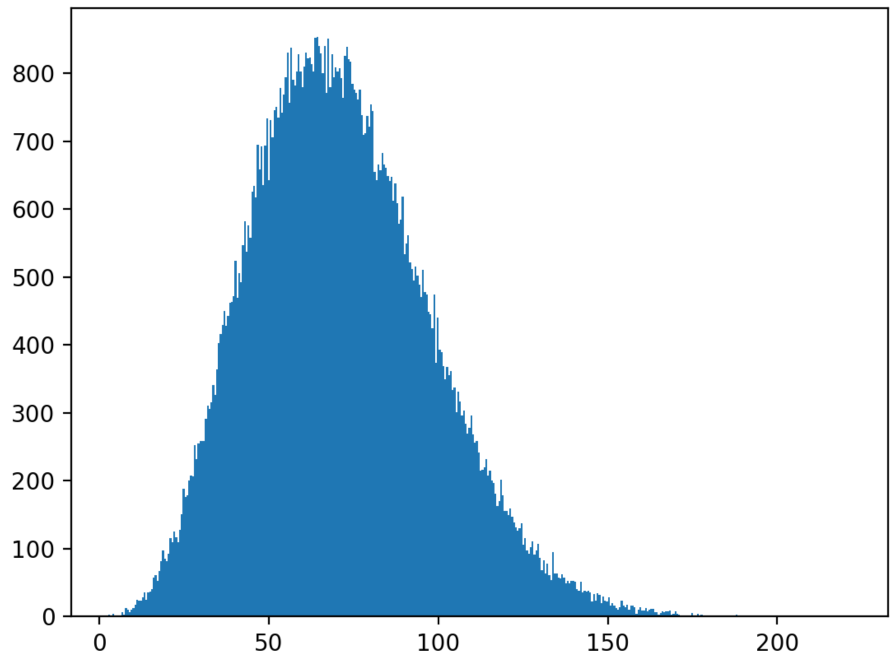
Scatter plot of values of sample means (Mn) vs # of samples (n)

This scatter plot has a clear trend of becoming more concentrated around the red line as the x-axis or number of samples increases. This aligns exactly with the Law of Large Numbers which shows the sporadic set of sample means at early sample sizes and the more concentrated trend as the number of samples increased. The red line is a solid straight line with the value of the expected value (mean) calculated using the Rayleigh distribution at the beginning of this paper. Therefore, the observation to be made is that as the number of samples used increases, the mean of the population comes closer to the real average of the overall population which aligns exactly with the Law of Large Numbers.

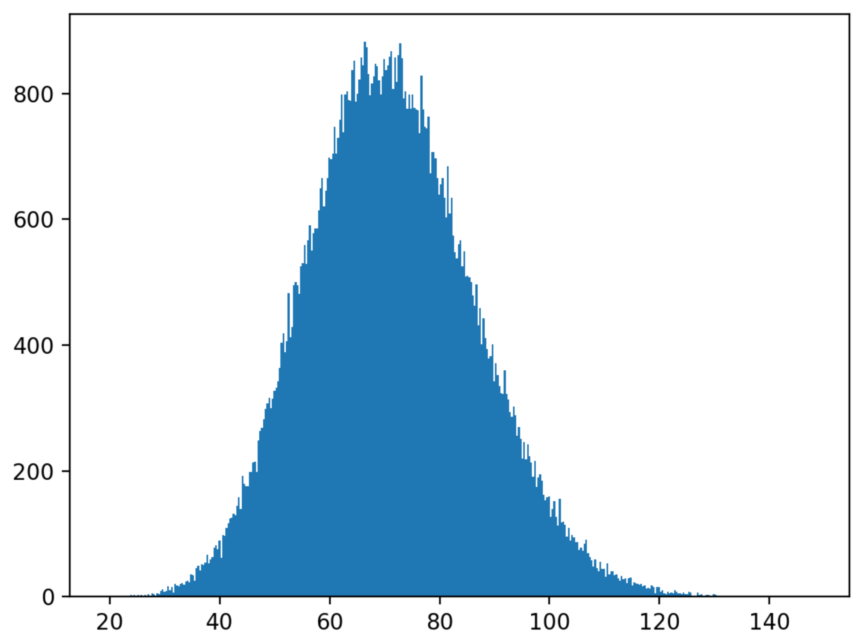
One way to make use of this property is in a real-world experiment where the information wanted is the number of times the newspaper needs to be dropped to calculate the sample mean of X such that the mean is within 6 inches of the true mean. In order to make sure that the mean is certainly within 6 inches of the true mean given a number, n*, times of dropping the newspaper, it is best to drop the newspaper 26 or more times or n*>=26. The confidence coefficient of using n*>=26 is a probability of .9999997 or almost 100% success calculated using the process shown below. The reason this value and confidence interval are extremely high is because upon running the simulation for a sample mean of lower n* (probability of success of 96%) there seemed to be a significant probability of simulation error due to the low amount of times the experiment was conducted. Since the cost of the experiment is high and failure is not acceptable it is best to find the lowest value that provides certainty that the sample mean is within 6 inches of the true mean.

After understanding the value of the Law of Large Numbers, it is important to look at the simulation under the Central Limit Theorem to verify more than the expected value. The Central Limit Theorem is defined to be able to normalize any distribution into a normal Gaussian distribution using large samples of means to create a cumulative distribution function and make claims on the overall experiment.

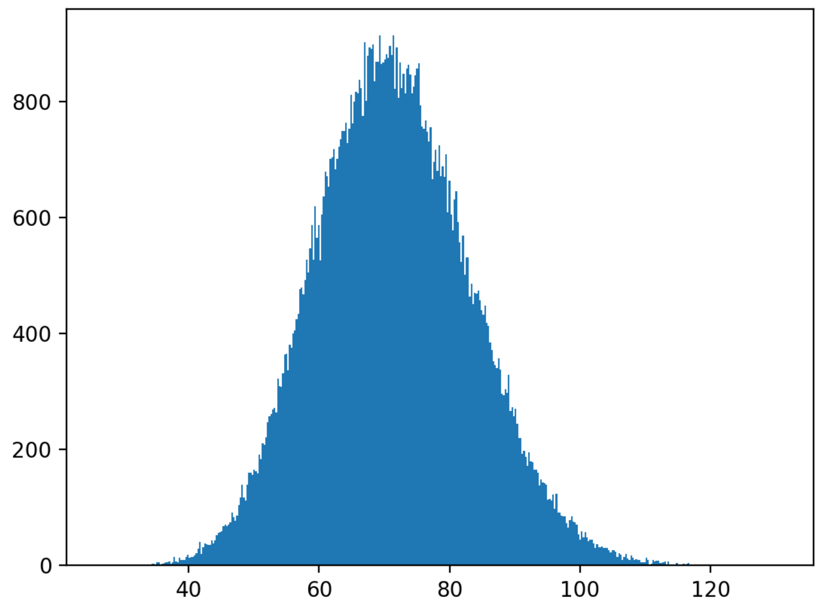
The three graphs below are filled with the respective frequencies at which the sample means lay at a specific value by using 100,000 sample means of 2, 6, and 10 values in the means respectively. Upon inspecting the graphs, it is not extremely difficult to notice the peak or area with the most frequency in each graph. This peak coincides with the actual expected value of the population very well which lies around 71. Upon inspection of the graphs, as the size of the samples increase, the normal distribution graph becomes even more refined and more clearly peaks around 71. This shows how simply gathering many samples of data or many sample means of data and plotting them in a histogram will create a normal distribution which is what allows the Central Limit Theorem to claim that any distribution can fit under a normal distribution.



Histogram of 100,000 Sample Means (n=2)



Histogram of 100,000 Sample Means (n=6)



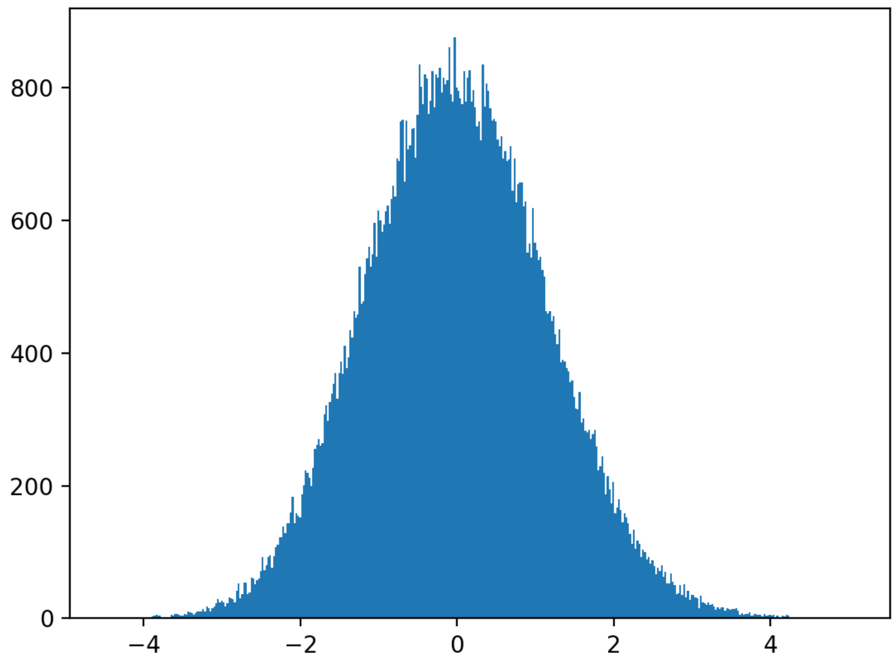
Histogram of 100,000 Sample Means (n=10)

Viewing these three graphs makes it increasingly evident how it is possible to implement the Central Limit Theorem in this situation which standardizes X. In fact, when 100,000 values of the sample mean at a given number of samples (for example, 30 as used in the graph shown below) are amassed and normalized, the distribution truly resembles a normal Gaussian PDF. Each value is standardized by subtracting the value by the total sample mean and dividing it by the standard deviation of the sample mean which was obtained by calculating the square root of the unbiased estimator for sample variance (see below).

$$\text{(b)} \quad V_n'(X) = \frac{1}{n-1}\sum_{i=1}^{n}\left[X_i - M_n(X)\right]^2 \quad \text{(unbiased estimator).}$$

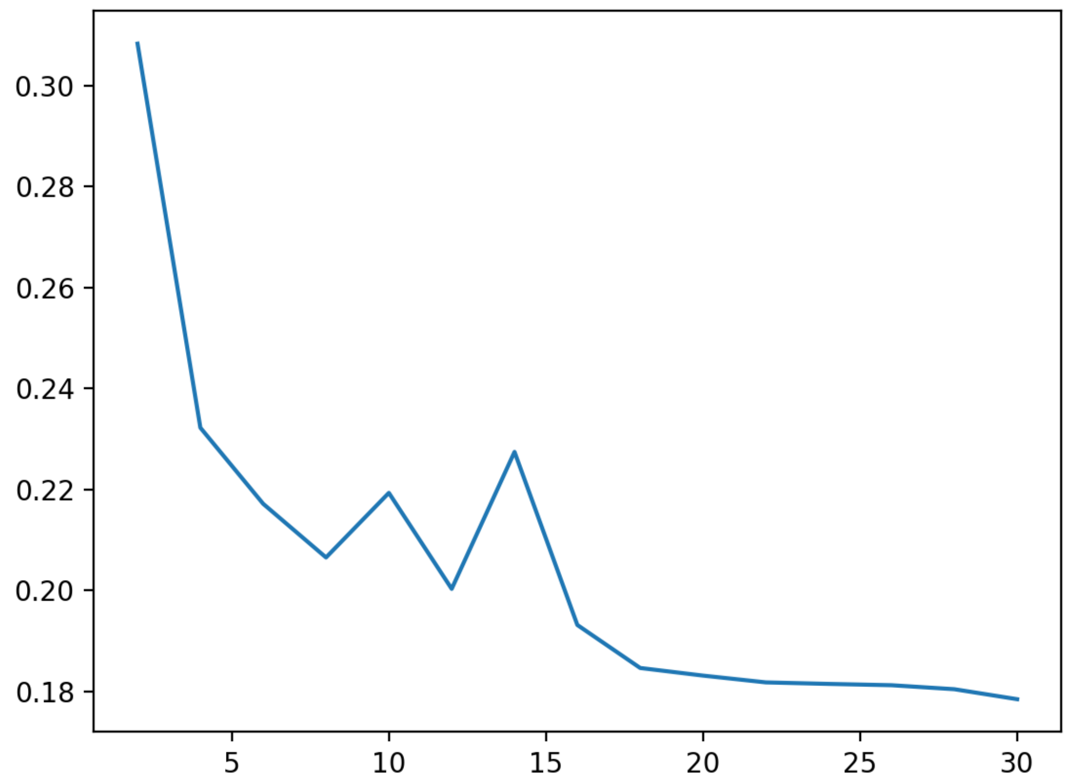
Sample Variance (Unbiased Estimator)

After standardizing each value in this form and creating a new list of normal values, the histogram plot would reveal a normal distribution resembling a Gaussian PDF and whose area under the curve can be used to calculate the cumulative distribution. It is essentially a plot of the standardized frequency as distributed among its new standardized values instead of the distance values for the histograms of the sample means above.



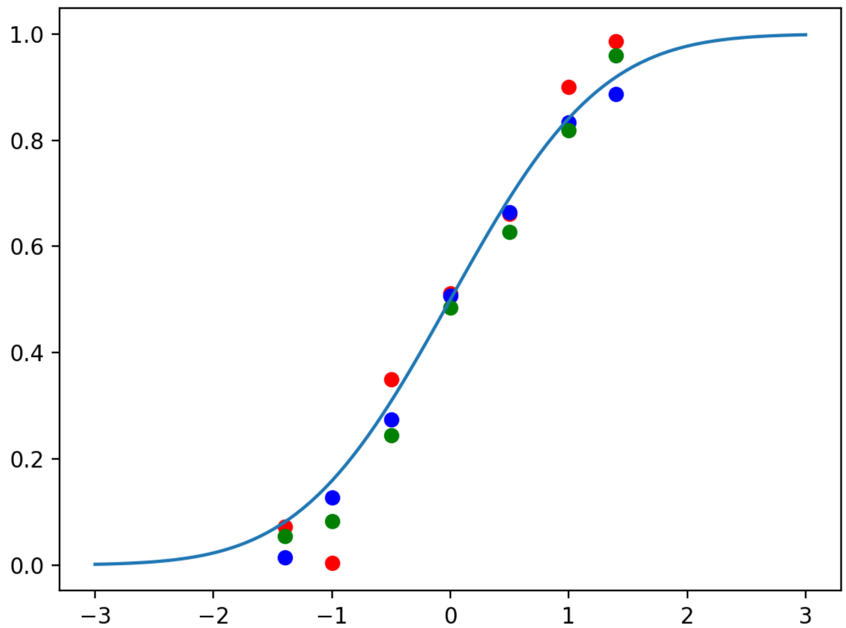
Histogram of Standardized Frequency of Sample Means

With this information it becomes possible to find the probabilities of an event occurring. In the case where the probability that the distance X is less than 71.4339, the mean, then the values can be normalized to instead find the probability that Z is less than 0. In a similar way, the simulation finds the indices at which the upper bound probability occurs and divides it by the total number of sample means taken (100,000). This creates an experimental probability of an event occurring which should be close to the actual probability of that event occurring given the large sample size. This can be tested by finding the absolute difference between the experimental probabilities and empirical probabilities. In fact, another facet of the Central Limit theorem is supported by this concept. The mean absolute deviation (MAD) was taken by finding the largest absolute difference between real and experimental probability for sample means of 2, 4, …, 28, and 30 at the probability values of -1.4, -1, -0.5, 0, 0.5, 1, and 1.4. By plotting the change in MAD as the number of samples used in the same mean increased there was a general decreasing trend in the inaccuracy of the sample mean. It would go to say, as the Central Limit Theorem states, that given a higher sample mean batch of values (i.e. 100), the MAD would be even lower and a more sophisticated and steady decline could be observed.



Mean Absolute Deviation vs Sample Mean Size

Lastly, to wrap up the simulation and further support the use of the Central Limit Theorem, some experimental values of different sample mean baskets (2, 6, 10) were superimposed on the empirical, actual CDF of the normal Gaussian distribution. The red dots correspond to sample mean of 2 used, blue to a sample mean of 6 used, and green to a sample mean of 10 used. The trend in how close the dots are in comparison to the empirical graph follows the same pattern with the red dots being somewhat inaccurate, the blue dots being slightly more accurate, and the green dots being the most accurate among the compared sample means. The Central Limit Theorem would state that larger sample mean baskets used would have dots in a tighter position with the empirical CDF graph as the current trend suggests.



7 Probabilities of Sample Mean Baskets Superimposed on Empirical CDF of Normal Gaussian. Probability vs Normal Distribution

Rediscovering the value of the Law of Large Numbers and Central Limit Theorem in a practical sense rather than a theoretical one helped provide valuable insight into the backbone and properties of both concepts. The practicality of these methods to both test, verify, and confirm the provided probability distributions has many applications including and beyond the drone-newspaper delivery seen in this simulation. In conclusion, this simulation covered some of the ways expected values and probabilities can be compared and checked with alternative methods that can be more practical in some contexts of the world.