Statistic Analysis Project

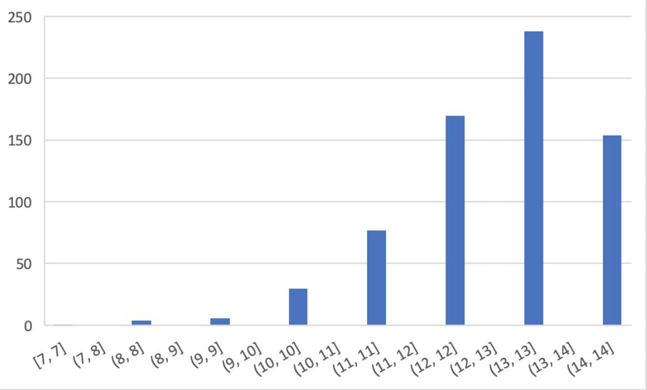
Jay Shah
Alexander Kravtsov
Walker Battey

11/26/2019

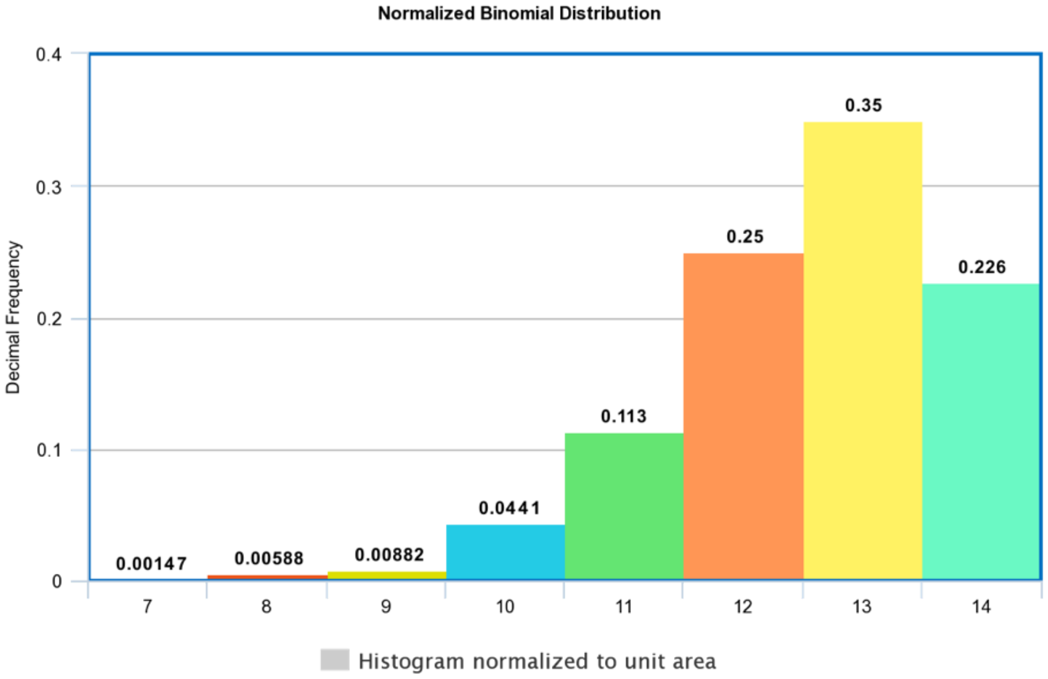
Column A Tasks - Jay Shah

A.1 For the type of distribution presented in the first column, plot a histogram (normalized to unit area) of the data.

The first thing I did with the data, knowing it is a binomial distribution, was to find the minimum and maximum values, which were 7 and 14 respectively. I then counted the total number of data points which was 680. Since part A is asking for a histogram normalized to unit area, I used Excel's built in histogram creator to graph a chart of my data set based on the frequency of each value from 7-14 appearing in the data set.



To create a Normalized Binomial Distribution, I had to then divide the frequency of each value in the range by 680, since 680 was the number of data points in my set. This gave me a decimal value for each data point that I could use to create a normalized distribution. I then made my y column of the histogram by finding the frequencies by taking each value for the number of times each # appeared across all of my data points and dividing them by the total number of data points. Then I plotted N vs the frequencies to make up the histogram.



A.2 Estimate the mean and the variance for the column A distribution. For comparison, the exact variance is given in the cell labelled (here) Distribution Variance.

Using my dataset, I selected every item in column A, and calculated the mean, variance and standard deviation using the given equations.

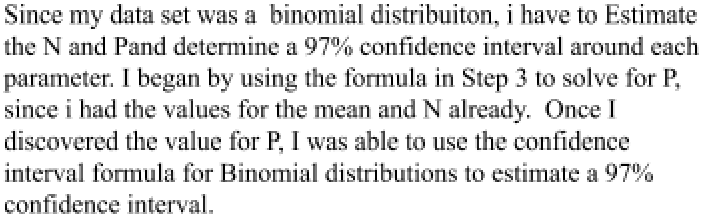
$$\overline{X} = (1/680) \sum_{i=1}^{680} X_i = 12.544118 \simeq \mathbf{12.544}$$

$$S^2 = (1/(680-1)) \sum_{i=1}^{680} (X_i - 12.544118)\text{^}2 = 1.4519683 \simeq \mathbf{1.462}$$

$$S = \sqrt{S^2} = 1.2049764728 \simeq \mathbf{1.208}$$

A.3 The distribution found in the first column depends on two parameters as follows: a and b. Estimate the two parameters associated with the distribution given in Column A and determine a 96% confidence interval around each parameter.

N = 680, P = ?
$\overline{X} = NP$
$P = (\overline{X}/N) = (12.54/680) = \mathbf{0.018441}$



Since my data set was a binomial distribuiton, i have to Estimate the N and Pand determine a 97% confidence interval around each parameter. I began by using the formula in Step 3 to solve for P, since i had the values for the mean and N already. Once I discovered the value for P, I was able to use the confidence interval formula for Binomial distributions to estimate a 97% confidence interval.

Confidence Interval

$$\overline{X}/N - Z_{\alpha/2}\sqrt{(}\,(\overline{X}/N(1-\overline{X}/N)/N) < P < \overline{X}/N + Z_{\alpha/2}\sqrt{(}\,(\overline{X}/N(1-\overline{X}/N)/N)$$

$$\alpha = 1 - 0.97 = 0.03$$
$$\alpha/2 = .015$$
$$Z_{\alpha/2} = \mathbf{2.17}$$
$$\overline{X}/N = 12.54 / 680 = 0.018441$$

Error Estimate $= Z_{\alpha/2}\sqrt{(}\,(\overline{X}/N(1-\overline{X}/N)/N)$

$= 2.17\sqrt{(}\,(0.018441(1-0.018441))/680)$

$\simeq 0.011198$

Therefore, the confidence interval = $\overline{X}/N - 0.011196 < P < \overline{X}/N + 0.011196$ =

$0.007245 < P < 0.029637$

A.4 How large a data set is needed to get 97% confidence intervals of width 0.01 or smaller around the two parameters? (Assume $\bar{X}$ and S2 do not change significantly with N when N is large.)

Since the width given in the problem statement states 0.01 or smaller, we would need an error estimate of less than or equal to 0.005. We can use the Sample size determination formula to generate a value for N. We use this formula to estimate N since we have information about probability already.

$N = 680$
$P = 0.018441$
Error Estimate = $E = 0.005$
$Z_{\alpha/2} = 2.17$

$$\mathbf{N \geq P(1-P)(Z_{\alpha/2}/E)^2}$$

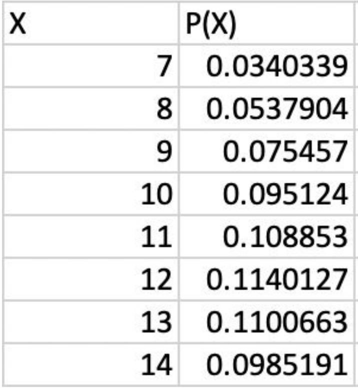
$$N \geq (0.018441(1-0.018441)(2.17/0.005)^2$$
$$N \geq 3409.45$$

If width is 0.01, for an error estimate of 0.005, then $N \geq 3409.45$ ≈ **3409** since N must be an integer. This estimate looks very similar to a Poisson Distribution due to N being a large sample size, and probability being so low.

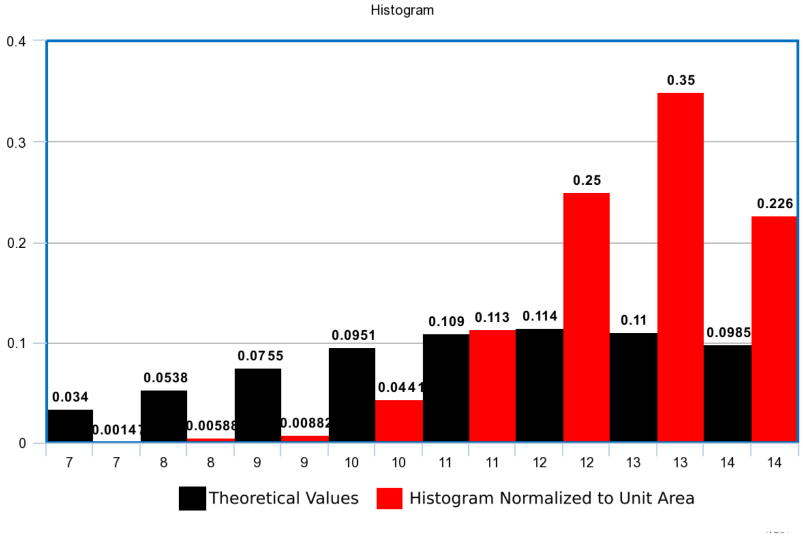
A.5 Plot a graph of the density function for the distribution in column A using the estimated parameter values determined in part A.3. Compare this graph to your normalized histogram.

For this part of the project, I wrote a binomial probability mass function in excel that calculates the binomial probability. The function takes three parameters: X, T, P and C. X is the number of successes in the trial, T represents the number of trials, P represents the probability of success of each trial, and lastly, C represents a boolean value for a cumulative distribution.

Using the function, I calculated the density for the distribution ranging from 7 to 14, and received these values for the probability.



X	P(X)
7	0.0340339
8	0.0537904
9	0.075457
10	0.095124
11	0.108853
12	0.1140127
13	0.1100663
14	0.0985191



Here we can see that the histogram normalized to the unit area is left tail skewed, whereas the density function (aka theoretical values) is right and left tail skewed with a concentration in the middle of the graph.

Column B Tasks - Alexander Kravtsov

B.1 From the B column select two non-overlapping chunks of consecutive data points. The first chunk should contain a large number of data points. The second chunk should contain exactly 25 data points. Estimate the mean and variance of this normal distribution using your first (large) chunk of data.

The **large** chunk of data contains data points B4 through B2000.

N_L=1997

μ_L=3.849231

Variance$_L$=0.198023

The **small** chunk of data contains points B2001 through B2025

- In order to calculate the mean, the function AVERAGE was used.

  i.e. μ_L=AVERAGE(B4:B2000)

- In order to calculate the variance, the function VAR was used.

  i.e. Variance_L=VAR(B4:B2000)

B.2 Compute 98% confidence intervals around each of the parameters μ and σ based on the large chunk of data.

98% confidence intervals around μ

Distribution $\frac{\overline{X}-\mu}{\frac{S}{\sqrt{N}}}$ is the **t distribution** with N-1 degrees of freedom**.** First, standard deviation has to be found, using STDEV function.

1) Standard deviation=S=STDEV(B4:B2000)=0.444997
2) $P(-t \le T \le t)=0.98$
3) $\alpha = 1-0.98 = 0.02 \Rightarrow \alpha/2 = 0.01$

In order to find t, the function T.INV(1- α/2,N-1) was used:

4) t=T.INV(0.99,1996)=2.328218

Confidence interval around μ follows the equation $X + t_{N-1,\alpha/2}(S/\sqrt{N})$ for the upper estimate and $X - t_{N-1,\alpha/2}(S/\sqrt{N})$ for the lower estimate:

5) Upper estimate= 3.872421
6) Lower estimate=3.826041

98% confidence intervals around σ

Distribution $\frac{S^2(N-1)}{\sigma^2}$ is the χ^2 **distribution** with N-1 degrees of freedom**.**

In order to find χ^2 , the function CHISQ.INV.RT was used:

1) $\chi^2_{\alpha/2}$ = CHISQ.INV.RT(0.01,1996) = 2145.918
2) $\chi^2_{1-\alpha/2}$ = CHISQ.INV.RT(0.99,1996) = 1851.964

Upper Estimate equation is $\sqrt{\frac{(N-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}}}$ => Upper Estimate=0.461862

Lower Estimate equation is $\sqrt{\frac{(N-1)S^2}{\chi^2_{\frac{\alpha}{2}}}}$ => Lower Estimate=0.429064

B.3 Test the claim μ ≤ 4 using the second (small) chunk of data in a significance test. State the null and research hypotheses. Describe the location of the critical region. Give the P-value and the Z, T, or χ2-stat as appropriate. Find the region where the power of the test exceeds 0.95.

The **small** chunk of data contains points B2001 through B2025:

N_s=25

μ_s=3.830493

Variance_s=0.238491

Standard deviation=S=STDEV(B2001:B2025)=0.488356

Distribution $\frac{\overline{X}-\mu}{\frac{S}{\sqrt{N}}}$ is the **t distribution** with N-1 degrees of freedom.

Null hypothesis: H0=$\mu \ge \mu_0$, where $\mu_0 = 4$

Research hypothesis: H1=$\mu < \mu_0$, where $\mu_0 = 4$

The critical region is located to the right of μ_s

α=P($\overline{X} \le$ C|H0 is false)

=> α=P($\frac{\overline{X}-\mu}{S/\sqrt{N}} \le \frac{C-\mu}{S/\sqrt{N}}$ | μ = 4) , where μ=4 because it is the limit for the problem

$$T_{N-1} = \frac{\overline{X}-\mu}{S/\sqrt{N}} = -1.735488$$

=> α=P($T_{N-1} \le \frac{3.830493-4}{0.238491/\sqrt{25}}$)

In order to find the P-value, the function T.DIST is used.

=> T.DIST($\frac{\overline{X}-\mu}{S/\sqrt{N}}$, 24, TRUE)=0.04774

Because the P-value is less than 0.05, the null hypothesis is rejected as there is no probability of a Type I error.

Power of the test:

$\beta = P(\overline{X} \ge C | H_0$ is false)

$$\beta = P(\overline{X} \geq C \mid \mu = \mu^*)$$

$$\Rightarrow \beta = 1 - \mathrm{P}\left(T_{N-1} \leq \frac{C-\mu^*}{S/\sqrt{N}}\right)$$

$$1 - \beta > 0.95 \Rightarrow \mathrm{P}\left(T_{N-1} \leq \frac{C-\mu^*}{S/\sqrt{N}}\right) > 0.95$$

$$\mathrm{C} = \overline{X}$$

$$\overline{X} - \mathrm{S} * \sqrt{N} \; * T_{N-1}{}^{-1} * (0.95) > \mu^*$$

$T_{N-1}{}^{-1}$ can be found using the function T.INV(0.95,24)

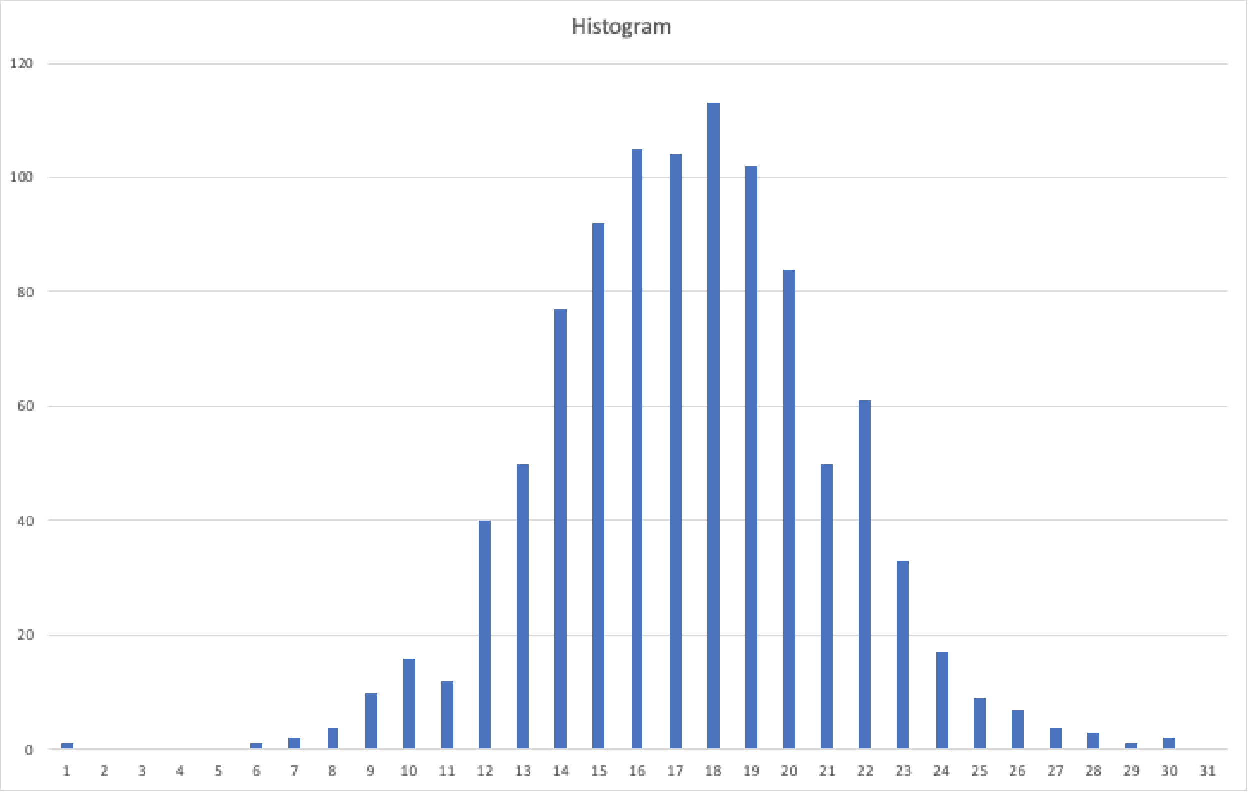
$$T_{N-1}{}^{-1} = 1.710882$$

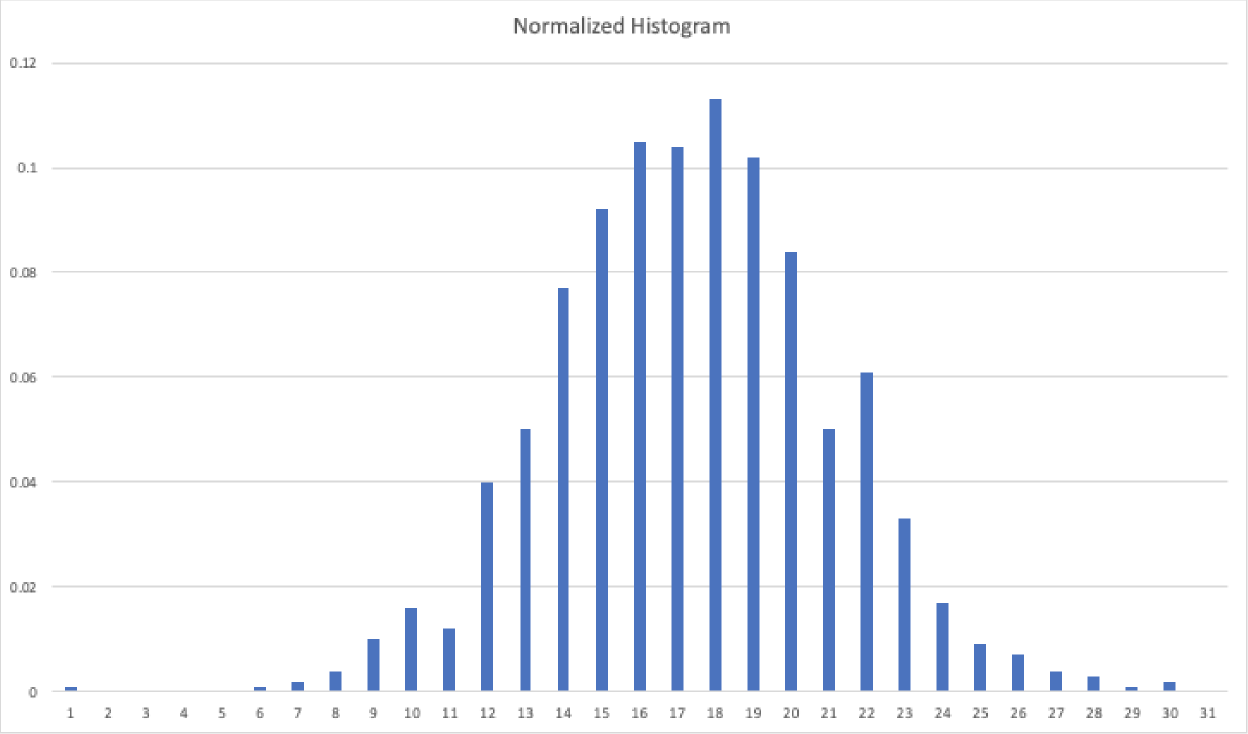
$$\Rightarrow \mu^* < 3.841251$$

μ* represents the region where the power of the test exceeds 0.95

C.1 Histogram

A histogram of the data was created to take initial guess at the distribution of the data as it was given as the mystery distribution data set. This histogram has 30 bins, from 0 to 29, of width *w = 0.3987124*, *a = 1.46179331*, *b = 5.53729092*. From the shape of the histogram, it was predicted that the distribution of the data set would be normal due to the shape of the distribution being unimodal and symmetric. The left side of the distribution shows to drop off a little more steeply than the right side, seemingly signaling a gamma distribution but this was dismissed as it appeared as though the steepness of the drop seemed insignificant as it looks somewhat, although not entirely similar to that of the right side of the data set.





C.2 Goodness of Fit Assuming Normal Distribution

Operating under the assumption that the distribution is normal, $\overline{X}$ was used to approximate the mean and s was used to approximate the standard deviation. Using these estimators μ = 5.53729092 and σ = 2.13683967. The hypotheses used to perform a goodness of fit test were:

$$H_o : \textit{ distribution is normal}$$

$$H_1 : \textit{ distribution is not normal}$$

The critical region for this test was given by [C, ∞), where

$$C = F^{-1}_{x^2_v}(1 - \alpha)$$

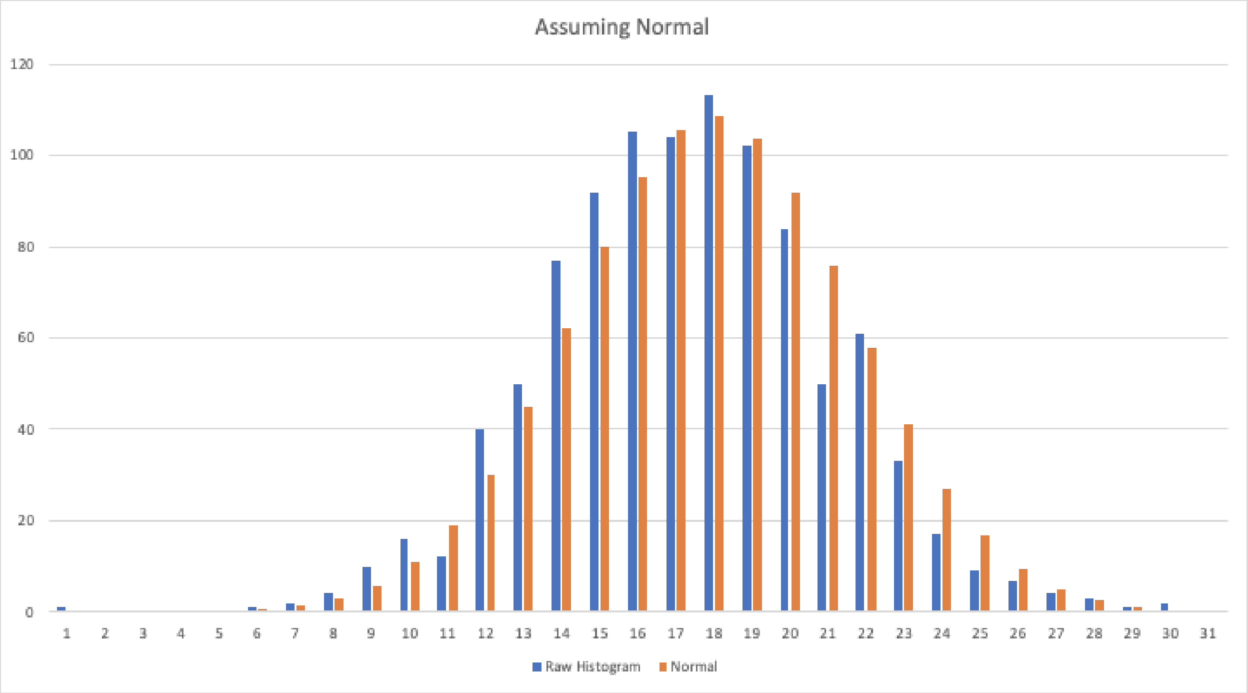
For this distribution, $\alpha = 0.05$ and $\nu = 30 - 3 = 27$, so $C = 40.1$, hence the critical region is [40.1, ∞). The statistic further used to reject or accept the Goodness of Fit hypothesis is G_f, given by:

$$G_f = \sum_{j=1}^{N} \frac{(o_j - e_j)^2}{e_j}$$

O_j is the actual number of samples in a bin while e_j is the expected number for a normal distribution with the same bin width, mean and standard deviation. G_f is also the χ^2 - statistic and is equal to 294.9529547. The P-value was found by using the equation:

$$P = 1 - F_{x^2_v}(G_f)$$

The resulting P-value was 3.36241E-47, so at a rejection level of 0.05, the null hypothesis is rejected because 3.36241E-47 < 0.05.

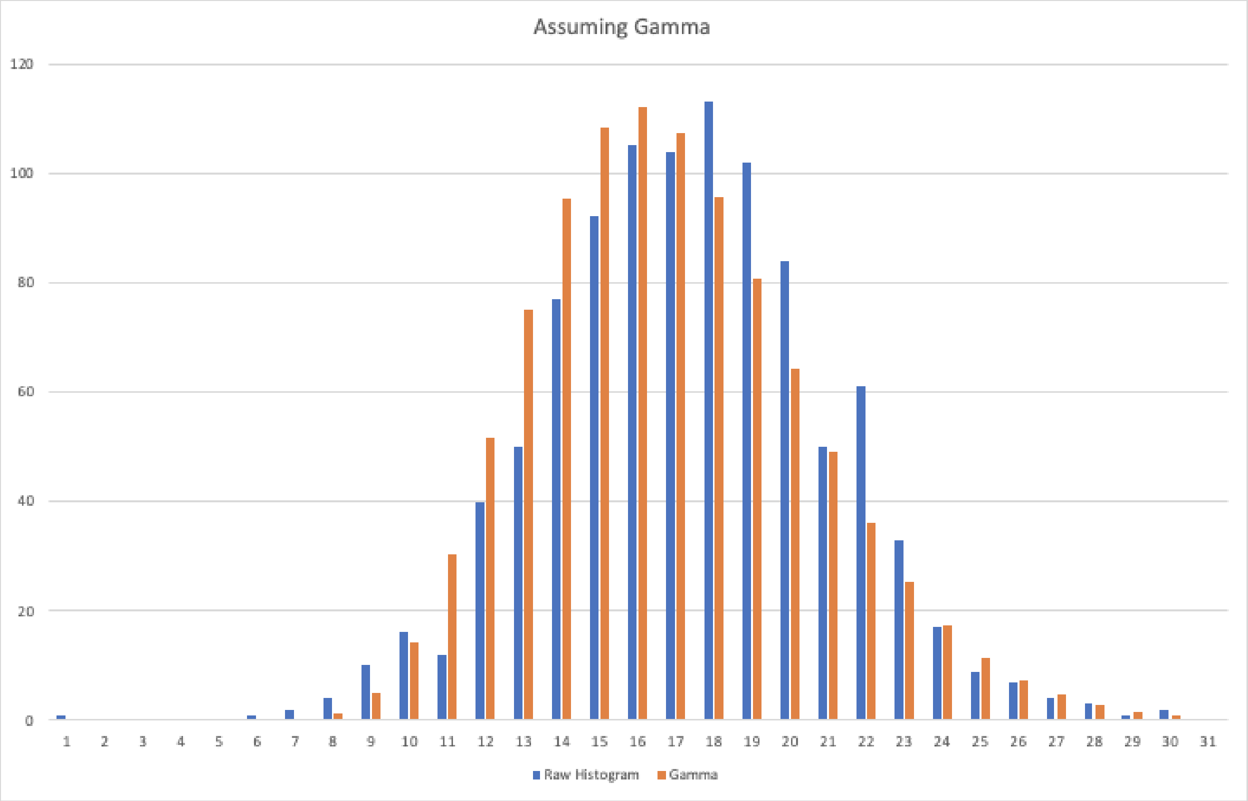


C.3 Goodness of Fit Assuming Gamma Distribution

Operating under the assumption that the distribution is gamma, $\overline{X}$ was used to approximate the mean and s was used to approximate the standard deviation. Using these estimators $\mu = 5.53729092$ and $\sigma = 2.13683967$. Next, we found he values for both the alpha and beta estimators $\alpha = 14.34903667$ and $\beta = 0.385899838$. The hypotheses used to test the Goodness of Fit were:

$$H_o : \text{ distribution is gamma}$$
$$H_1 : \text{ distribution is not gamma} .$$

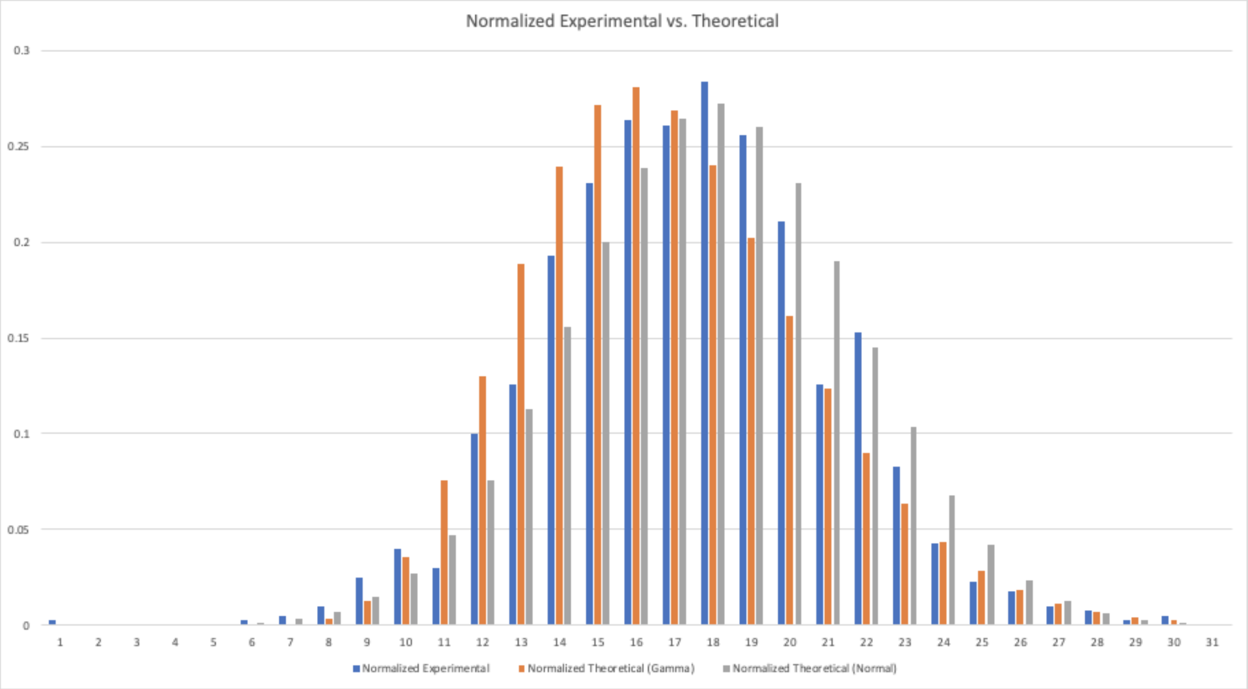
Using the same processes as described in C.2, C = 40.1, and $G_f = 181.2576868$ (e_j values were found assuming a gamma distribution with the α and β values noted earlier). The resulting P-value calculated was 3.73302E-25, so at a rejection level of 0.05, the null hypothesis is rejected because 3.73302E-25 < 0.05.



C.4 Mystery Distribution

The Goodness of Fit tests rejected the hypothesis that the data was either a Gamma Distribution or a Normal Distribution. The Gamma Distribution, however, yielded a P-value that was about 10^{20} higher in magnitude so it can be said that the data given is closer to a Gamma Distribution than a Normal Distribution. This opposed the initial guess that the Mystery Distribution was a Normal Distribution, but did not come as a surprise due to our initial observations of the steepness of the left tail of the distribution. The histogram below plots our Normalized

Experimental data against the Normalized Theoretical data of both Gamma and Normal Distributions.



Although the data appears to be closer to that of the Normal Distribution on the right side of the histogram, we can not say that it is more normal due to how the data acts on the left side of the histogram The data follows more closer to that of the path of a Gamma Distribution with a right-skew, therefore helping us show further that the data is closer to that of a Gamma Distribution.