Ziyue Chen PID: A14524683

Math 183 – Project 2

According to www.marketingcharts.com, the average 18–24-year-old has 649 Facebook friends. A student wanted to test if the mean number is higher at his school. Using the given data, set appropriate hypothesis in part a) and answer all the questions. Find the data file "Data Facebook Friends" on Canvas under Modules to view the data for the number of Facebook friends.

(Note: Round all your findings to two decimal places)

a) (2 points) Write appropriate hypothesis for the test.

$$H_0 : \mu = 649$$
$$H_A : \mu > 649$$

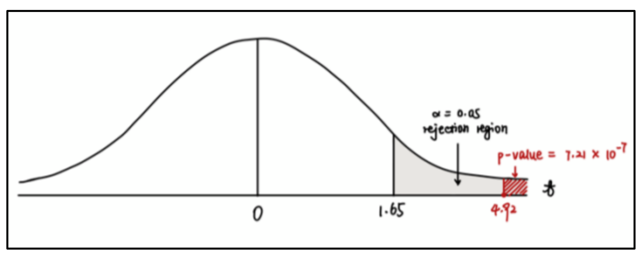
b) (2 points) The test statistics is:
(Include the formula)

$$\text{Unknown population variance: } T = \frac{\overline{X} - \mu}{\sqrt{S^2/n}} \sim t_{293}$$

c) (1 point) The p-value of the test statistics is:
(draw a diagram to illustrate)

$$\text{One-tail t-Test: } p - value = 7.21 \times 10^{-7}$$

t-Test: Mean		
		Column 1
Mean		751.4864
Standard Deviation		357.1355
Hypothesized Mean		649
df		293
t Stat		4.9205
P(T<=t) one-tail		7.21214E-07
t Critical one-tail		1.6501
P(T<=t) two-tail		0
t Critical two-tail		1.9681



d) (1 point) State your decision using both rejection region approach and p-value approach:
(Use α = .05)

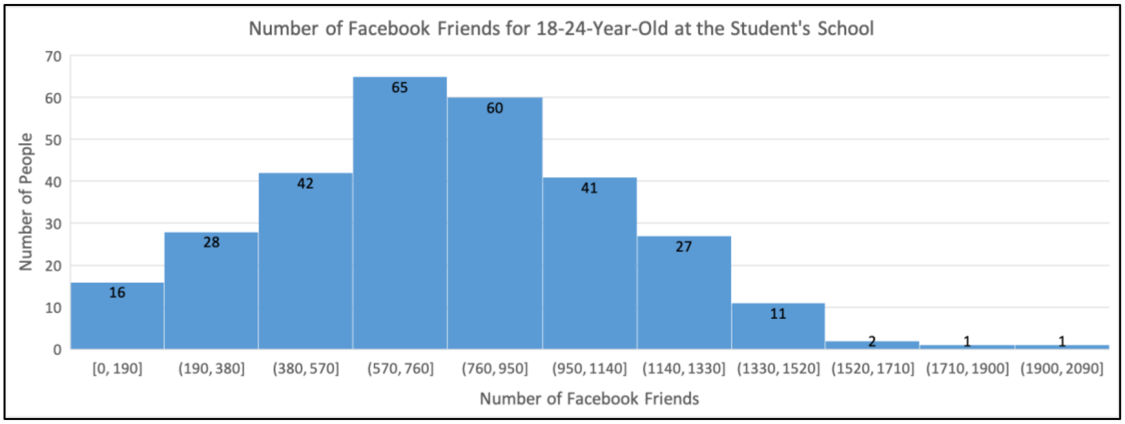
$$\text{Rejection region: reject } H_0 \text{ if } t_0 = \frac{\overline{x} - \mu_0}{\sqrt{s^2/n}} > t_{\alpha,293}$$
$$\text{p-value: reject } H_0 \text{ if } p - value = P(T > t_0) < \alpha$$

e) Interpret your decision made in part d).

$$\text{Rejection region: } t_0 = 4.92 > t_{\alpha,293} = 1.65$$
$$\text{p-value:} p - value = 7.21 \times 10^{-7} < \alpha = 0.05$$

Therefore, H_0 is rejected at the 5% level of significance, and there is sufficient evidence to conclude that the mean number of Facebook friends for students at the school is higher than 649.

f) (1 point) Provide a histogram of the sample.



g) (1 point) Comment on distribution shape of the sample.

The distribution of number of Facebook Friends for 18-24-year-old at the student's school is approximately normal.

h) (2 points) Apply an appropriate test to statistically confirm if data is normally distributed.

Chi-squared Test for Normality: H_0: The data is normally distributed.
H_A: The data is not normally distributed.

$$p - value = 0.85 > \alpha = 0.05$$

Since the p-value is greater than α, H_0 is not rejected at the 5% level of significance, and there is no evidence to conclude that the data is not normally distributed.

Chi-Squared Test of Normality			
	Column 1		
Mean	751.4863946		
Standard deviation	357.1355		
Observations	294		
Intervals	Probability	Expected	Observed
(z <= -2)	0.02275	6.6885	7
(-2 < z <= -1)	0.135905	39.95607	40
(-1 < z <= 0)	0.341345	100.35543	102
(0 < z <= 1)	0.341345	100.35543	97
(1 < z <= 2)	0.135905	39.95607	43
(z > 2)	0.02275	6.6885	5
chi-squared Stat	0.8118		
df	3		
p-value	0.8466		
chi-squared Critical	7.8147		

i) (2 points) Calculate 90% and 99% Confidence Intervals for the population mean. Interpret each of the Confidence Intervals.

$$90\%\text{ CI: } \bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} = 751.49 \pm t_{0.05,293}\frac{357.14}{\sqrt{294}}$$
$$= 751.49 \pm 1.65 \times \frac{357.14}{\sqrt{294}}$$
$$= (717.12, 785.86)$$
$$99\%\text{ CI: } \bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} = 751.49 \pm t_{0.005,293}\frac{357.14}{\sqrt{294}}$$
$$= 751.49 \pm 2.59 \times \frac{357.14}{\sqrt{294}}$$
$$= (697.48, 805.49)$$

Column1	
Mean	751.4863946
Standard Error	20.82856706
Median	747.5
Mode	0
Standard Deviation	357.1355296
Sample Variance	127545.7865
Kurtosis	0.031814353
Skewness	0.203431026
Range	2002
Minimum	0
Maximum	2002
Sum	220937
Count	294
Confidence Level(90.0%)	34.36861003

Column1	
Mean	751.4863946
Standard Error	20.82856706
Median	747.5
Mode	0
Standard Deviation	357.1355296
Sample Variance	127545.7865
Kurtosis	0.031814353
Skewness	0.203431026
Range	2002
Minimum	0
Maximum	2002
Sum	220937
Count	294
Confidence Level(99.0%)	54.00249085

The 90% CI means that if we repeat the samplings of the same size and generate CIs in the same way, 90% of the CIs will contain the true mean number of Facebook friends for 18-24-year-old. Similarly, the 99% CI means that if we repeat the samplings of the same size and generate CIs in the same way, 99% of the CIs will contain the true mean number of Facebook friends for 18-24-year-old.

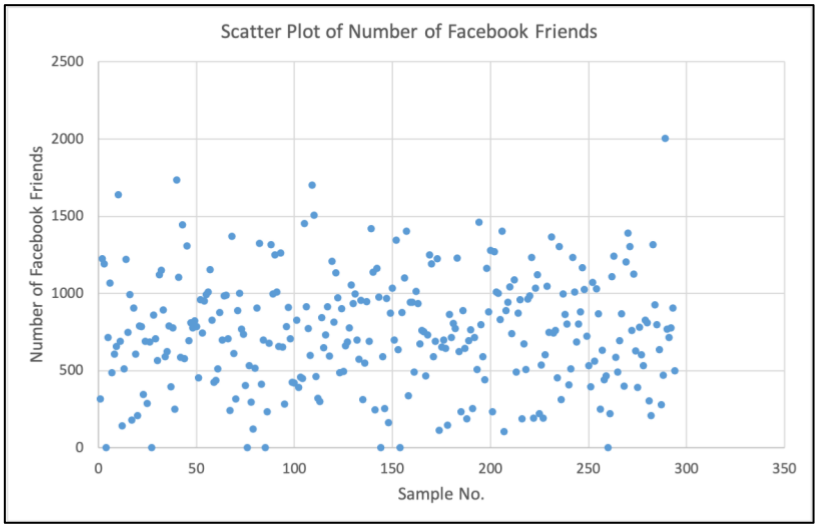
j) (1 point) Compare the 90% and 99% Confidence Intervals for the population mean. Which one gives a better idea about what the population mean is? Explain.

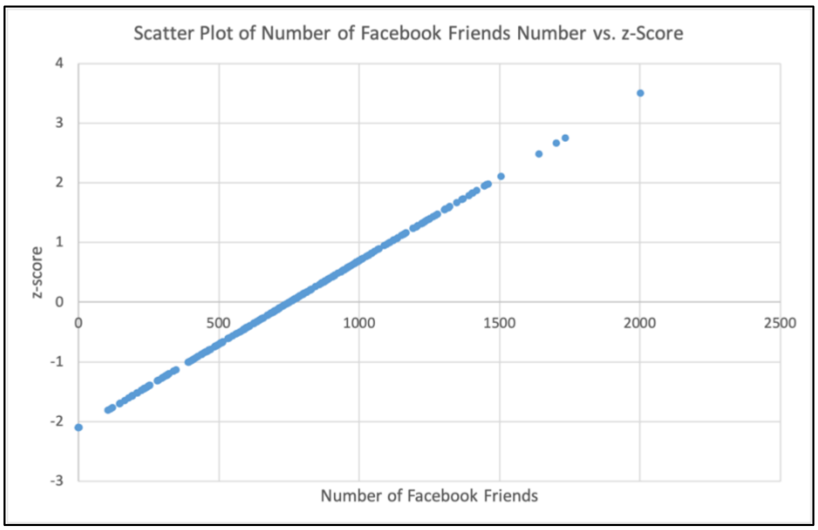
The 90% CI is narrower than the 99% CI. Since the 99% CI is wider, it is more likely to contain the true mean and thus gives a better idea about what the population mean is.

k) (2 points) Provide scatter plot of the sample observations. Identify the outliers in this sample. (Hint: Compute z-scores of your observations and consider any z-scores bigger than 3 or less than -3 as outliers)

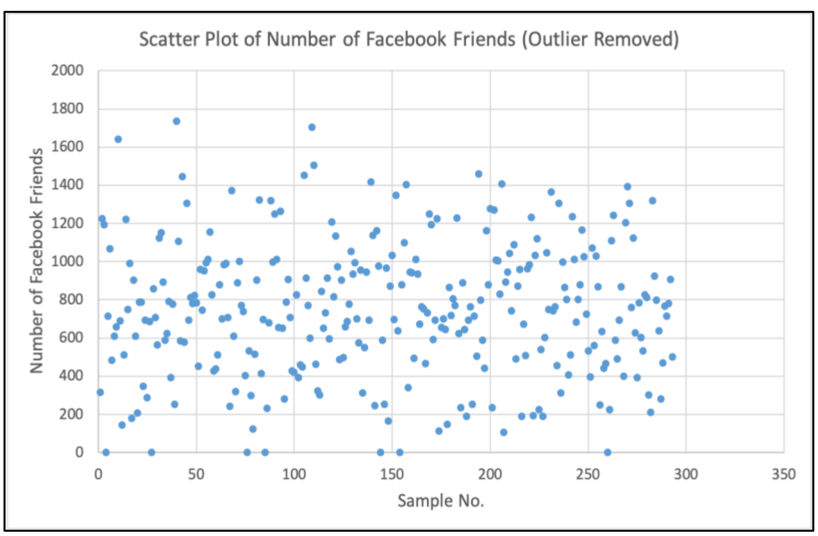
Sample No. 289, who has 2002 Facebook friends, is an outlier because the corresponding z-score of 3.50 is greater than 3.

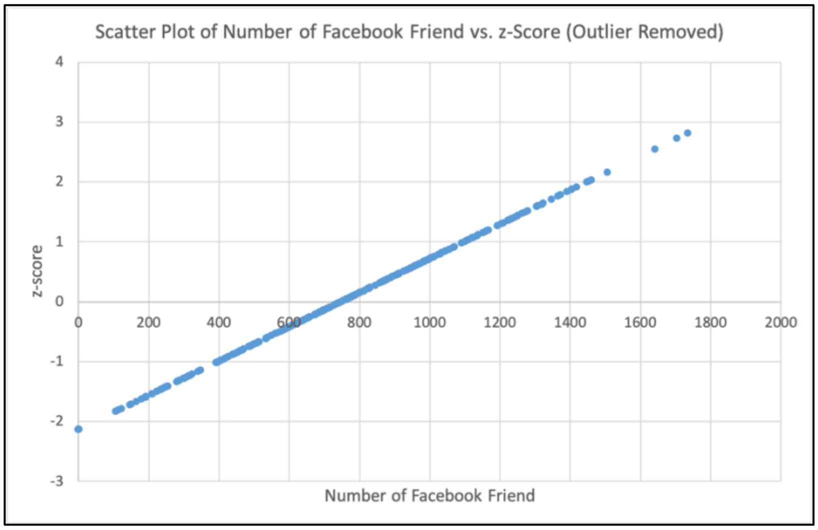
Ziyue Chen PID: A14524683





l) (1 point) Drop the identified outliers in part k) and provide scatter plot of the observations again.





m) (3 points) After dropping the outliers, answer parts h), i) and j) again. Explain if removing the outliers affected your findings in part h), i) and j)? Explain all the statistical differences that you might have observed, if any.

Chi-squared Test for Normality: H_0: The data is normally distributed.
H_A: The data is not normally distributed.

$$p-value = 0.92 > \alpha = 0.05$$

Since the p-value is greater than α, H_0 is not rejected at the 5% level of significance, and there is no evidence to conclude that the data is not normally distributed.

Chi-Squared Test of Normality			
	Column 1		
Mean	747.21843		
Standard deviation	350.1555		
Observations	293		
Intervals	Probability	Expected	Observed
(z <= -2)	0.02275	6.66575	7
(-2 < z <= -1)	0.135905	39.820165	41
(-1 < z <= 0)	0.341345	100.014085	99
(0 < z <= 1)	0.341345	100.014085	97
(1 < z <= 2)	0.135905	39.820165	43
(z > 2)	0.02275	6.66575	6
chi-squared Stat	0.4733		
df	3		
p-value	0.9247		
chi-squared Critical	7.8147		

$$\text{90\% CI: } \bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} = 747.22 \pm t_{0.05,292}\frac{350.16}{\sqrt{293}}$$
$$= 747.22 \pm 1.65 \times \frac{350.16}{\sqrt{293}}$$
$$= (713.46, 780.97)$$

$$\text{99\% CI: } \bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} = 747.22 \pm t_{0.005,292}\frac{350.16}{\sqrt{293}}$$

$$= 747.22 \pm 2.59 \times \frac{350.16}{\sqrt{293}}$$
$$= (694.18, 800.26)$$

Column1	
Mean	747.21843
Standard Error	20.45630352
Median	746
Mode	0
Standard Deviation	350.1555135
Sample Variance	122608.8836
Kurtosis	-0.269900355
Skewness	0.096341362
Range	1735
Minimum	0
Maximum	1735
Sum	218935
Count	293
Confidence Level(90.0%)	33.7547155

Column1	
Mean	747.21843
Standard Error	20.45630352
Median	746
Mode	0
Standard Deviation	350.1555135
Sample Variance	122608.8836
Kurtosis	-0.269900355
Skewness	0.096341362
Range	1735
Minimum	0
Maximum	1735
Sum	218935
Count	293
Confidence Level(99.0%)	53.03850851

The 90% CI means that if we repeat the samplings of the same size and generate CIs in the same way, 90% of the CIs will contain the true mean number of Facebook friends for 18-24-year-old. Similarly, the 99% CI means that if we repeat the samplings of the same size and generate CIs in the same way, 99% of the CIs will contain the true mean number of Facebook friends for 18-24-year-old.

The 90% CI is narrower than the 99% CI. Since the 99% CI is wider, it is more likely to contain the true mean and thus gives a better idea about what the population mean is.

After removing the outlier of 2002, the null hypothesis that the data is normally distributed is still not rejected, and the p-value is even greater, indicating that it is more possible that the null hypothesis is true. This is reasonable because the outlier, which was originally pulling up the mean and standard deviation, is removed to generate a more normally distributed set of data. Similarly, both 90% and 99% CIs become narrower than before, and the lower bounds and upper bounds all have smaller values due to the removal of the outlier which has a large value. The 90% CI is still narrower than the 99% CI, and the 99% CI should still give a better idea about the population mean as it is wider.