Name:
E-number:
Sum of E-number:
Math 1530 section number

MATH 1530 CAPSTONE TECHNOLOGY PROJECT SPRING 2019

Problem 1: Sampling In the survey data, the variable "**AGE**" is the age in years for each student.

a. Starting with the first observation, select every 20th observation until you have 10 observations. Assume the first observation was randomly chosen as the starting point. Type the 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 10 observations and report the value below.

N	1	2	3	4	5	6	7	8	9	10
AGE	**55**	**32**	**25**	**22**	**21**	**20**	**20**	**20**	**20**	**19**

b. The mean age of the above 10 students is **25.4**. (Type the value into the space provided.)

c. What type of sampling was used in part (a)? **systematic sampling**

d. 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 C18-C19 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	**413**	**357**	**147**	**394**	**465**	**13**	**468**	**399**	**27**	**6**
AGE	**18**	**19**	**20**	**18**	**18**	**34**	**18**	**18**	**29**	**44**

e. Calculate and report the mean age for your random sample of 10 students. The sample mean age is **23.6**.

ANSWERS WILL VARY.

f. What type of sampling was used in part (d)? **simple random sample**

g. 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 19.839. Discuss (two or more complete sentences) the **differences and similarities** between 19.839 and the answers you got in (b) and (e).

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

The 'systematic sample' mean found in (b.) *25.4* is above the 'population' mean 19.839. The 'SRS' mean found in part (d.) of *23.6* is *above* the population mean of 19.839. However, in the 'long run,' the distribution of sample means centered around the population mean in both types of sampling.

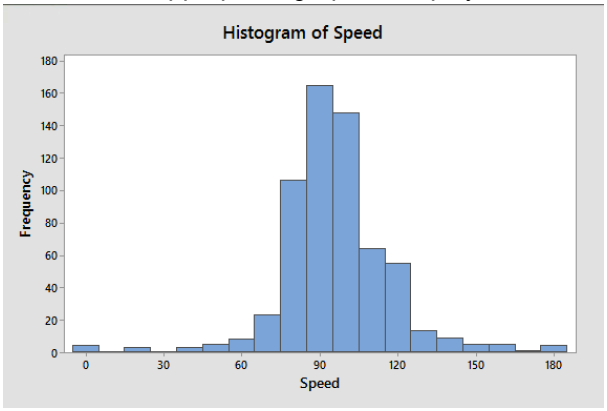
One may comment that the data was sorted from largest to smallest and thus would cause problems in the systematic sampling.

Name:
E-number:
Sum of E-number:
Math 1530 section number

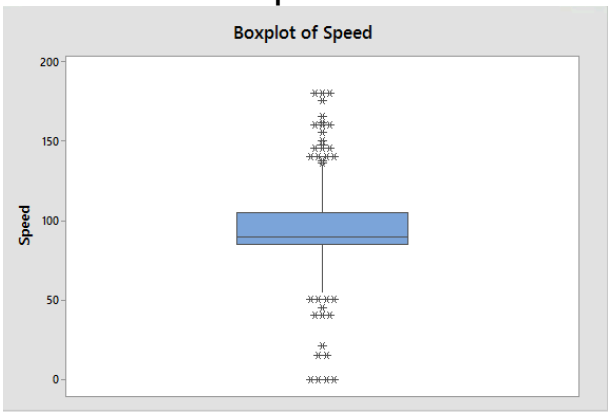
Problem 2(e): If your E number ends in an even number (0, 2, 4, 6, or 8) then do this question. (Omit this page/problem if your E# ends with an odd number.)

Question 8 of the SPRING 2019 survey asked students, "What is the fastest you have ever driven an automobile on the highway or interstate (in mph)?"

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



OR



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

Skewed left Uniform Skewed right Bimodal Symmetric

Instructors: A student might feel the distribution is slightly skewed right since the mean is larger than the median. Please give credit for both. The answers given in blue will be assuming the data is skewed right.

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

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum	IQR
Speed	621	0	94.717	0.853	21.268	0.000	85.000	90.000	105.000	180.000	20.000

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

d. Which statistic will you use to describe the center of the distribution? **Mean (or median);** Median

e. In one or two sentences, describe why this statistic was chosen. **Since the shape of the distribution is symmetric, the mean should be used to describe the center of the distribution. However, since the mean and the median are approximately the same, one could report either value.**

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.

Name:
E-number:
Sum of E-number:
Math 1530 section number

f. What is the value of that statistic? $\bar{x}$ = **94.717 (median = 90); median = 90**

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

h. What is (are) the value(s) of that (those) statistic(s)? **s = 21.268; Q1 = 85, Q3 = 105, and possible IQR = 20**

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.

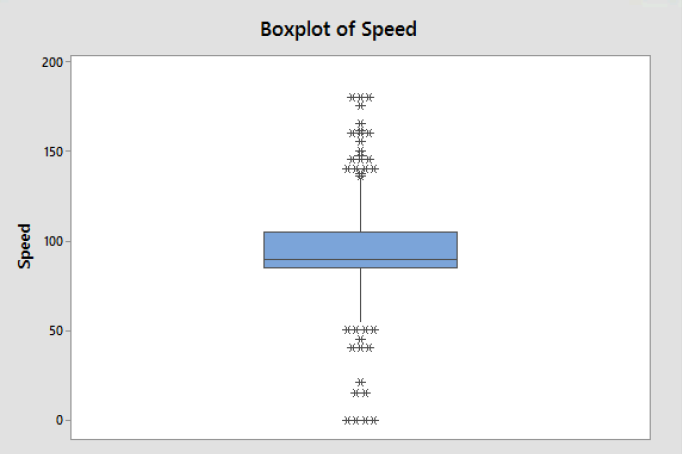
IQR = Q3 – Q1 = 105 –85 = 20, so 1.5 * IQR = 1.5 * 20 = 30.

Q1 – 1.5 * IQR = 85 – 30 = 55 and Q3 + 1.5 * IQR = 105+ 30 = 135

Any value of 'Speed' *below* 55 or *above* 135 would be considered outliers.

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

Minitab shows outliers with * on a boxplot:

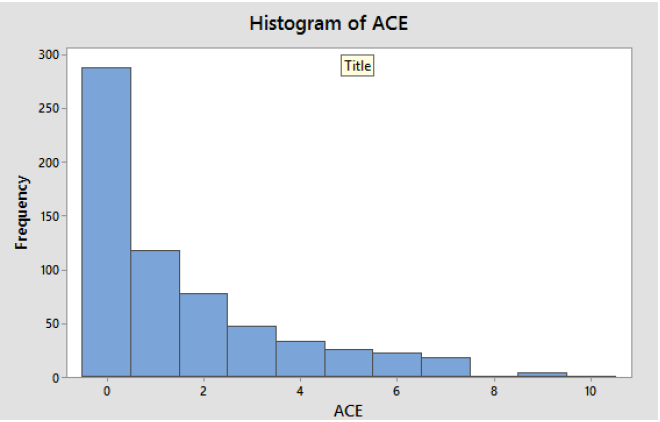


Name:
E-number:
Sum of E-number:
Math 1530 section number

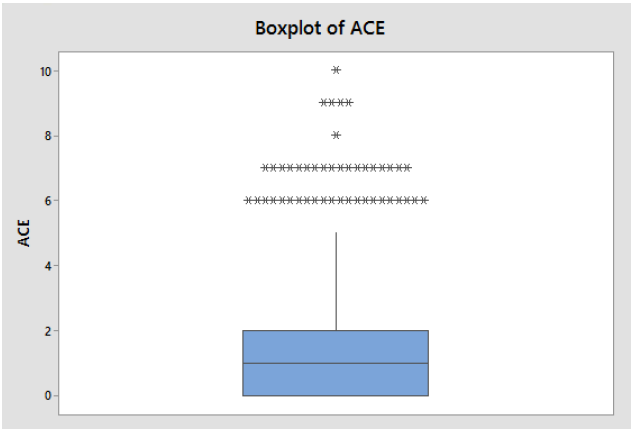
Problem 2(o): If your E number ends in an odd number (1, 3, 5, 7, or 9) then do this question. (Omit this page/problem if your E# ends with an even number.)

Question 4 of the SPRING 2019 survey asked students, "There are many health issues associated with Adverse Childhood Experiences (ACE). Go to the following link and take the short ACE quiz: https://www.npr.org/sections/health-shots/2015/03/02/387007941/take-the-ace-quiz-and-learn-what-it-does-and-doesnt-mean. What was your ACE quiz score?" Note, you do not need to take this quiz.

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



OR



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 **ACE**. Copy and paste all of the Minitab output here.

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum	IQR
ACE	635	0	1.5543	0.0812	2.0463	0.0000	0.0000	1.0000	2.0000	10.0000	2.0000

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

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? **1**

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 = 0, Q3 = 2, and possible IQR = 2**

Name:
E-number:
Sum of 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.

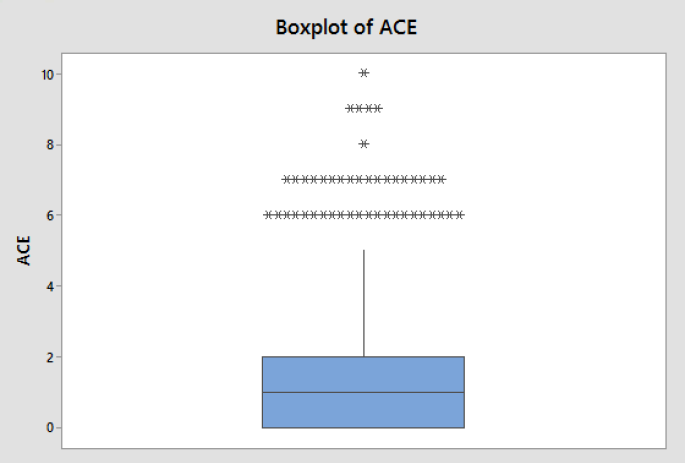
IQR = Q3 – Q1 = 2 – 0 = 2, so 1.5 * IQR = 1.5 * 2 = 3.

Q1 – 1.5 * IQR = 0 – 3 = –3 and Q3 + 1.5 * IQR = 2 + 3 = 5

Any value of 'ACE' *below* -3 or *above* 5 would be considered outliers.

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

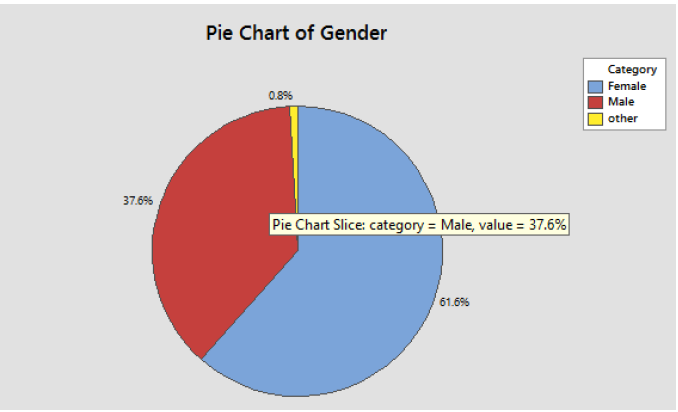
Minitab shows outliers with * on a boxplot:



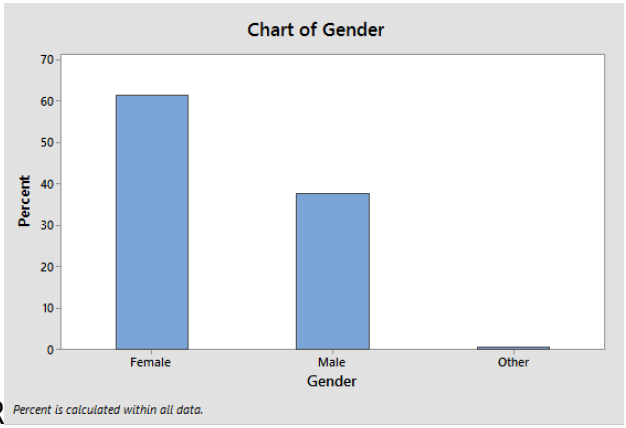
Name:
E-number:
Sum of E-number:
Math 1530 section number

Problem 3: Gender versus ACE. Question 2 of the survey asked students, "What gender do you identify with? (Female, Male, Other)" Question 4 of the SPRING 2019 survey asked students, "What was your ACE quiz score?"

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



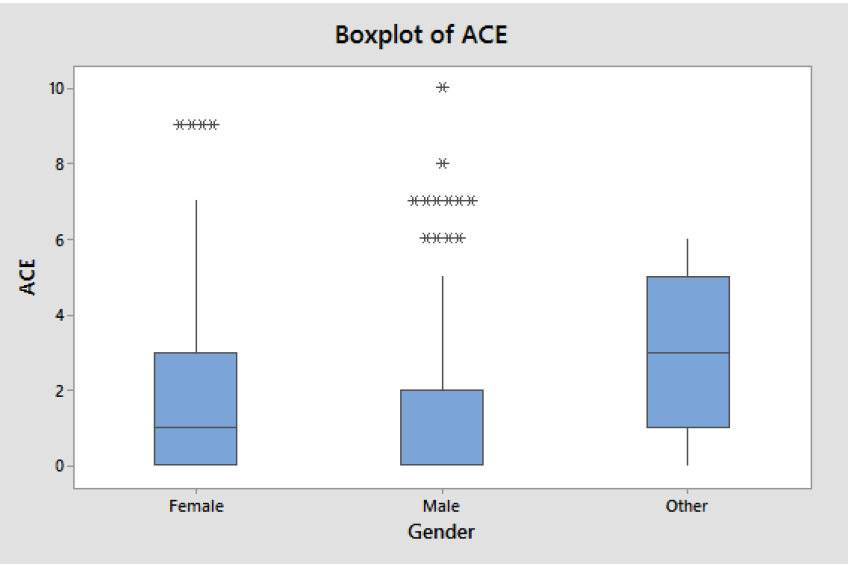
OR



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

Female Male Other

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



Name:
E-number:
Sum of E-number:
Math 1530 section number

Instructors: Please note the following summary was not required

Variable	Gender	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
ACE	Female	391	0	1.691	0.106	2.086	0.000	0.000	1.000	3.000	9.000
	Male	239	0	1.301	0.126	1.949	0.000	0.000	0.000	2.000	10.000
	Other	5	0	3.00	1.00	2.24	0.00	1.00	3.00	5.00	6.00

Variable	Gender	IQR
ACE	Female	3.000
	Male	2.000
	Other	4.00

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

d. Which gender group has the largest variability in terms of their ACE score? **Other but they all seem to be fairly the same**

e. Which gender group has the highest median ACE score? **Other**

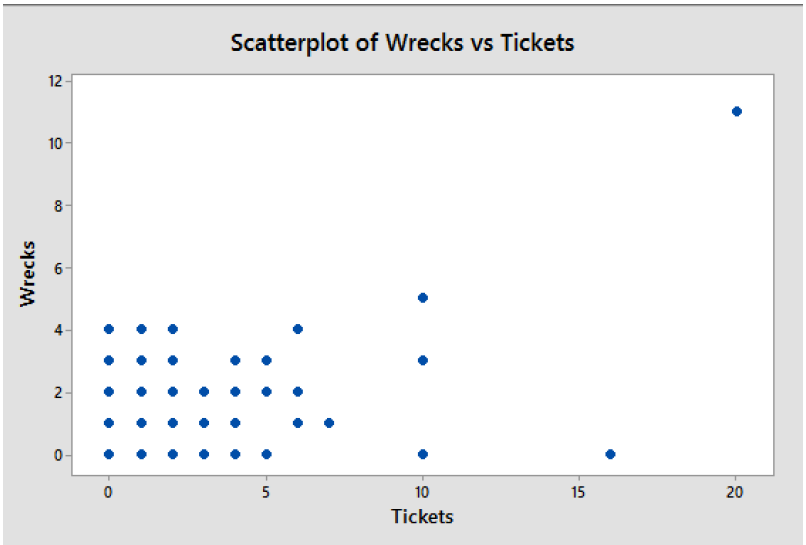
f. Which gender group has the largest IQR in terms of their ACE score? **Other**

g. Discuss (two or more complete sentences) what this plot tells you. **Both females and males have outliers unlike those whose chose "other" for gender. Those who identify as "other" tend to have a higher ACE score than those who identify as males and females with males appearing to have the lowest ACE score.**

Name:
E-number:
Sum of E-number:
Math 1530 section number

Problem 4: REGRESSION On the SPRING 2019 Math 1530 survey, question 9 asked students, "How many wrecks have you been in where you were the driver?" and question 10 asked students, "How many speeding tickets have you received?" We are interested in seeing whether we can use the number of speeding tickets to predict the number of wrecks a student has been in where they were the driver.

a. Create an appropriate graph to display the relationship between **Tickets** and **Wrecks**. 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 tickets increases, the number of wrecks increases.

Instructors: Some students may feel there is no association and the few outliers is causing to association to appear positive. Please give credit for this answer as well.

c. Describe the *form* of the relationship between **Tickets** and **Wrecks.**

Moderately linear

Instructors: For those that said no association, then the form should be no association (random scatter)

d. Report the value of the correlation between this pair of variables? ***r* = 0.411**

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 moderate to weak.

Name:
E-number:
Sum of E-number:
Math 1530 section number

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

Regression Equation

Wrecks = 0.4075 + 0.2466 Tickets

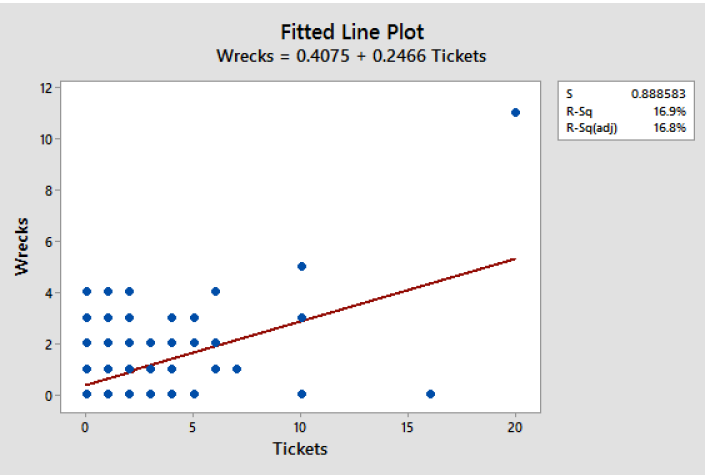
g. Interpret the value of the slope in the least squares regression equation you found in part (f).

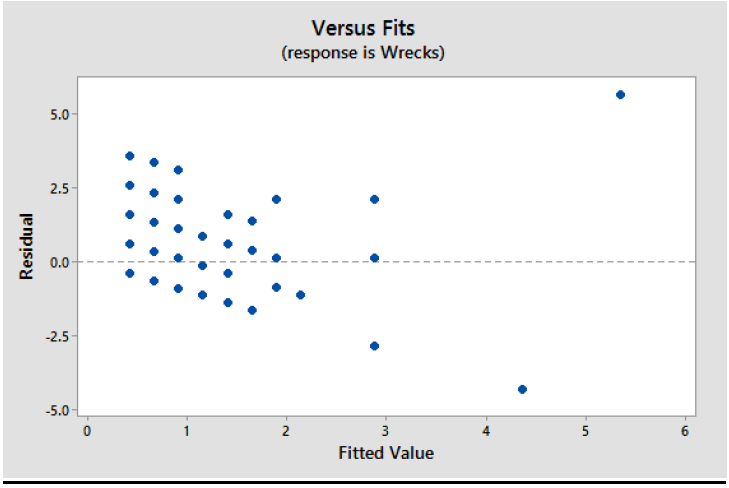
As the number of tickets increases by one ticket, the estimated number of wrecks increases by 0.2466 wrecks.

h. Use the regression equation in part (f) to predict the number of wrecks for a student that has had 3 speeding tickets. (Show your math.)

Predicted number of wrecks = 0.4075+0.2466*3 = 1.1473 wrecks

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





The association is very weak and can be seen clearly in the fitted line plot. There are a few points that are scattered far away from the regression line. The squared correlation (R^2) indicates that 16.9% of the variation we observed in wrecks is explained by the linear relationship with tickets.

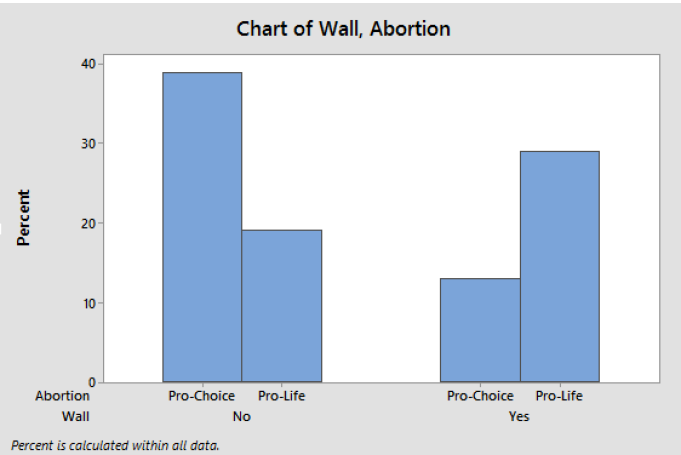
Note: Another scatterplot that is useful to see whether the model makes sense is the residual plot. This helps in determining the appropriateness of the regression model. Recall that the residuals are Residual = Observed Data – Predicted Data. The residual plot shouldn't have any interesting features, 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 appears to have possibly some cause to worry.

Name:
E-number:
Sum of E-number:
Math 1530 section number

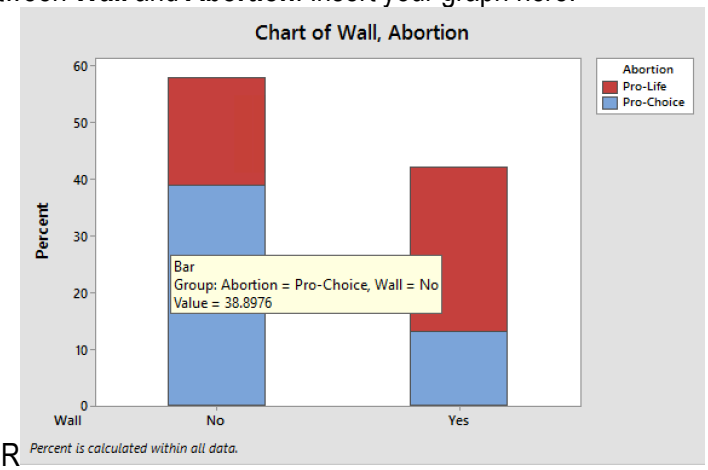
<u>Problem 5 (T):</u> If the sum of the digits in your E number is an even number then do this question. (Omit this page/problem if the sum of the digits in your E number is an odd number.)

Wall and Abortion: Question 14 from the SPRING 2019 Math 1530 survey asked students, "Do you support the idea of building a border wall between the U.S. and Mexico? (Yes, No)" and Question 15 of the survey asked students, "Would you identify yourself as pro-choice or pro-life on abortion? (Pro-choice, Pro-life)" We want to check if there is a relationship between Wall and Abortion among ETSU students. Assume the students who took the (SPRING 2019 Math 1530) class survey are from an SRS of ETSU students.

a. Create an appropriate **graph** to display the relationship between **Wall** and **Abortion**. 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: Wall Columns: Abortion

	Pro-Choice	Pro-Life	All
No	247	121	368
Yes	83	184	267
All	330	305	635

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

c. What is the probability that this student supports the idea of building a border wall between the U.S. and Mexico *and* is pro-life?

P = 184/635 = 0.29 = 29%

d. What is the probability that this student supports the idea of building a border wall between the U.S. and Mexico *or* is pro-life?

P = (305 + 267 – 184)/635 = 0.611 = 61.1%

Name:
E-number:
Sum of E-number:
Math 1530 section number

e. What is the probability that this student does not support the idea of building a border wall between the U.S. and Mexico *given* the student is pro-life?

P = 121/305 = 0.3967 = 39.67%

f. BONUS: Carry out a test for the hypothesis that there is no relationship between **Wall** and **Abortion**. Use a significance level of α = 0.05.

i. State the null and alternative hypotheses.

H_0: There is no relationship between Wall and Abortion
H_a: There is a relationship between Wall and Abortion

ii. Perform the test and include any output from Minitab here.

Rows: Wall Columns: Abortion

	Pro-Choice	Pro-Life	All
No	247	121	368
	191.2	176.8	
Yes	83	184	267
	138.8	128.2	
All	330	305	635

Cell Contents
Count
Expected count

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	80.488	1	0.000

iii. Which test statistic are you using and what is its value? **A chi-square test statistic and its value is 80.488**

iv. State your decision and conclusion for the test. **Based on the chi-square test results, the p-value is 0.000. Therefore, at a 5% level of significance, we reject the null hypothesis and conclude there is a significant relationship between wall and abortion.**

v. Examine the data. Are the conditions for inference in part (ii) violated? Explain.

Conditions for inference about a chi-square test:
*** No more than 20% of the expected counts are less than 5 and all individual expected counts are greater than 0.** All the expected counts are greater than 5.

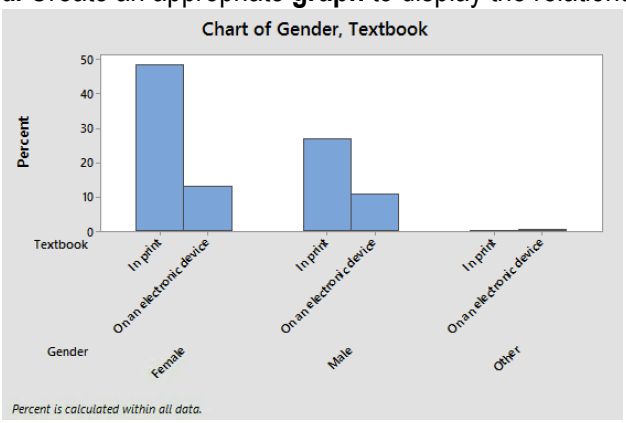
*** The data is a random sample from the population.** Here the problem states to assume the students who took the class survey are from an SRS of ETSU Students.

Name:
E-number:
Sum of E-number:
Math 1530 section number

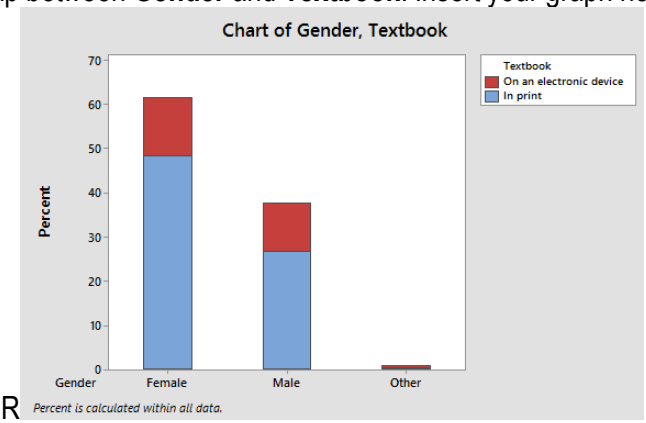
<u>Problem 5 (H):</u> If the sum of the digits in your E number is an odd number then do this question. (Omit this page/problem if the sum of the digits in your E number is an even number.)

Gender and Textbook: Question 2 from the SPRING 2019 Math 1530 survey asked students, "What gender do you identify with? (Female, Male, Other)" and Question 5 of the survey asked students, "Do you prefer to read textbooks in print or on an electronic device (such as a smart phone, tablet, computer, or e-reader)? (In print, On an electronic device)" We want to check if there is a relationship between **Gender and Textbook** among ETSU students. Assume the students who took the (SPRING 2019 Math 1530) class survey are from an SRS of ETSU students.

a. Create an appropriate **graph** to display the relationship between **Gender** and **Textbook**. 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: Gender Columns: Textbook

	In print	On an electronic device	All
Female	308	83	391
Male	170	69	239
Other	2	3	5
All	480	155	635

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

c. What is the probability that this student identifies as a female *and* says they prefer to read a textbook in print?

P = 308/635= 0.485 = 48.5%

d. What is the probability that this student identifies as a female *or* says they prefer to read a textbook in print?

P = (391 + 480 – 308)/635 = 0.8866= 88.66%

Name:
E-number:
Sum of E-number:
Math 1530 section number

e. What is the probability that this student identifies as other *given* the student prefers to read a textbook on an electronic device?

P = 3/155 = 0.0194= 1.94%

f. BONUS: Carry out a test for the hypothesis that there is no relationship between **Gender** and **Textbook**. Use a significance level of α = 0.05.

i. State the null and alternative hypotheses.

H_0: There is no relationship between Gender and Textbook
H_a: There is a relationship between Gender and Textbook

ii. Perform the test and include any output from Minitab here.

Rows: Gender Columns: Textbook

	In print	On an electronic device	All
Female	308	83	391
	295.56	95.44	
Male	170	69	239
	180.66	58.34	
Other	2	3	5
	3.78	1.22	
All	480	155	635

Cell Contents
Count
Expected count

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	8.155	2	0.017

iii. Which test statistic are you using and what is its value? **A chi-square test statistic and its value is 8.155**

iv. State your decision and conclusion for the test. **Based on the chi-square test results, the p-value is 0.017. Therefore, at a 5% level of significance, we reject the null hypothesis and conclude there is a significant relationship between gender and textbook.**

Name:
E-number:
Sum of E-number:
Math 1530 section number

v. Examine the data. Are the conditions for inference in part (ii) violated? Explain.

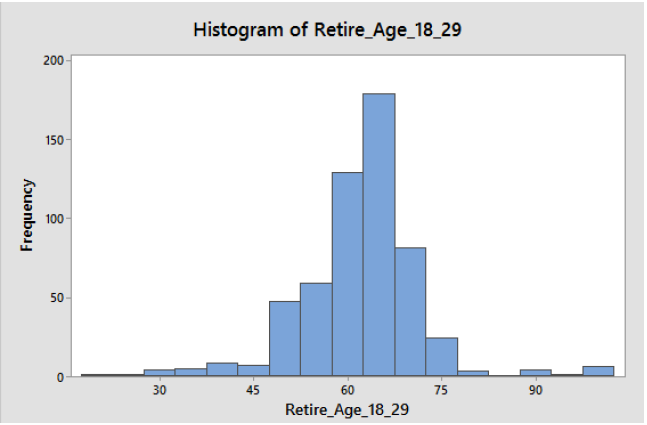
Conditions for inference about a chi-square test:
*** No more than 20% of the expected counts are less than 5 and all individual expected counts are greater than 0.** **We see this assumption is not satisfied because the expected counts for "other and in print" and "other and on an electronic device" are less than five.**

*** The data is a random sample from the population.** **Here the problem states to assume the students who took the class survey are from an SRS of ETSU Students.**

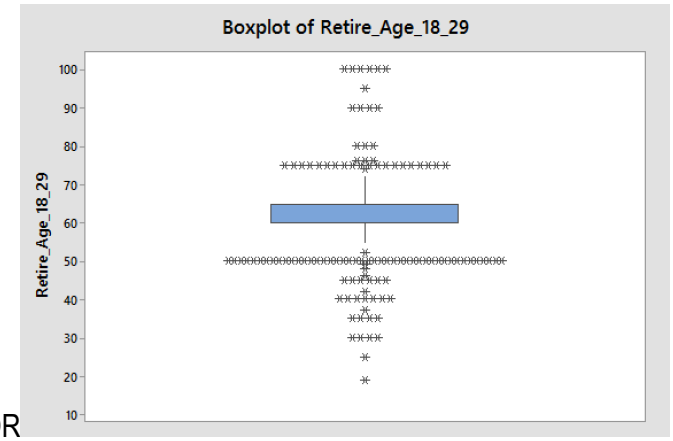
Name:
E-number:
Sum of E-number:
Math 1530 section number

Problem 6: In 2018, Gallup conducted a survey on age of retirement (https://news.gallup.com/poll/234302/snapshot-americans-project-average-retirement-age.aspx). One question asked was the same as question 13 from the SPRING 2019 MATH 1530 survey. Question 13 of the survey asked students, "At what age do you expect to retire?" Gallup asked this question for three different age groups. It was found that average age one expects to retire (for age group 18-29) was 63. Since the majority of students taking MATH 1530 fall in the age group 18-29, the variable **Retire_Age_18_29** was created. This variable contains the values of what age a student expects to retire for those students aged 18 to 29. On average, is the age that ETSU students (aged 18-29) expect to retire the same as reported in the survey, 63?

a. Create a suitable graph to display the distribution of **Retire_Age_18_29** reported by our sample of college students and insert it here.



OR



Using α = 0.05, perform a test of significance to see if, on average, the age that ETSU students (aged 18-29) expect to retire the same as reported in the survey, 63.

b. Write the correct null and alternative hypotheses for the test:

H_0: $\mu = 63$

H_a: $\mu \neq 63$

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

Descriptive Statistics

N	Mean	StDev	SE Mean	95% CI for μ
559	62.079	9.595	0.406	(61.282, 62.876)

μ: mean of Retire_Age_18_29

Test

Null hypothesis H_0: μ = 63
Alternative hypothesis H_1: μ ≠ 63

T-Value	P-Value
-2.27	0.024

Name:
E-number:
Sum of E-number:
Math 1530 section number

d. What is the name of your test statistic and what is its value? ***t* test statistic, t = -2.27**

e. What is the P-value for the test? **P = 0.024**

f. State your decision regarding the hypotheses being tested. **Because the P-value is small (less than $\alpha = 0.05$), we *reject* the null hypothesis.**

g. State your conclusion. USE COMPLETE SENTENCES.

Based on the sample data provided, we reject the null hypothesis. We believe, on average, the age that ETSU students (aged 18-29) expect to retire is not the same as reported in the survey, 63.

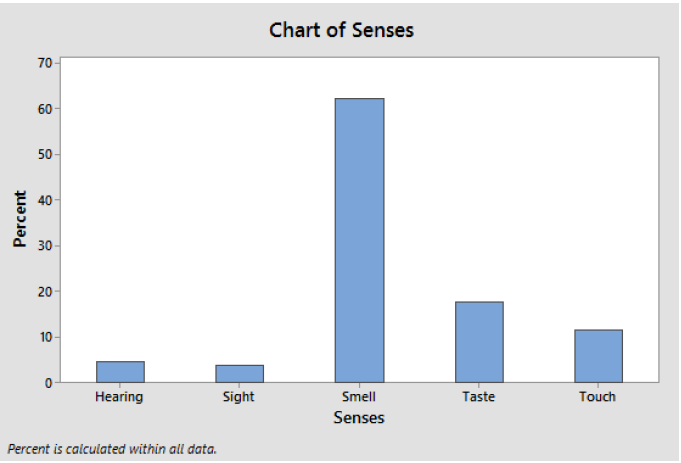
h. What assumptions are you making in order to carry out this test? Is the P-value valid in this case?

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 = 635, is large enough for the t-statistic to be valid. However, one may notice that there is an extreme number of outliers in this data and thus questioning the validity of the t-procedure.

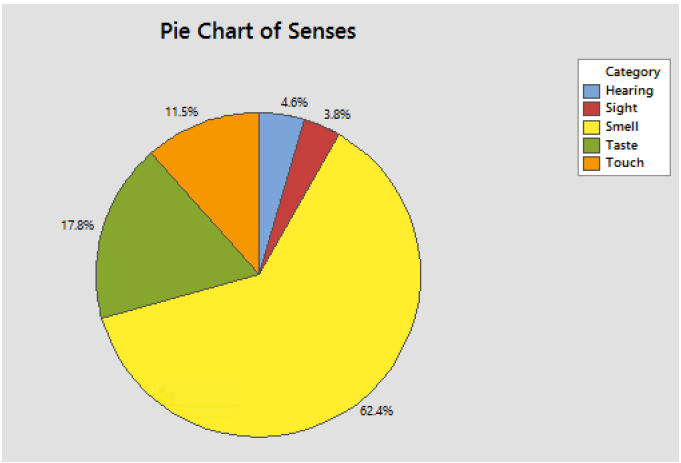
Name:
E-number:
Sum of E-number:
Math 1530 section number

Problem 7 Question 12 from the SPRING 2019 Math 1530 survey asked students "If you had to lose one of your five senses, which would you choose? (Hearing, Sight, Touch, Smell, Taste)." Many surveys have been conducted on which of the five senses would you want to lose if you had to lose one. One studied stated that 57% of people said smell would be the sense they would choose to lose (https://www.quibblo.com/poll/1JoM1SK/If-you-HAD-to-give-up-a-sense-which-would-it-be). Is the same true for the population of all U.S. college/university students?

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



OR



b. How many of the students surveyed said "Smell" was the sense they would choose to lose? **396**

Senses	Count
Hearing	29
Sight	24
Smell	396
Taste	113
Touch	73
N=	635

c. What proportion of our sample said "Smell" was the sense they would choose to lose? **62.36**

Senses	Count	Percent
Hearing	29	4.57
Sight	24	3.78
Smell	396	62.36
Taste	113	17.80
Touch	73	11.50
N=	635	

Name:
E-number:
Sum of E-number:
Math 1530 section number

d. Assume (for the purpose of this problem) that we may treat the SPRING 2019 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 "Smell" to the survey question (based on our sample data). Copy and paste the Minitab output here.

Normal Approximation:

Descriptive Statistics

N	Event	Sample p	95% CI for p
635	396	0.623622	(0.585940, 0.661304)

Exact:

Descriptive Statistics

N	Event	Sample p	95% CI for p
635	396	0.623622	(0.584648, 0.661441)

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

With 95% confidence, the true proportion of students who would chose "Smell" to the survey question is between 58.6% and 66.1%.

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

Because the value 57% is not within the calculated confidence interval, our sample suggests that the proportion of US college/university students that chose "Smell" to the survey question is not 57%. Therefore, 57% is not within the 95% CI so our results did not appear to be in agreement with the survey conducted.

Instructors: A student may choose to answer this using a hypothesis test. The P-value is less than 0.05, and thus we reject the null hypothesis and thus, our results did not appear to be in agreement with the survey conducted.

Test

Null hypothesis	H_0: p = 0.57
Alternative hypothesis	H_1: p ≠ 0.57

Z-Value	P-Value
2.73	0.006