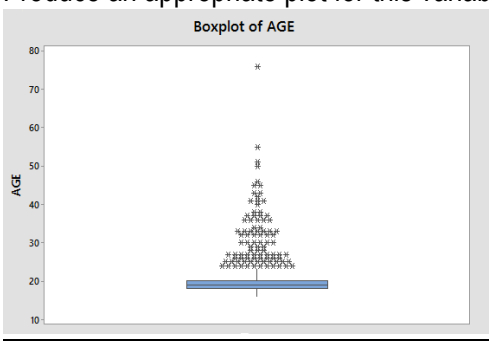
Name:
E number:
Math 1530 section number

MATH 1530 CAPSTONE TECHNOLOGY PROJECT FALL 2017

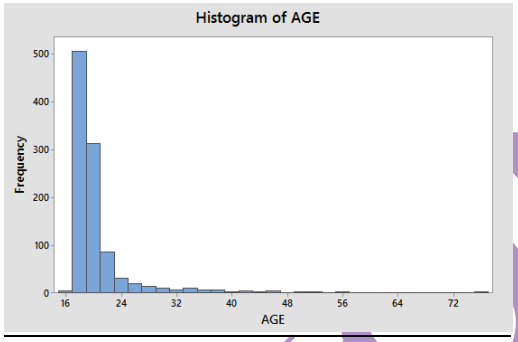
Problem 1: Identify Variable Type. Which of these questions from the class survey produced variables that are categorical and which are quantitative? Use your word processor to underline/highlight the best option.

a. AGE Categorical Quantitative Neither

Produce an appropriate plot for this variable.

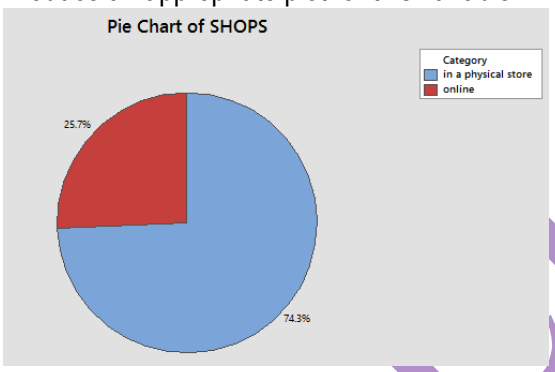


OR

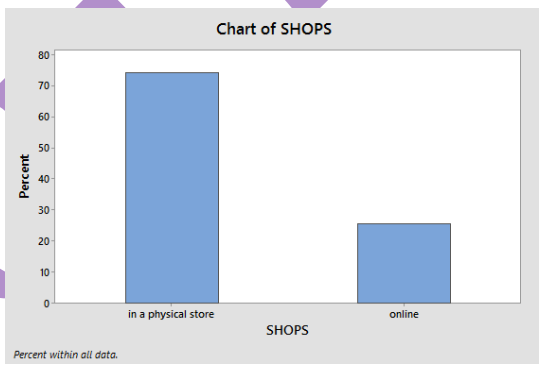


b. SHOPS Categorical Quantitative Neither

Produce an appropriate plot for this variable.

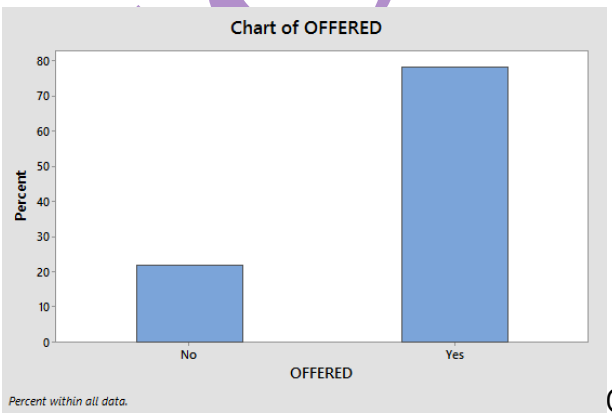


OR

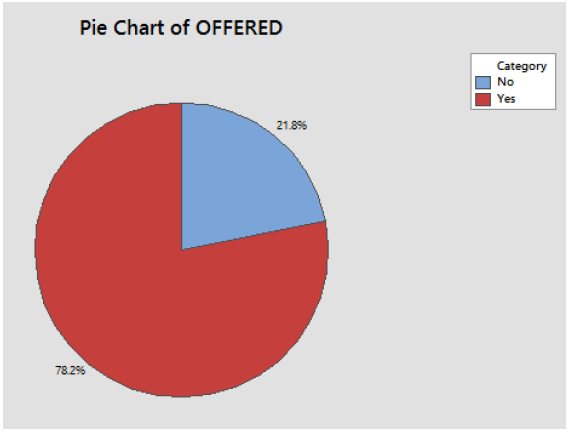


c. OFFERED Categorical Quantitative Neither

Produce an appropriate plot for this variable.



OR



Name:
E number:
Math 1530 section number

Problem 2: Sampling. In the survey data, the variable "**AGE**" is the age in years for each student.

a. Type the first 10 observations from the column representing the variable **AGE** into the table below, and use this as your sample data for part (b). Then calculate the mean age of these first 10 observations and report the value below.

N	1	2	3	4	5	6	7	8	9	10
AGE	**43**	**18**	**18**	**18**	**18**	**25**	**18**	**20**	**25**	**21**

b. The mean age of the first 10 students is **22.4.** (Type the value into the space provided.)

c. Next, select a random sample of size n = 10 (Go to Calc > Random Data > Sample from Columns). Type the number 10 in the "Number of rows to Sample" slot. Enter the variable "**ID**" and "**AGE**" into the "From columns" slot. Enter C17-C18 into the "Store samples in" slot. Record the data for your sample in the table below.

N	1	2	3	4	5	6	7	8	9	10
ID	**864**	**632**	**164**	**860**	**134**	**618**	**720**	**847**	**73**	**2**
AGE	**19**	**22**	**19**	**19**	**20**	**17**	**20**	**18**	**18**	**18**

d. Calculate and report the mean age for your random sample of 10 students. The sample mean age is **19**.
ANSWERS WILL VARY.

e. Suppose we think of *all* the students who responded to the survey as a *population* for the purposes of this problem. In that case, the *population mean* age is 20.001. Discuss (two or more complete sentences) the **differences and similarities** between 20.001 and the answers you got in (b) and (d).

Instructors will need a bit of flexibility in how to interpret this one's answer.

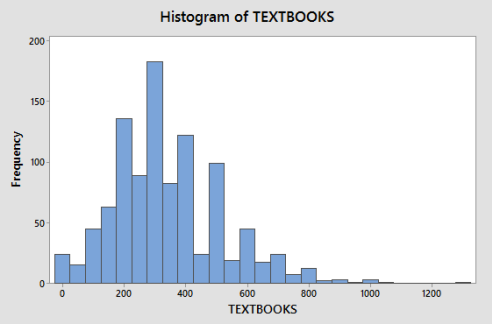
The 'convenience sample' mean found in (b.) 22.4 overestimates the 'population' mean 20.0001, but that hardly matters. As it is not a random sample, there is no long-run guarantee that means from such samples would or would not come close to the population mean. The 'SRS' mean found in part (d.) of *19* is a little *below* the population mean of 20.0001. However, in the 'long run,' the distribution of sample means centered around the population mean. *Students* may further remark that more samples would have a more even mix of $\bar{x}$ values above and below the population mean.

Name:
E number:
Math 1530 section number

Problem 3(e): If your E number ends in an even number (0, 2, 4, 6, or 8) then do this question. (Omit this page/problem if your E# ends with an odd number.)

1. Question 3 of the FALL 2017 survey asked students, "Approximately how much did you spend on textbooks this semester (Round to the nearest whole dollar)?"

a. Create an appropriate graph to display the *distribution* of the variable called **TEXTBOOKS** and insert it here.



b. Which of the following best describes the shape of the distribution? Underline (or highlight) your answer.

Skewed left Uniform Skewed right Bimodal Symmetric

c. Using Minitab, calculate the basic statistics for the data collected on **TEXTBOOKS**. Copy and paste all of the Minitab output here.

Descriptive Statistics: TEXTBOOKS

Variable	N	N*	Mean	StDev	Variance	Minimum	Q1	Median	Q3	Maximum
TEXTBOOKS	1017	0	341.13	179.86	32348.18	0.00	200.00	300.00	433.00	1300.00

Variable	IQR
TEXTBOOKS	233.00

Choose statistics that are appropriate for the shape of the distribution to describe the center and spread of **TEXTBOOKS**.

d. Which statistic will you use to describe the center of the distribution? **Median**

e. In one or two sentences, describe why this statistic was chosen. **Since the shape of the distribution is skewed right, the median should be used to describe the center of the distribution instead of the mean because the median is robust to outliers while the mean is highly affected by outliers.**

f. What is the value of that statistic? **300**

g. Which statistic(s) will you use to describe the spread of the distribution? **Q1, Q3, and possible IQR**

h. What is (are) the value(s) of that (those) statistic(s)? **Q1 = 200, Q3 = 433, and possible IQR = 233**

Name:
E number:
Math 1530 section number

i. Are there any outliers in this distribution? Justify your answer using the IQR rule or an appropriate plot.

IQR rule says that any value *below* Q1 – 1.5*IQR or *above* Q3 + 1.5*IQR are outliers.

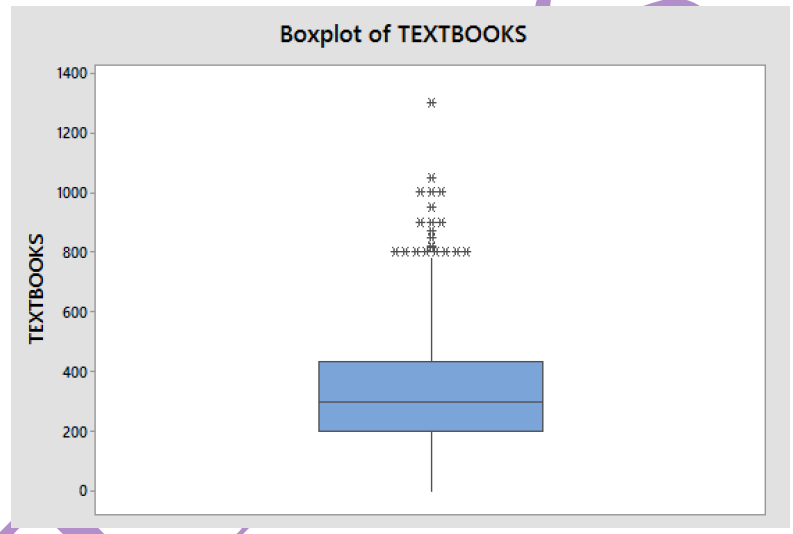
IQR = Q3 – Q1 = 433 – 200 = 233, so 1.5 * IQR = 1.5 * 233 = 349.5.

Q1 – 1.5 * IQR = 200 – 349.5 = –149.5 and Q3 + 1.5 * IQR = 433 + 349.5 = 782.5

Any value of 'TEXTBOOKS' *below* -149.5 or *above* 782.5 would be considered outliers.

Yes, there are definitely outliers in the distribution of 'TEXTBOOKS' on the upper end.

Minitab shows outliers with * on a boxplot:

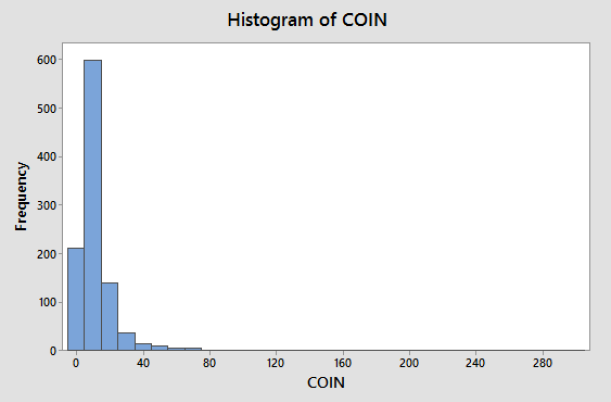


Name:
E number:
Math 1530 section number

Problem 3(o): If your E number ends in an odd number (1, 3, 5, 7, or 9) then do this question. (Omit this page/problem if your E# ends with an even number.)

Question 15 of the FALL 2017 survey asked students, "Flip a fair coin until you get three heads in a row. How many flips did it take before you got three heads in a row?"

a. Create an appropriate graph to display the *distribution* of the variable called **COIN** and insert it here.



b. Which of the following best describes the shape of the distribution? Underline (or highlight) your answer.

Skewed left Uniform Skewed right Bimodal Symmetric

c. Using Minitab, calculate the basic statistics for the data collected on **COIN**. Copy and paste all of the Minitab output here.

Descriptive Statistics: COIN

Variable	N	N*	Mean	StDev	Variance	Minimum	Q1	Median	Q3	Maximum	IQR
COIN	1017	0	10.733	13.467	181.369	0.000	5.000	7.000	12.000	300.000	7.000

Choose statistics that are appropriate for the shape of the distribution to describe the center and spread of **COIN**.

d. Which statistic will you use to describe the center of the distribution? Median

e. In one or two sentences, describe why this statistic was chosen. **Since the shape of the distribution is skewed right, the median should be used to describe the center of the distribution instead of the mean because the median is robust to outliers while the mean is highly affected by outliers.**

f. What is the value of that statistic? 7

g. Which statistic(s) will you use to describe the spread of the distribution? Q1, Q3, and possible IQR

h. What is (are) the value(s) of that (those) statistic(s)? Q1 = 5, Q3 = 12, and possible IQR = 7

Name:
E number:
Math 1530 section number

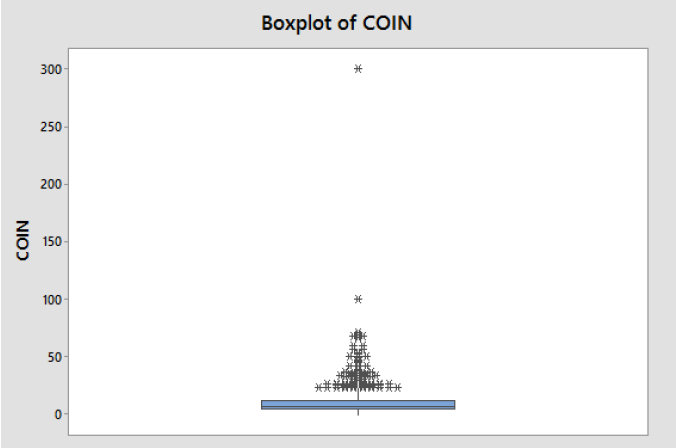
i. Are there any outliers in this distribution? Justify your answer using the IQR rule or an appropriate plot.

IQR rule says that any value *below* Q1 – 1.5*IQR or *above* Q3 + 1.5*IQR are outliers.

IQR = Q3 – Q1 = 12 – 5 = 7, so 1.5 * IQR = 1.5 * 7 =10.5.

Q1 – 1.5 * IQR = 5 – 10.5 = –5.5 and Q3 + 1.5 * IQR = 12 + 10.5 = 22.5

Any value of 'COIN' *below* -5.5 or *above* 22.5 would be considered outliers.



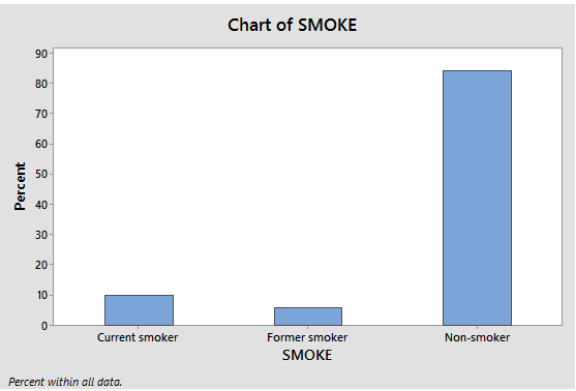
Yes, there are definitely outliers in the distribution of 'COIN' on the upper end.

Minitab shows outliers with * on a boxplot:

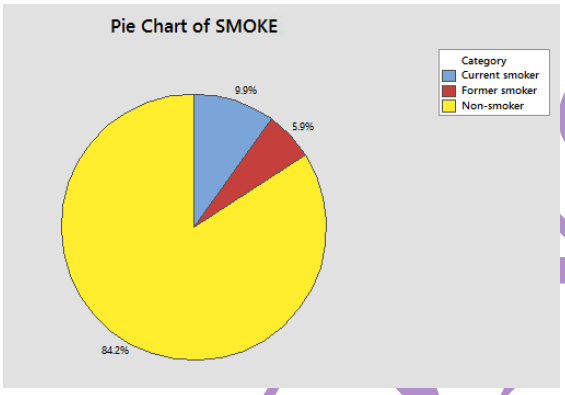
Name:
E number:
Math 1530 section number

Problem 4: SMOKE versus AGE. Question 1 of the survey asked students, "What is your age (in years)?" Question 12 of the survey asked students, "What is your current smoking status? (Current smoker, Former smoker, Non-smoker)"

a. Create a suitable graph to display the *distribution* of **SMOKE** and insert it here.



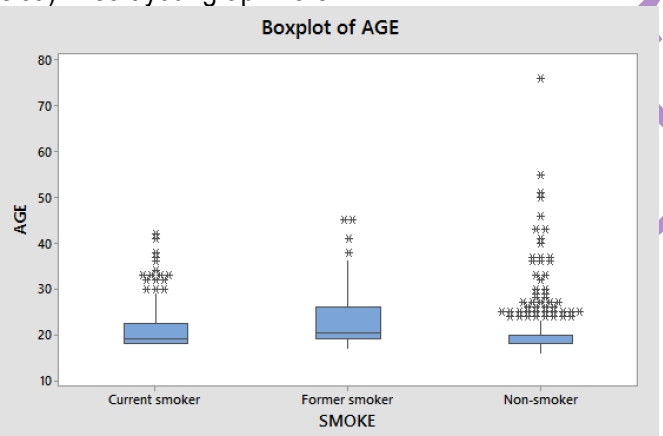
OR



b. What is the mode of this distribution? Underline (or highlight) one option.

Current Smoker Former Smoker Non-smoker

c. Create a side-by-side boxplot to display the age of students for the different levels of **SMOKE**. (Go to Graph > Boxplot > One Y with Groups > OK. Select **AGE** for the "Graph variables" slot and **SMOKE** for the "Categorical variables for grouping" slot.) Insert your graph here.



Use the side-by-side boxplot found in part (c) to answer the following questions.

d. Which smoking status has the oldest student? Non-smoker

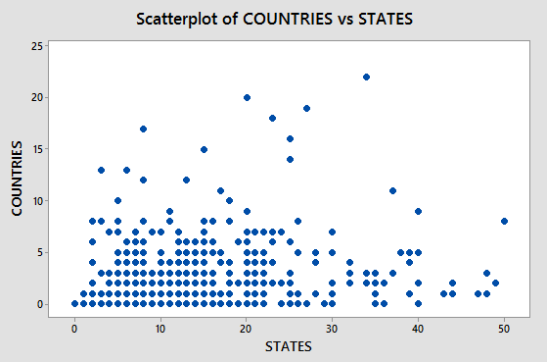
e. Which smoking status has the highest median age? Former smoker

f. Which smoking status has the smallest IQR? Non-smoker

Name:
E number:
Math 1530 section number

Problem 5: STATES vs. COUNTRIES. On the FALL 2017 Math 1530 survey, question 8 asked students, "How many U.S. states have you visited, not including the state you were born? (Enter zero if you have never traveled outside the state you were born.)" and question 9 asked students, "How many countries have you visited, not including the country you were born? (Enter zero if you have never traveled outside the country you were born.)" We are interested in seeing whether we can use the number of states visited to predict the number of countries visited.

a. Create an appropriate graph to display the relationship between **STATES** and **COUNTRIES**. Insert it here.



b. Does the plot show a positive association, a negative association, or no association between these two variables? EXPLAIN what this means with respect to the variables being studied.

Positive association: As the number of states visited increases, the number of countries visited increases.

c. Describe the *form* of the relationship between **STATES** and **COUNTRIES.**

Linear

d. Report the value of the correlation between this pair of variables? ***r* = 0.282**

e. Based on the information displayed in the graph and the correlation you just reported, how would you describe the *strength* of the association?

The strength is very weak.

f. Using Minitab, obtain the equation for the least squares regression of **STATES** on **COUNTRIES**. Copy & paste the output here.

```
Regression Equation

COUNTRIES = 0.467 + 0.0957 STATES
```

g. Interpret the value of the slope in the least squares regression equation you found in part (f).

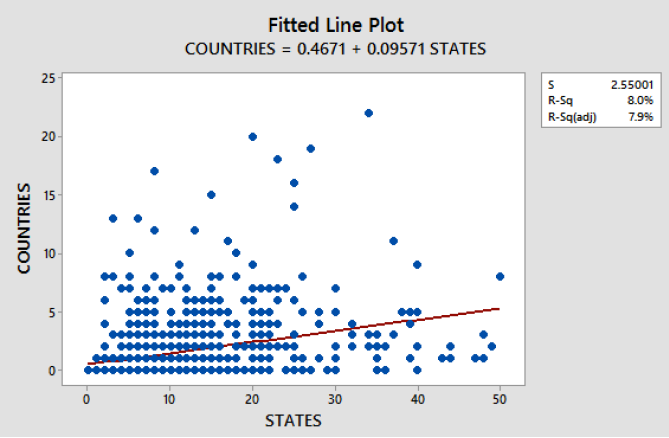
For every additional state visited, the estimated number of countries visited increases by 0.0957 country.

Name:
E number:
Math 1530 section number

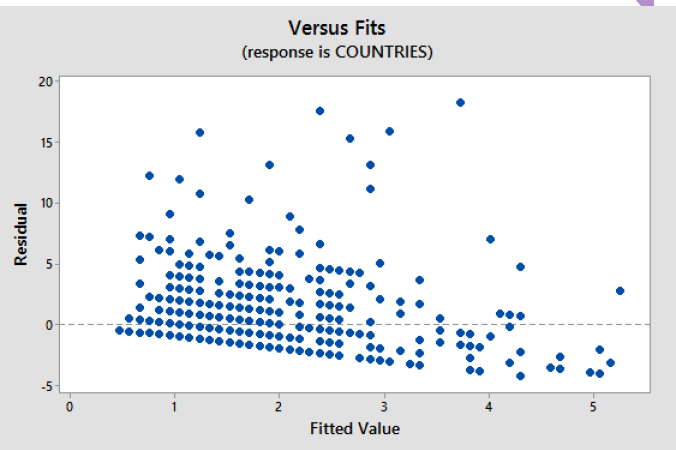
h. Use the regression equation in part (f) to predict the number of countries visited for a student that has visited 8 U.S. states. (Show your math.)

Predicted number of countries visited = 0.467 + 0.0957*8 = 1.2326 countries

i. How well does the regression equation fit the data? Explain. Justify your answer with appropriate plot(s) and summary statistics.



The association is a weak one and can be seen clearly in the fitted line plot. There are several points that are scattered far away from the regression line. The squared correlation (R^2) indicates that 7.98% of the variation we observed in the number of countries visited is explained by the linear relationship with the number of states visited.



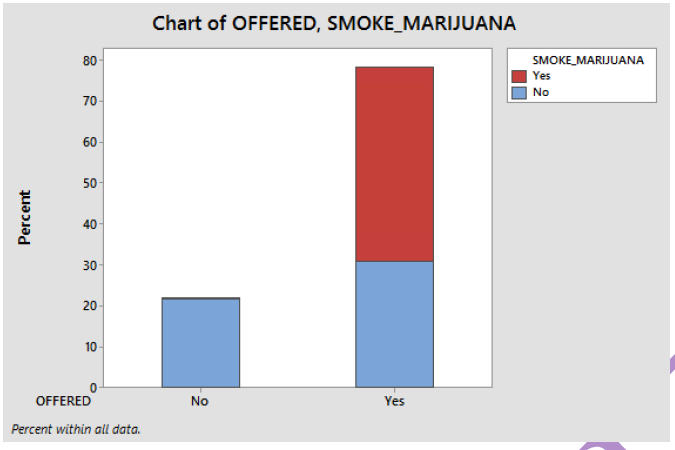
Note: Another scatterplot that is useful to see whether the model makes sense is the residual plot. This helps in determining the appropriateness of the regression model. Recall that the residuals are Residual = Observed Data – Predicted Data. The residual plot shouldn't have any interesting features, like direction or shape. It should stretch horizontally with about the same amount of scatter about the horizontal line at 0. There should be no bends and no outliers. We see that the plot above may possibly be cause to worry.

Name:
E number:
Math 1530 section number

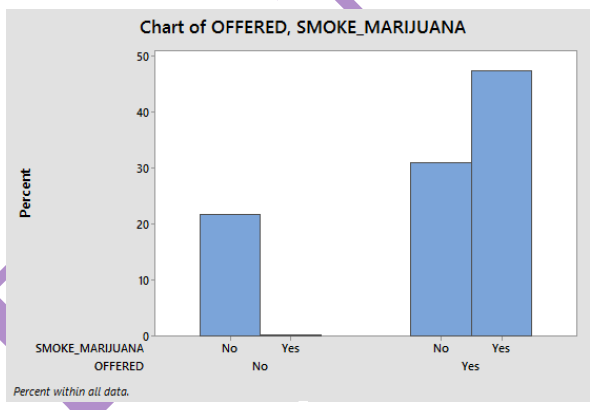
Problem 6 (T): Flip a fair coin. If it lands on tails do this problem (Omit this page/problem AND DO PROBLEM 6(H) if it lands on heads.)

OFFERED AND SMOKE_MARIJUANA: Question 13 from the FALL 2017 Math 1530 survey asked students "Have you ever been offered to smoke marijuana? (Yes, No)" and Question 14 of the survey asked students, "Have you ever smoked marijuana? (Yes, No)." We want to check if there is a relationship between **OFFERED AND SMOKE_MARIJUANA** among ETSU students. Assume the students who took the (FALL 2017 Math 1530) class survey are from an SRS of ETSU students.

a. Create an appropriate **graph** to display the relationship between **OFFERED** and **SMOKE_MARIJUANA**. Insert your graph here.



OR



b. Create an appropriate two-way table to summarize the data. Insert your table here. (**IN MINITAB: STAT → TABLES → CROSS TABULATION AND CHI-SQUARE. Make sure to select "Options" and click "No variables" under the "*Display missing values for*").**

Rows: OFFERED Columns: SMOKE_MARIJUANA

	No	Yes	All
No	220	2	222
Yes	314	481	795
All	534	483	1017

Cell Contents: Count

SUPPOSE WE SELECT ONE STUDENT AT RANDOM: (Calculate the following probabilities and show your work.)

c. What is the probability that this student has been offered marijuana *and* has smoked marijuana?

P = 481/1017 = 0.4729597 = 47.3%

d. What is the probability that this student has been offered marijuana *or* has smoked marijuana?

P = (795 + 483 – 481)/1017 = 0.7836775 = 78.37%

Name:
E number:
Math 1530 section number

e. What is the probability that this student has not smoked marijuana given that the student has been offered marijuana?

P = 314/795=0.3949686 = 39.5%

f. What is the probability that this student has been offered marijuana given that the student has not smoked marijuana?

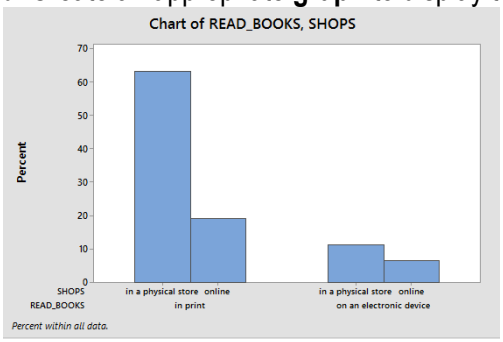
P = 314/534 =0.588015 = 58.8%

Name:
E number:
Math 1530 section number

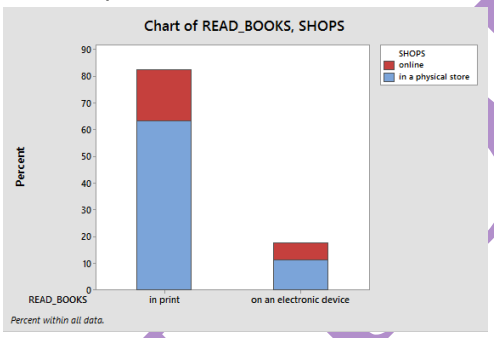
Problem 6 (H): Flip a fair coin. If it lands on heads do this problem (Omit this page/problem AND DO PROBLEM 6(T) if it lands on tails.)

READ_BOOKS AND SHOPS: Question 5 from the FALL 2017 Math 1530 survey asked students "Do you prefer to read books (excluding textbooks) in print or on an electronic device (such as a smart phone, tablet, computer, or e-reader)? (In print, On an electronic device)" and Question 7 from the FALL 2017 Math 1530 survey asked students "Do you prefer to shop for non-grocery items online or in a physical store? (Online, In a physical store)." We want to check if there is a relationship between **READ_BOOKS** and **SHOPS** among ETSU students. Assume the students who took the (FALL 2017 Math 1530) class survey are from an SRS of ETSU students.

a. Create an appropriate **graph** to display the relationship between **READ_BOOKS** and **SHOPS**. Insert your graph here.



OR



b. Create an appropriate two-way table to summarize the data. Insert your table here. (**IN MINITAB: STAT → TABLES → CROSS TABULATION AND CHI-SQUARE. Make sure to select "Options" and click "No variables" under the "*Display missing values for*").**

Rows: READ_BOOKS Columns: SHOPS

	in a physical store	online	All
in print	642	195	837
on an electronic device	114	66	180
All	756	261	1017

Cell Contents: Count

SUPPOSE WE SELECT ONE STUDENT AT RANDOM: (Calculate the following probabilities and show your work.)

c. What is the probability that this student prefers to read books on an electronic device *and* the student prefers to shop for non-grocery items online?

P = 66/1017 = 0.06489676 = 6.49%

d. What is the probability that this student prefers to read books on an electronic *or* the student prefers to shop for non-grocery items online?

P = (180 + 261 – 66)/1017 = 0.3687316 = 36.87%

Name:
E number:
Math 1530 section number

e. What is the probability that this student prefers to read books in print given that the student prefers to shop for non-grocery items in a physical store?

P = 642/756 = 0.8492063 = 84.92%

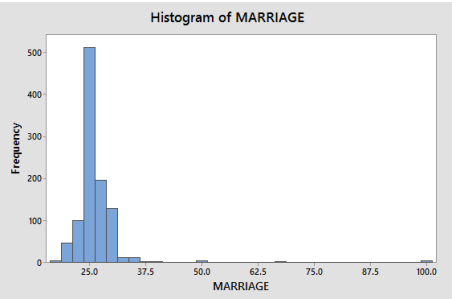
f. What is the probability that this student prefers to shop for non-grocery items in a physical store given that this student prefers to read books in print?

P = 642/837 = 0.7670251 = 76.7%

Name:
E number:
Math 1530 section number

Problem 7: In 2017, Business Insider did a story on using a mathematical theory to determine the best age to get married. This story reported the best age to get married was 26. (http://www.businessinsider.com/best-age-to-get-married-is-26-2017-2). Question 10 of the survey asked students, "At what age do you think is the best to get married in order to have a successful marriage?" On average, is the age that ETSU students believe is ideal to get married in order to have a successful marriage the same as reported in the story, 26?

a. Create a suitable graph to display the distribution of **MARRIAGE** reported by our sample of college students and insert it here.



Perform a test of significance to see if, on average, the age that ETSU students believe is ideal to get married in order to have a successful marriage the same as reported in the story, 26, using α = 0.05.

b. Write the correct null and alternative hypothesis for the test: **H_0: μ = 26 years versus H_a: μ ≠ 26 years**

c. Use Minitab to perform the appropriate test. Copy and paste the output for the test here.

```
Test of μ = 26 vs ≠ 26


Variable     N    Mean  StDev  SE Mean       95% CI          T      P
MARRIAGE  1017  26.224  5.193    0.163  (25.905, 26.544)  1.38  0.169
```

d. What is the name of your test statistic and what is its value? ***t* test statistic, t = 1.38**

e. What is the P-value for the test? **P-value = 0.169**

f. State your decision regarding the hypotheses being tested.

Because the P-value = 0.1689 is not small, we do not *reject* the null hypothesis.

g. State your conclusion. USE COMPLETE SENTENCES.

Based on the sample data provided, we did not reject the null hypothesis. We believe that ETSU students' ideal age to marry in order to have successful marriage, on average, is not different than the reported 26 years old.

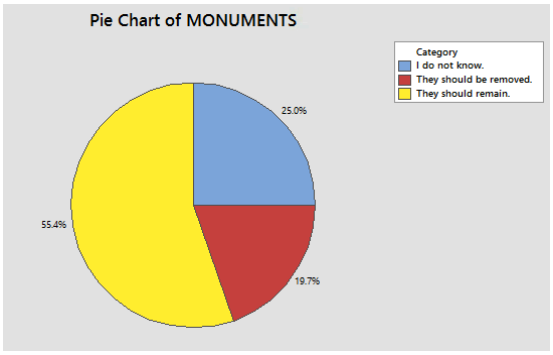
h. Is the P-value valid in this case? **See answer for part (i)**
i. What assumptions are you making in order to carry out this test?

Combined answer for part (h) and (i): ASSUMING the sample of ETSU college students from the Math 1530 survey can be treated as a random/representative sample of college students, the sample size, n = 1017, is large enough for the l-statistic to be valid.

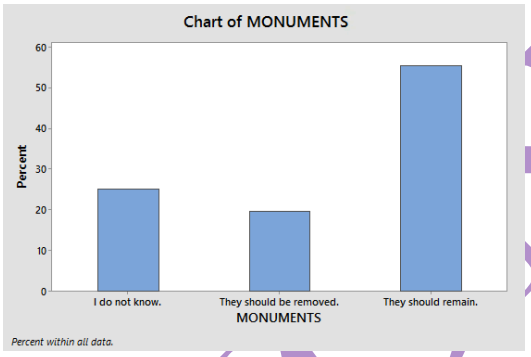
Name:
E number:
Math 1530 section number

Bonus Problem: Question 11 on the FALL 2017 Math 1530 asked, "Which comes closest to your opinion about Confederate monuments in public spaces: "They should be removed.", "They should remain.", or "I do not know."? (They should be removed, They should remain, I do not know)" An online survey conducted by the software company icitizen reported that 43% of adults said the monuments should remain (https://hyperallergic.com/397792/polls-americans-confederate-statues-removal/). Is the same true for the population of all U.S. college/university students?

a. Create an appropriate graph to display the distribution of **MONUMENTS** and insert it here.



OR



b. How many of the students surveyed said "They should remain?" **563**

MONUMENTS	Count
I do not know.	254
They should be removed.	200
They should remain.	563
N=	1017

c. What proportion of our sample said "They should remain?" **563/1017 = 0.553589 = 55.36%**

MONUMENTS	Percent
I do not know.	24.98
They should be removed.	19.67
They should remain.	55.36

d. Assume (for the purpose of this problem) that we may treat the FALL 2017 sample of Math-1530 students as a simple random sample drawn from the population of all U.S. college/university students. Use Minitab to calculate a 95% confidence interval for the proportion of students in the population who chose "They should remain" to the survey question (based on our sample data). Copy and paste the Minitab output here.

Test and CI for One Proportion

Sample	X	N	Sample p	95% CI
1	563	1017	0.553589	(0.522427, 0.584440)

Test and CI for One Proportion

Sample	X	N	Sample p	95% CI
1	563	1017	0.553589	(0.523036, 0.584142)

Using the normal approximation.

Name:
E number:
Math 1530 section number

e. Interpret the confidence interval you reported in part (d).

With 95% confidence, the true proportion of students who would chose "They should remain" to the survey question is between 52.3% and 58.4%.

f. What do you think? Do our results contradict the results obtained from survey by icitizen or do they appear to agree with it? EXPLAIN.

Because the value 43% is not in the calculated confidence interval, our sample suggests that the proportion of US college/university students that chose "They should remain" to the survey question is not 43%. Therefore, 43% is not within the 95% CI so our results did not appear to be in agreement with the survey conducted by icitizen.

Instructors: A student may choose to answer this using a hypothesis test.

Test and CI for One Proportion

```
Test of p = 0.43 vs p ≠ 0.43


Sample    X     N  Sample p         95% CI         Z-Value  P-Value
1       563  1017  0.553589  (0.523036, 0.584142)     7.96    0.000

Using the normal approximation.
```

Or

```
Test of p = 0.43 vs p ≠ 0.43


                                                    Exact
Sample    X     N  Sample p         95% CI        P-Value
1       563  1017  0.553589  (0.522427, 0.584440)   0.000
```

Because the P-value is zero, we reject the null hypothesis and thus, our results did not appear to be in agreement with the survey conducted by icitizen.