AMTH 108 Final Stats Project
Team: Gamma 3

Tamir Enkhjargal, Justin Ling, Soren Madsen

March 2019

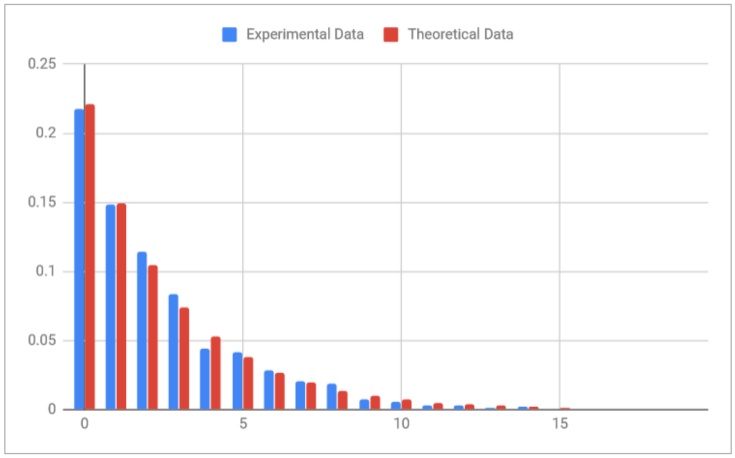
1 Column A

We all worked on each column in sections and were constantly helping each other. Therefore, we would say the analysis for all three columns were done by all of us equally.

Beginning from 0 to the maximum, we determined the lower bounds of a bin, by taking the range of the data set and dividing by the bin amount, and used `COUNTIFS` to count how many values from the data set that fit between the bounds. For example, 263 data points fit between 0 and 1.3421, 179 data points fit between 1.3421 and 2.6879, etc.

Lower Bounds	Bin Number:	Histogram:	Experimental Data	Theoretical Data
0	0	263	0.2179650842	0.221029844
1.342116093	1	179	0.1483488596	0.1496497621
2.682799732	2	138	0.1143695119	0.1048985967
4.02348337	3	101	0.08370522245	0.07429123017
5.364167009	4	53	0.04392452267	0.05287999103
6.704850648	5	50	0.04143822894	0.03775247132
8.045534287	6	34	0.02817799568	0.0270061408
9.386217926	7	25	0.02071911447	0.01934615763
10.72690156	8	23	0.01906158531	0.01387354757
12.0675852	9	9	0.007458881209	0.009957223115
13.40826884	10	7	0.005801352051	0.007151136705
14.74895248	11	4	0.003315058315	0.005138613042
16.08963612	12	4	0.003315058315	0.003694125728
17.43031976	13	2	0.001657529157	0.002656698796
18.7710034	14	3	0.002486293736	0.001911235785
20.11168704	15	1	0.0008287645787	0.001375335518
21.45237067	16	1	0.0008287645787	0.0009899429572
22.79305431	17	0	0	0.0007126990658
24.13373795	18	1	0.0008287645787	0.0005131994801
25.47442159	19	2	0.001657529157	0.0003696080119
26.81510523	20	NOT INCLUDED		

From the data, the data seems to be following an exponential distribution, starting from high to low. After normalization and estimating our α and β using $\overline{x}$ and the given variance, we were able to plot the theoretical gamma data using `=GAMMADIST(LOWER_BOUND, ALPHA, BETA, FALSE)`



Using the data from the `EXPERIMENTAL DATA` and `THEORETICAL DATA` and plotting them together, we see that the experimental distribution follows the same height and trend as the theoretical distribution. This means that the estimated α and β match closely to the experimental data.

Experimental	
Alpha	0.9408563033
Beta	4.123807099
Estimated Var:	14.84572172
Estimated Mean	3.879909903
Known Var:	16

These values of α and β were estimated using the relationships $\alpha = \mu^2/\sigma^2$ and $\beta = \sigma^2/\mu$ and using the given variance for σ^2 and $\overline{x}$ for μ.

Confidence Interval for Alpha and Beta from Mu			
mu_hat	3.879909903	Alpha(mu)=	mu^2/sig^2
sigma:	4	Upper Alpha	1.078348854
sqrt(N):	30	Lower Alpha	0.8127368296
z_mu/2:	2.053748909	alpha+-	0.1328060122
mu=	mu_hat	Beta(mu)=	sig^2/mu
CI Upper Bound	4.153743091	Upper Beta	4.436954969
CI Lower Bound	3.606076715	Lower Beta	3.851947425
x +-	0.2738331879	beta +-	0.2925037724

To setup the confidence interval for α and β, we need to first find the confidence interval on μ. Using the equation $\mu = \overline{x} \pm z_{\alpha/2} * \sigma/\sqrt{N}$ we found the bounds of μ and we can use that found bounds back in our relationship $\alpha = \mu^2/\sigma^2$ and $\beta = \sigma^2/\mu$. Therefore, from our confidence intervals, we can state with 96% confidence level that the true α lies between 0.8127 and 1.0783 and the true β lies between 3.8519 and 4.4370.

N for 0.01 width around alpha:
634947.7272
N for 0.01 width around beta:
3049504.964

Testing to find the sample size necessary to get a 0.01 width around the parameters alpha and beta consisted of just reversing the equation, solving for μ: $0.01 \leq \overline{x} \pm z_{\alpha/2} * \sigma/\sqrt{N}$ and using the same relationships again, we found the sample size necessary to have a confidence interval at 96% and width 0.01 to be around 635,000 for α and 3,049,500 for β.

2 Column B

We can use the average of the data ($\overline{x}$) and the sample variance ($\S^2$) as unbiased estimators the true μ and σ^2 because we have a large amount of samples (N = 9975).

Avg first 9,975	3.54523089	N-1	9974
Avg last 25	4.039442587	Chi-Sq Alpha/2	9,648.38
Estimated Mean (large)	3.54523089	Chi-Sq 1-Alpha/2	10,305.51
Estimated Var (large)	3.48107735	L1	3.598560271
z-value	2.326347874	L2	3.369098696
Sq Root of N	99.87492178	Sqrt L1	1.896987156
Upper Bound	3.626314277	Sqrt L2	1.835510473
Lower Bound	3.464147503	S	1.865764548
x+-	0.0810833866	S+-	0.03073834166

The confidence interval on μ at 98% confidence level would mean solving for $\overline{\mu} = \overline{x} \pm z_{0.01} * S/\sqrt{N}$. Finding the z-value, square root of N, and S gave us the upper and lower bounds of $\overline{x}$. The confidence interval states that with a 98% confidence, our true μ lies between 3.4641 and 3.6263. Testing a confidence interval on σ uses the chi-square distribution, $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$. Using this, we determined with 98% confidence that our true σ lies between 1.8355 and 1.8970.

Null Hypothesis	mu >= 4			
Alt. Hypothesis	mu <4			
Critical Region C	mu = 4			
P-Value	0.4609928812	T-stat	0.09896788951	
.95 Pwr Crit Reg	4.169811988	<--- values below this fall under critical region		

Using the last chunk of 25 data points, we have some data that shows variability away from our larger 9975 sample sized data. For example, the average value of the last 25 was 4.0394, while the average value of the first 9975 was 3.5452. The null hypothesis we want to set up is:

$$H_0 : \mu \geq 4 \tag{1}$$

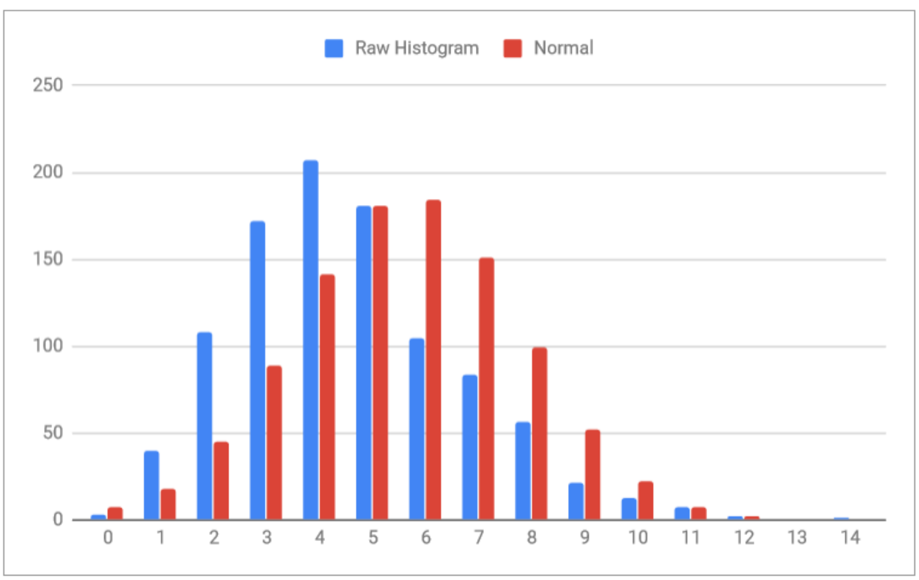
$$H_a : \mu < 4 \tag{2}$$

The critical region, the region where we reject the null hypothesis is when $\mu < 4$. To take the worst case scenario for α we find it on the boundary of the critical region, at $\mu = 4$. We used a t-statistic because we have a normal distribution with no known sigma, and not a high amount of n. The t-statistic used was calculated from $(\overline{x} - c)/(s/\sqrt{N})$.

We can not state with certainty that there's a chance of making a Type I error, due to a middle-ground result of P-value. The location where the power is greater than 0.95 is at 4.1698. This was found by taking the `=TINV(0.95, 24)` and adjusting it at the c level of 4. This critical region just states that we can make a 95% confidence of rejecting the null hypothesis if our observed $\hat{\mu}$ estimate was above 4.1698.

3 Column C

Minimum:	1.084990799	mu_hat	6.340728176
Maximum:	16.54494554	Sample Variance:	4.698103312
Bins:	15	Estimated b	2.167510856
Width:	1.030663649	Estimated a	6.340728176
Sample Count:	1000		
Bin Number:	Lower Bounds:	Raw Histogram	
0	1.084990799	3	
1	2.115654448	40	
2	3.146318098	108	
3	4.176981747	172	
4	5.207645397	207	
5	6.238309046	181	
6	7.268972695	105	
7	8.299636345	84	
8	9.330299994	56	
9	10.36096364	21	
10	11.39162729	13	
11	12.42229094	7	
12	13.45295459	2	
13	14.48361824	0	
14	15.51428189	1	
15	16.54494554		



Following the standard procedure, finding the range of the data, setting an arbitrary bin count, and finding the width, we can create our histogram by counting the number of values in our data that fall between these bins.

We plotted the distribution of the ***normal*** data from ***part C.2*** next to it, with the found $\overline{x}$ and sample standard deviation s. We determined that out of the two choices, it was most likely a ***Gamma*** distribution and not normal.

Assuming Normal				
Z value	Area Below	Area Within	Normal	(O-E)^2/E
-2.424780186	0.007658829128	0.007658829128	7.658829128	2.833943477
-1.94927454	0.02563132369	0.01797249456	17.97249456	26.99742066
-1.473768894	0.07027191305	0.04464058936	44.64058936	89.92746229
-0.9982632484	0.1590758619	0.08880394884	88.80394884	77.94228768
-0.5227576027	0.3005714736	0.1414956117	141.4956117	30.32479127
-0.04725195688	0.4811562091	0.1805847355	180.5847355	0.00095492350
0.4282536889	0.6657667841	0.184610575	184.610575	34.33088086
0.9037593347	0.8169384855	0.1511717014	151.1717014	29.84710384
1.37926498	0.9160934652	0.09915497968	99.15497968	18.78223643
1.854770626	0.9681855058	0.05209204065	52.09204065	18.55782533
2.330276272	0.990104223	0.02191871715	21.91871715	3.629022406
2.805781918	0.9974902674	0.007386044444	7.386044444	0.02017728348
3.281287564	0.9994833283	0.001993060902	1.993060902	0.000024159364
3.756793209	0.9999139477	0.000430619365	0.4306193657	0.4306193657
4.232298855	0.9999884343	0.000074486560	0.07448656064	11.49972719

At the $\alpha = 0.05$ level of significance and degrees of freedom 21, the critical χ^2 value is 32.7, taken from **Table IV** in the back of the textbook. Using the goodness of fit chi-squared test, we will be using the null and research hypotheses,

$$H_0 : \chi^2 \leq 32.7 \tag{3}$$

$$H_a : \chi^2 > 32.7 \tag{4}$$

Our χ^2 sum value was found to be 345.1245. The P-value was calculated (in Excel) to be 0. This means that there is a 0% probability we should accept our null hypothesis. Also, since our find χ^2 value is greater than the critical region of 32.7, we can state that we should reject the null hypothesis, and probably state that our distribution is **not normal**.

Assuming Gamma						
Estimated Mean	6.340728176	Bin Number:	Lower Bounds:	Raw Histogram	Gamma	(o-e)^2/e
Estimated Var	4.698103312	0	1.084990799	3	3.415113547	0.05045784109
Estimated Alpha	8.557673412	1	2.115654448	40	36.39801707	0.3564557106
Estimated Beta	0.7409406588	2	3.146318098	108	110.136552	0.04144722395
		3	4.176981747	172	178.5901528	0.2431831378
		4	5.207645397	207	199.2878738	0.298447114
Chi_square	9.94830601	5	6.238309046	181	173.3347246	0.3389767814
DoF: 12	Target X^2: 21.0	6	7.268972695	105	126.1665867	3.551054226
P value	0.6204958578	7	8.299636345	84	80.34395882	0.1663676686
		8	9.330299994	56	46.10857061	2.121956374
		9	10.36096364	21	24.34688049	0.4600839536
		10	11.39162729	13	12.00874145	0.0818231876
		11	12.42229094	7	5.595855111	0.3523362974
		12	13.45295459	2	2.485070185	0.09468267157
		13	14.48361824	0	1.058990875	1.058990875
		14	15.51428189	1	0.4354223837	0.7320429466
		15	16.54494554			

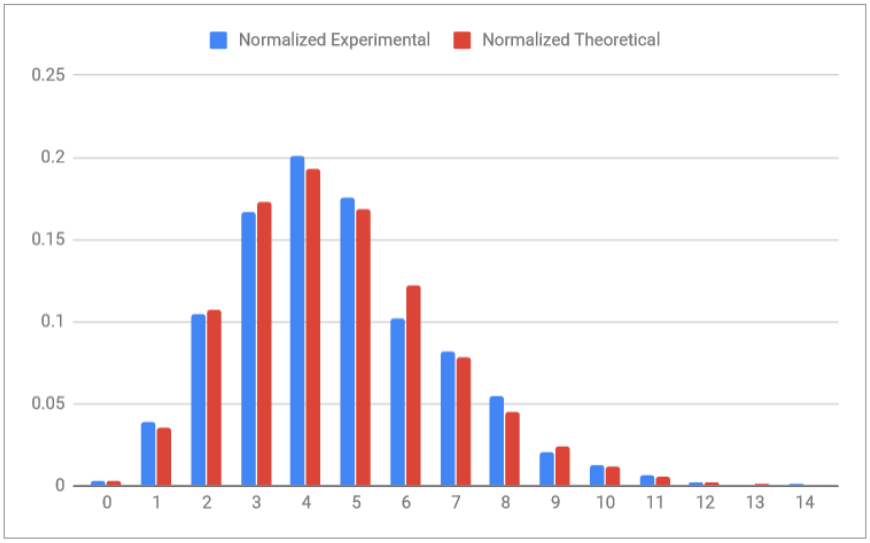
Similarly to the goodness of fit test on the *assumed normal* data, we will be assuming that this is **Gamma**. Using the sample mean and sample variance, we can determine what should be our theoretical Gamma distribution.

At the $\alpha = 0.05$ level of significance and degrees of freedom 12, the critical χ^2 value is 21.0, taken from **Table IV** in the back of the textbook. Using the goodness of fit test, our hypotheses are:

$$H_0 : \chi^2 \leq 21.0 \tag{5}$$

$$H_a : \chi^2 > 21.0 \tag{6}$$

Comparing our raw histogram data and theoretical gamma data, the χ^2 sum statistic we found is 9.9483, which is not within the critical region $\chi^2 > 21.0$. The P-value was found to be 0.6205, which is not statistically significant, meaning that we are able to accept our null hypothesis, and that our distribution is most likely ***Gamma*** over normal.



Normalized Experimental	Normalized Theoretical
0.002910745908	0.00331350926
0.03880994544	0.03531512641
0.1047868527	0.1068598393
0.1668827654	0.1732768521
0.2008414677	0.1933587877
0.1756150031	0.1681777801
0.1018761068	0.1224129586
0.08150088542	0.07795361645
0.05433392362	0.04473677774
0.02037522136	0.02362252759
0.01261323227	0.01165146502
0.006791740452	0.005429370788
0.001940497272	0.002411135957
0	0.001027484452
0.000970248636	0.0004224679739

Plotting our data, we also find that the graphs of the experimental and theoretical gamma data are very close to each other. Statistically, through the goodness of fit tests and the hypothesis testing, we found that the distribution is not normal, and *should* be gamma. Visually, the graph looks accurate to the theoretical gamma distribution.