Riley Payung

STAT 250-003

Data Analysis Assignment #3

Problem #1

From a July 2019 survey of 1186 randomly selected Americans ages 18-29, it was discovered that 248 of them vaped (used an e-cigarette) in the past week.

1A

Sample Proportion (p-hat): 0.2091

1B

The central limit theorem states that the problem must be randomly sampled and independent, of which it is, since it says in the problem that 1186 'randomly selected Americans…' were chosen.

The Central limit theorem also states that the number of successes (Americans who vaped) and the number of failures (Americans who did not vape) must be at least 10, which it is: success(N) = 248; failure(N) = 938.

Finally, the central limit theorem states that the population must be 10x greater than the sample size, which it is; the population of Americans is exorbitantly bigger than the sample size.

1C

SE = sqrt((0.2091(1-0.2091) / 1186) = 0.0118

CI = p-hat +- 1.645 * SE = 0.2091 +- 1.645 * 0.0118 = 0.2091 +- 0.019411 = (0.1897,0.2285)

1D

One sample proportion summary confidence interval:

p: Proportion of successes
Method: Standard-Wald

90% confidence interval results:

Proportion	Count	Total	Sample Prop.	Std. Err.	L. Limit	U. Limit
p	248	1186	0.20910624	0.011808649	0.18968274	0.22852974

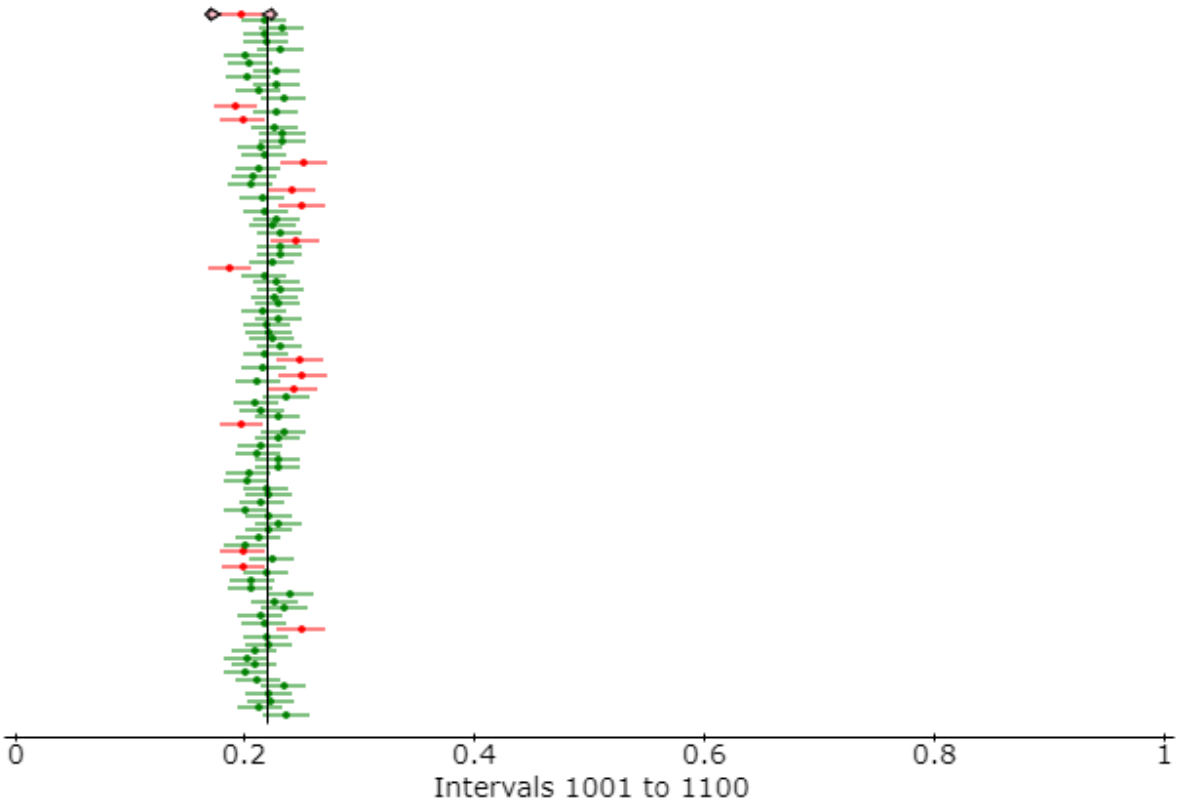
1E

We are 90% confident that the interval captures the unknown population proportion of Americans who vape.

1F

Confidence intervals for p, p=0.22, Type=Standard-Wald
Sample size=1186

CI Level	Containing p	Total	Proportion
0.9	969	1100	0.8809



1G

We are about 88.09% confident that the simulation of this confidence interval captures the known population proportion of 0.22 of Americans who vaped in the past week.

1H

In the long run, if we were to continue sampling, we expect to be about 90% confident that our interval will continue to capture the unknown population proportion of Americans who vape between the ages of 18-29.

Problem #2

GMU began a robot food delivery service in January 2019. It is expected that your food or drink will be delivered in around 30 minutes. The management team is considering a new policy for 2020: if you do not receive your items in at most 45 minutes, you will not have to pay the delivery fee*. To test this, the management team collected a random sample of 432 orders and stored the data in StatCrunch. The responses are 0 = delivery took less than or equal to 45 minutes and 1 = delivery took more than 45 minutes and the data set called "Food Delivery Robots." *(please note, this is not a real policy under consideration).

2A

Sample proportion of robots taking more than 45 minutes to deliver food (p-hat): 0.1856

2B

The population parameter is the proportion of deliveries that take more than 45 minutes out of all deliveries that are done.

2C

H_0 : mu = 0.14

H_A : mu > 0.14

2D

α = 0.01

2E

Since we have a sample size of 432 random orders, the problem satisfies the condition of independent and randomly sampled.

The problem satisfies the needed number of successes (i.e. orders that take longer than 45 minutes) of 80, and 352 failures.

Since were assuming the population is large, the third condition is automatically true.

2F

Test Statistic: Z = 0.1856 – 0.14 / sqrt(0.14 * 0.86 / 432) = 0.0456 / 0.016694 = 2.73.

2G

P-value = 1 – P(Z < 2.73) = 1 - .9968 = 0.0032

2H

Based on the P-value of 0.0032 being less than α = 0.01, p-value < α, we will reject the null hypothesis that mu = 0.14, and can conclude that the alternative, that mu > 0.14 is true.

2I

Based on our conclusion, we can conclude that the average proportion of deliveries that are completed take longer than 45 minutes is greater than 0.14.

2J

One sample proportion hypothesis test:

Outcomes in: Time
Success: 1
p : Proportion of successes
H_0 : p = 0.14
H_A : p > 0.14

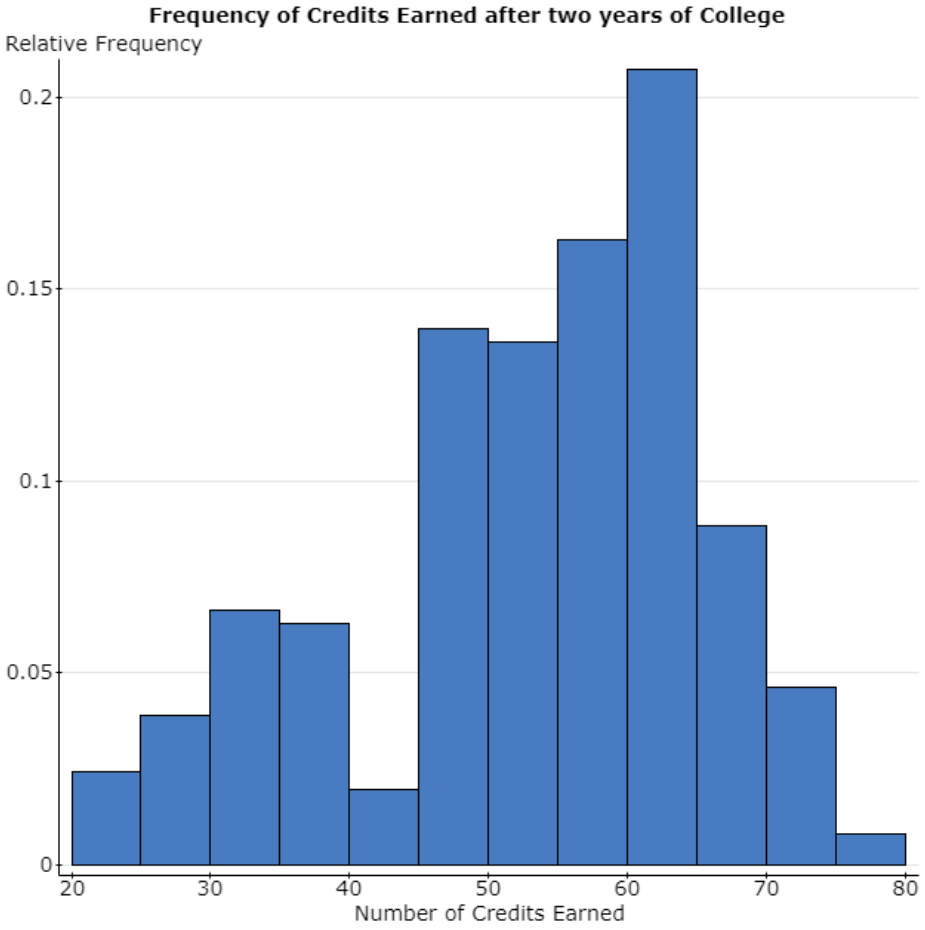
Hypothesis test results:

Variable	Count	Total	Sample Prop.	Std. Err.	Z-Stat	P-value
Time	80	432	0.18518519	0.016694421	2.7066039	0.0034

Problem #3

A midsized university collected data on all 10,128 seniors attending. One variable of interest was how many credits they earned after they completed two years of school (i.e. credits they earned by their junior year).

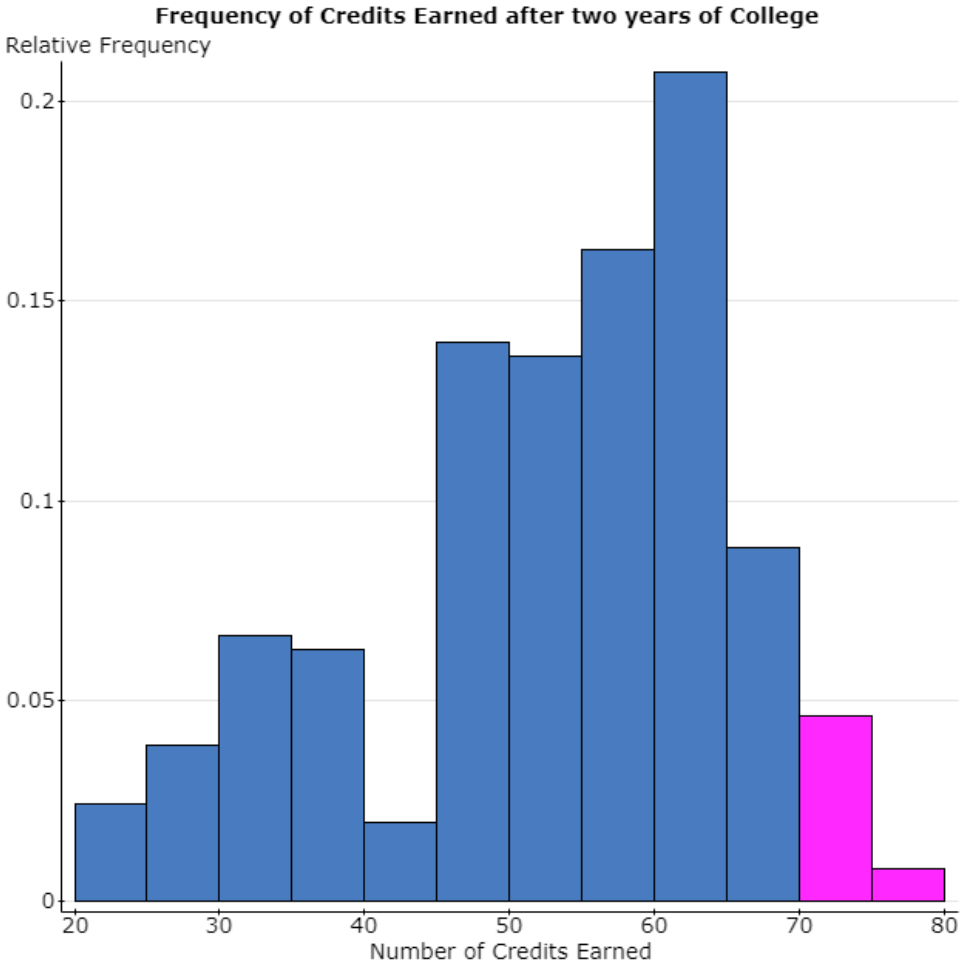
3A



3B

The shape of the graph is relatively normal, with a center around 60 credits and a minor bi-modal trend centered around 30 credits; it does not seem to be skewed.

The proportion of individuals who have completed 70 or greater credits is 0.054.



3C

Summary statistics:

Column	Mean	Std. dev.
Credits	52.44	12.54

These calculations are parameters since they are pulled directly from a population.

3D

Summary statistics:

Column	Mean	Std. dev.
Sample(Credits)	56.57	6.13

These calculations are statistics since they are sampled from a population.

3E

The sample mean in (d) does in fact come from a normal sampling distribution.

The first condition is true since we randomly sampled 7 individuals from a population of 10,128 seniors.

Secondly, our sample size is only 7, therefore we have too small a sample to conclude that the second condition holds.

Lastly, the population size of 10,128 students is large enough, so our third condition holds.

3F

Summary statistics:

Column	Mean	Std. dev.
Sample 2	56.36	9.71

These calculations are statistics.

3G

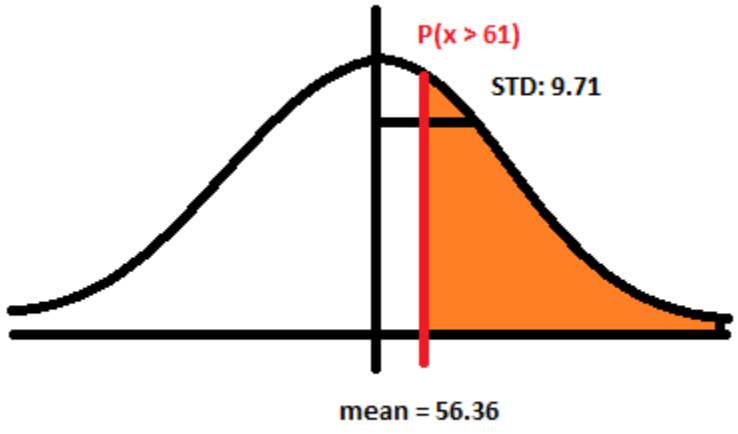
The sample mean in (f) does in fact come from a normal sampling distribution.

The first condition is true since we randomly sampled 7 individuals from a population of 10,128 seniors.

Secondly, our sample size is 28, therefore we have a significant sample size to conclude that the second condition holds.

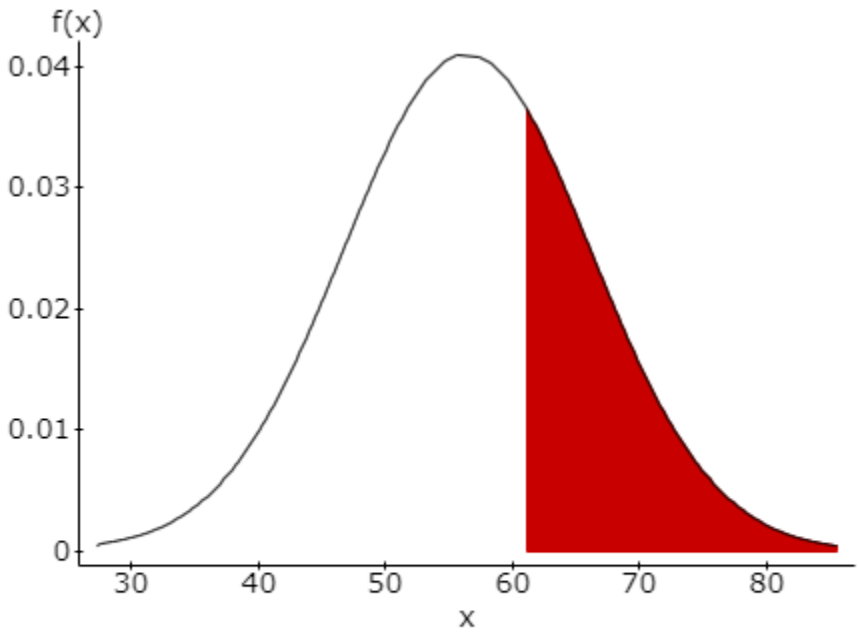
Lastly, the population size of 10,128 students is large enough, so our third condition holds.

3H



P (Z > 61) = 1 – P (Z <= 61) = 61 – 56.36 / 9.71 = 0.48 = 1 - .6844 = 0.3156

3I



3J

The probability we obtained in (b) was based on >= 70 credits, whereas the ones in (h) and (I) were based on >= 61 credits, but we can conclude that the sampling distribution has an accurate representation of the population proportion.

Problem #4

A random sample of 22 STAT 250 students was collected and the file size of Data Analysis 2 was recorded. The data was measured in megabytes. The instructors of the course claim that the file size will be different from 5 megabytes. Consider the population of all file sizes to be right skewed. Using α = 0.01, is there sufficient evidence to conclude that the mean file size of Data Analysis 2 is different from 5 megabytes? Conduct a full hypothesis test by following the steps below. Enter an answer for each of these steps in your document.

4A

The population parameter is the students who submitted Data analysis 2.

4B

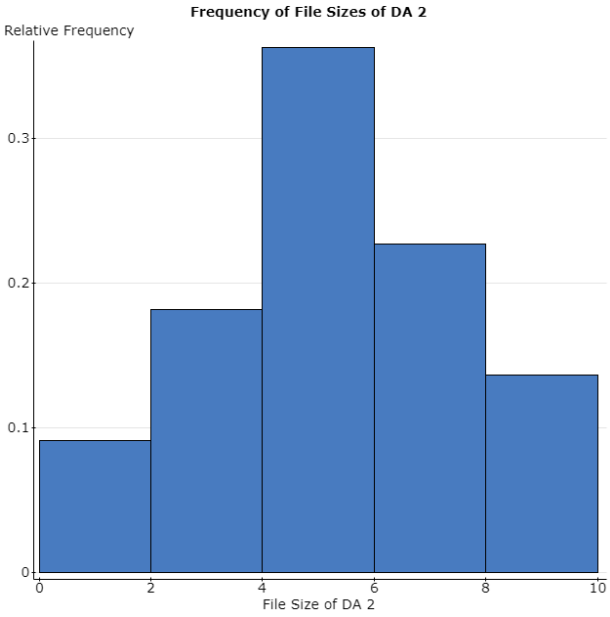
H_0 : mu = 5

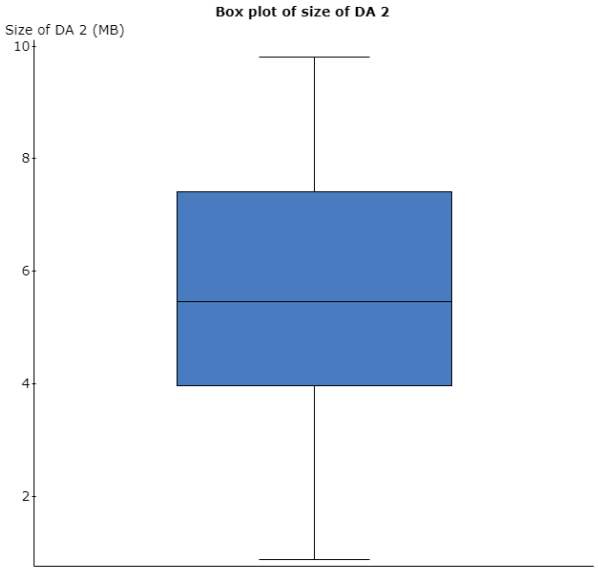
H_A : mu != 5

4C

The significance level is α = 0.01.

4D





4E

Because we have no outliers, and we have a normally shaped graph, we can confirm that we can perform a T-test on this problem.

4F

Mean = 5.3895455; Std = 2.3824507

T = 5 – 5.3895455 / (2.3824507 / sqrt (22)) = 0.3895455 / 0.507940 = 0.77

4G

One sample T hypothesis test:

μ: Mean of variable
H_0: μ = 5
H_A: μ ≠ 5

Hypothesis test results:

Variable	Sample Mean	Std. Err.	DF	T-Stat	P-value
Megabytes	5.39	0.51	21	0.77	0.4517

4H

P-value = 0.4517

4I

Since we have a p-value of 0.4517 which is greater than our alpha of 0.01 (P-value 0.4517 > α = 0.01), we cannot reject the null hypothesis.

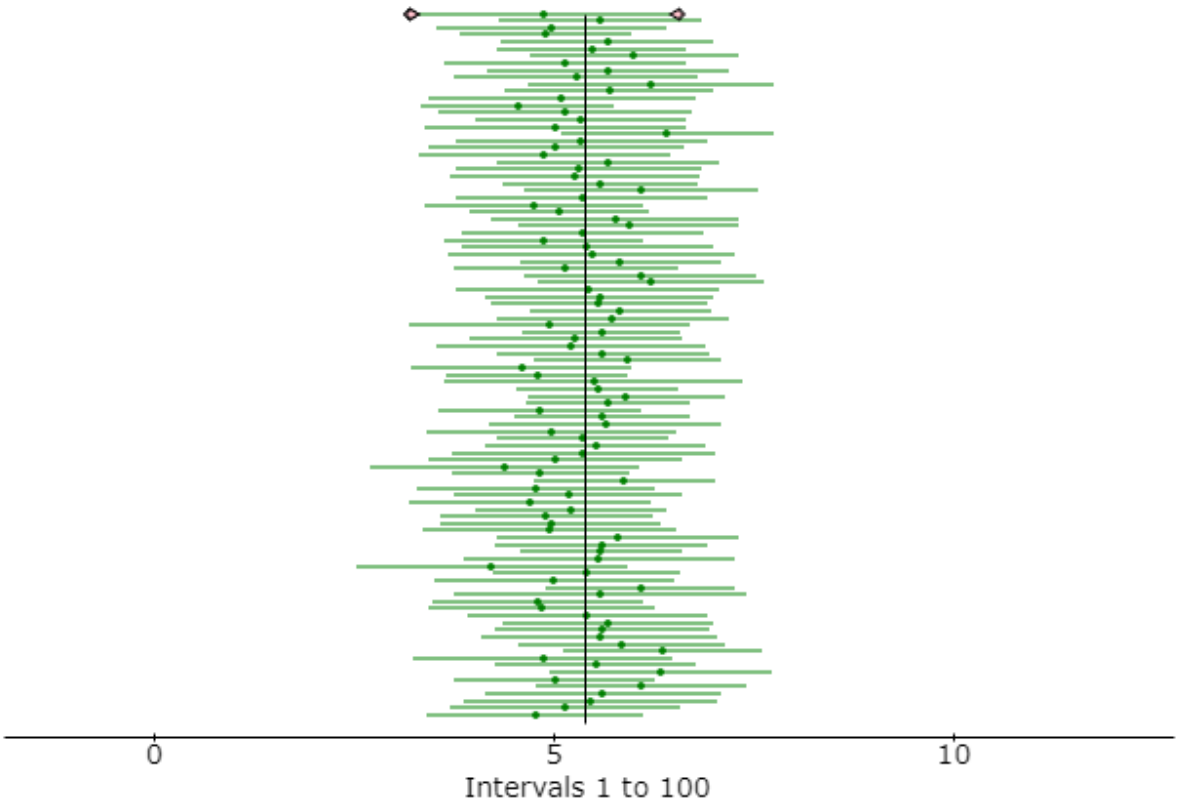
4J

In conclusion, we must accept that the file size will not be different from 5 megabytes since we were not able to reject our null hypothesis.

4K

Confidence intervals a mean: Megabytes (μ=5.39, σ=2.382) Type=T
Sample size=22

CI Level	Containing μ	Total	Proportion
0.99	100	100	1



4L

From this confidence interval, we can be 99% confident that we are capturing the known mean of 5 megabytes for each and every data analysis that was submitted here, furthering our argument that we cannot reject the null hypothesis.