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

MATH 1530 CAPSTONE TECHNOLOGY PROJECT FALL 2018

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. 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	**18**	**18**	**19**	**18**	**19**	**18**	**18**	**18**	**18**	**19**

b. The mean age of the above 10 students is **18.3**. (Type the value into the space provided.)

c. What type of sampling was used in part (a)? **convenience sampling unless one assumes a random starting place 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 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	**671**	**867**	**441**	**7**	**616**	**279**	**367**	**54**	**874**	**913**
AGE	**19**	**52**	**19**	**18**	**19**	**18**	**17**	**26**	**29**	**20**

e. Calculate and report the mean age for your random sample of 10 students. The sample mean age is **23.7**.

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 19.423. Discuss (two or more complete sentences) the **differences and similarities** between 19.423 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 'convenience sample' mean found in (b.) *18.3* underestimates the 'population' mean 19.423, 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 *23.7* is *above* the population mean of 19.423. 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.

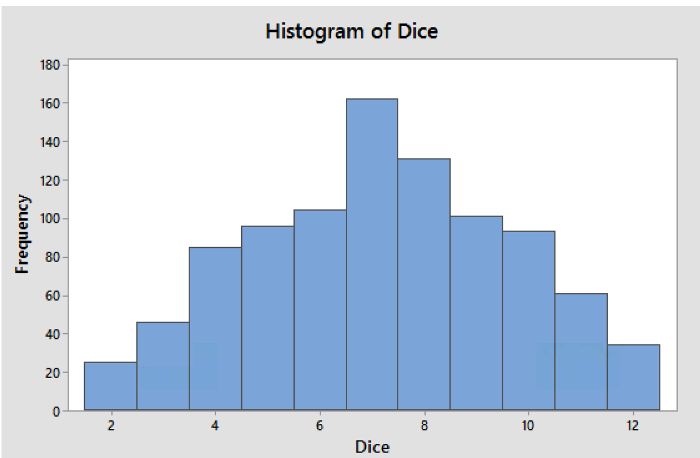
The 'systematic sample' mean found in (b.) *18.3* underestimates the 'population' mean 19.423. The 'SRS' mean found in part (d.) of *23.7* is *above* the population mean of 19.423. However, in the 'long run,' the distribution of sample means centered around the population mean in both types of sampling.

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.)

1. Question 6 of the FALL 2018 survey asked students, *"The following link will allow you to roll two virtual six-sided dice. Make sure the Roll is set to 2 virtual dice. Click Roll Dice. Here is the link: https://www.random.org/dice/. **What is the sum of the dots on the dice?**"* Note: You do NOT need to go to the website and roll the dice.

a. Create an appropriate graph to display the *distribution* of the variable called **Dice** 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 **Dice**. Copy and paste all of the Minitab output here.

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum	IQR
Dice	938	0	7.1770	0.0811	2.4842	2.0000	5.0000	7.0000	9.0000	12.0000	4.0000

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

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

f. What is the value of that statistic? **$\bar{x}$ = 7.177 (median = 7)**

g. Which statistic(s) will you use to describe the spread of the distribution? **Standard deviation**

h. What is (are) the value(s) of that (those) statistic(s)? **s = 2.4842**

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.

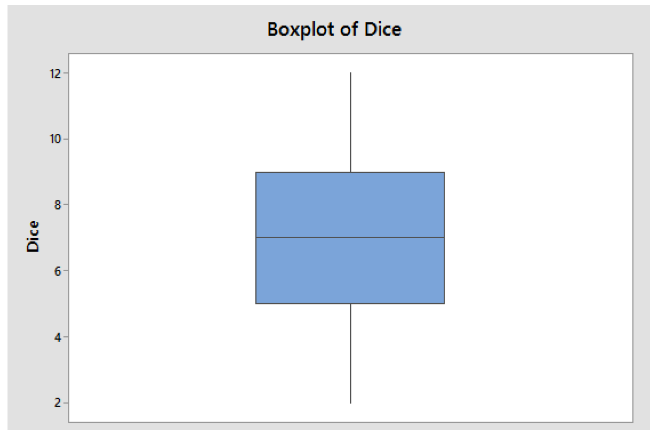
IQR rule says that any value *below* Q1 – 1.5*IQR or *above* Q3 + 1.5*IQR are outliers.

IQR = Q3 – Q1 = 9 – 5 = 4, so 1.5 * IQR = 1.5 * 5 = 6.

Q1 – 1.5 * IQR = 5 – 6 = -1 and Q3 + 1.5 * IQR = 9+ 6 = 15

Any value of 'DICE' *below* -1 or *above* 15 would be considered outliers.

No, there does not appear to be any outliers in the distribution of 'DICE'.

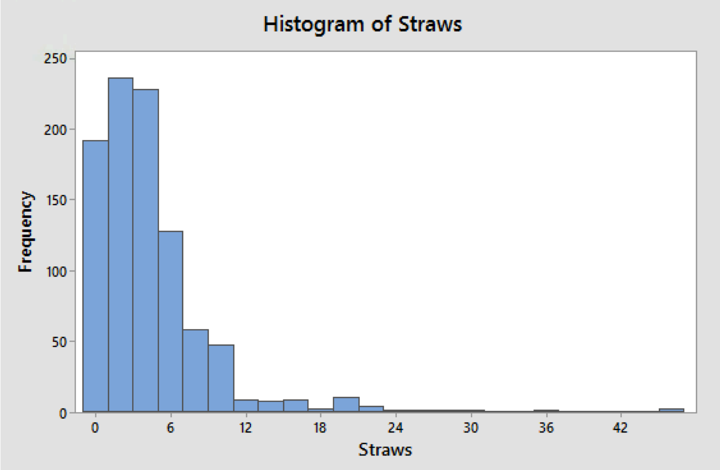


Name:
E-number:
Sum of E-number:
Math 1530 section number

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.)

2. Question 8 of the FALL 2018 survey asked students, "Approximately how many disposable plastic straws do you use during a typical week?"

a. Create an appropriate graph to display the *distribution* of the variable called **Straws** 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 **Straws**. Copy and paste all of the Minitab output here.

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum	IQR
Straws	938	0	3.857	0.152	4.645	0.000	1.000	3.000	5.000	47.000	4.000

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

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? **3**

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 = 1, Q3 = 5, and possible IQR =4**

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.

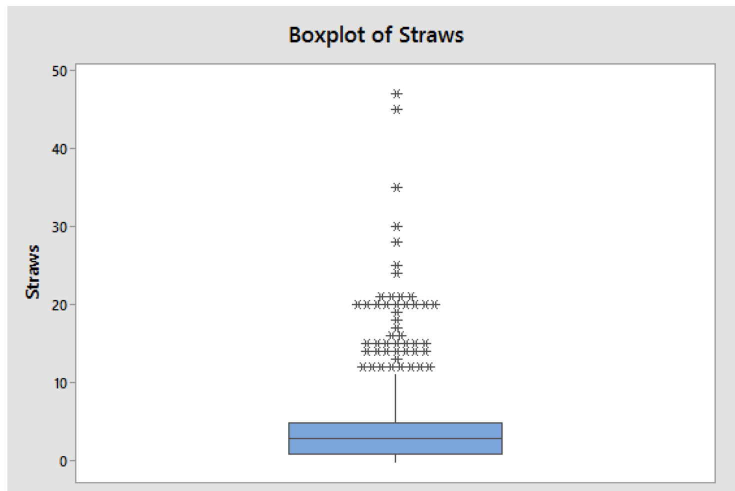
IQR = Q3 – Q1 = 5 – 1 = 4, so 1.5 * IQR = 1.5 * 4 = 6.

Q1 – 1.5 * IQR = 1 – 4 = –3 and Q3 + 1.5 * IQR = 5 + 6 = 11

Any value of 'STRAWS' *below* -3 or *above* 11 would be considered outliers.

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

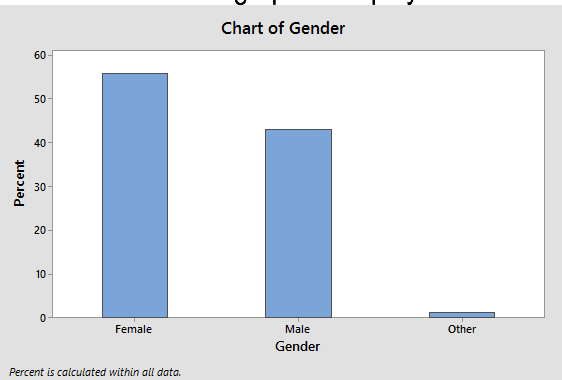
Minitab shows outliers with * on a boxplot:



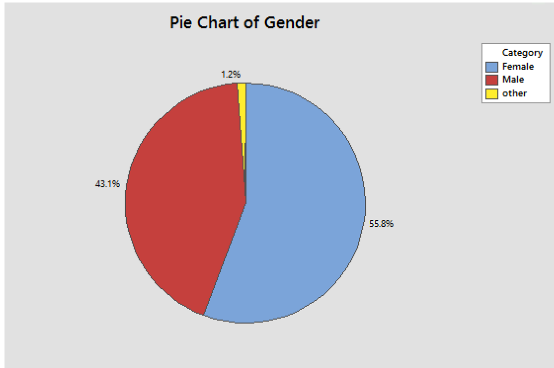
Name:
E-number:
Sum of E-number:
Math 1530 section number

Problem 3: Gender versus Foot. Question 2 of the survey asked students, "What gender do you identify with? (Female, Male, Other) " Question 3 of the survey asked students, "What is the length of your foot (in inches)?"

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



OR

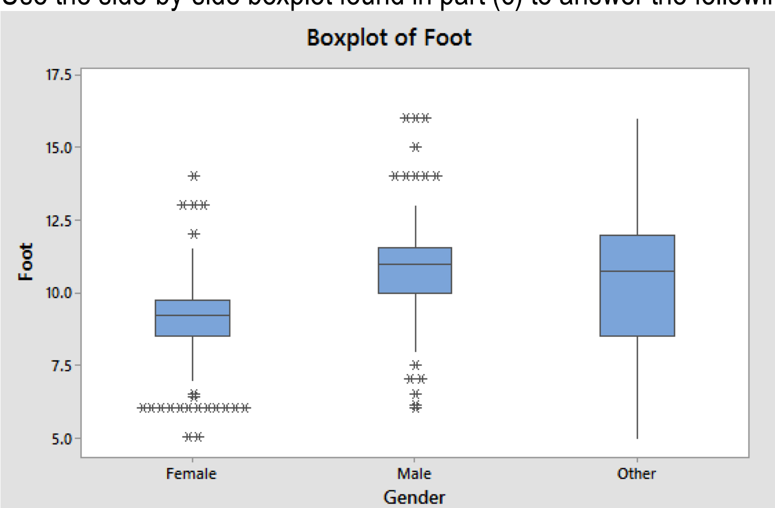


b. What is the mode of this distribution? Underline (or highlight) one option.

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 **Foot** for the "Graph variables" slot and **Gender** 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 gender group has the largest variability in terms of their foot length? **Other**

e. Which gender group has the highest median foot length? **Male**

f. Which gender group has the largest IQR in terms of their foot length? **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. Males seem to have a larger foot size then females.**

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

Problem 4: REGRESSION We are interested in creating a regression model between two numeric variables asked from the survey.

a. Which two quantitative variables have the strongest correlation?

	Age	Foot	Forearm	Water	Dice	Shooter
Foot	0.057					
Forearm	0.025	0.585				
Water	0.038	0.103	0.055			
Dice	0.009	0.061	0.136	-0.014		
Shooter	-0.075	-0.212	-0.171	-0.022	-0.031	
Straws	0.095	-0.028	0.004	-0.120	0.034	0.027

Cell Contents
Pearson correlation

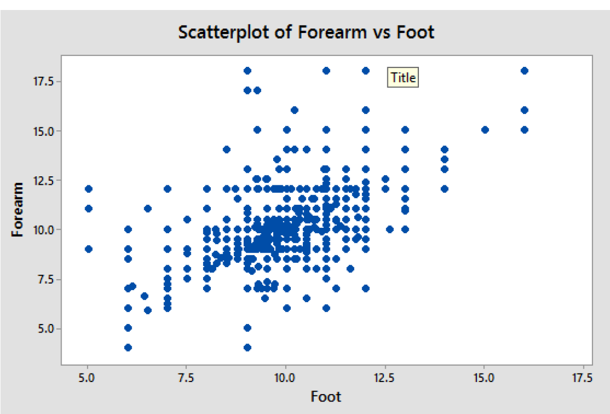
Based on the above correlation matrix, FOOT and FOREARM have the strongest correlation

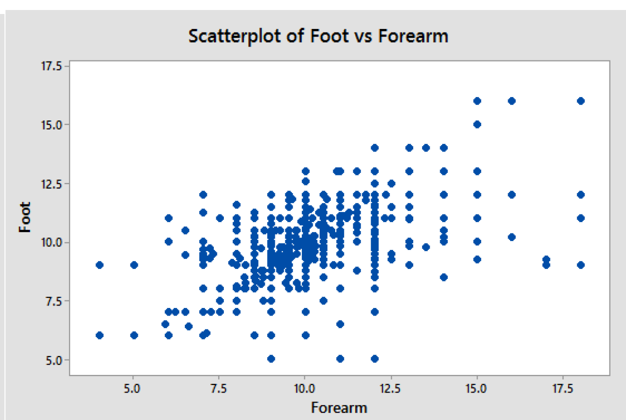
b. Report the value of the correlation between these two variables? ***r* = 0.585**

c. Create an appropriate graph to display the relationship between these two variables.

The first graph assumes FOOT is the explanatory variable and FOREARM is the response variable. The second graph assumes FOREARM is the explanatory variable and FOOT is the response variable.

Instructors: I feel a student could justify either situation in which variable they let be the response and explanatory.





d. 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.

(Y = FOREARM, X = FOOT): Positive association; As foot size increases, forearm length also increases.

(Y = FOOT, X = FOREARM): Positive association; As forearm length increases, foot size also increases.

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

e. Describe the *form* of the relationship between these two variables.

(Y = FOREARM, X = FOOT): Linear

(Y = FOOT, X = FOREARM): Linear

f. 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.

g. Using Minitab, obtain the equation for the least squares regression line. Use your best judgement to determine which is the response variable and which is the predictor variable.

The regression equation is
Forearm = 2.654 + 0.7479 Foot

OR

The regression equation is
Foot = 5.251 + 0.4579 Forearm

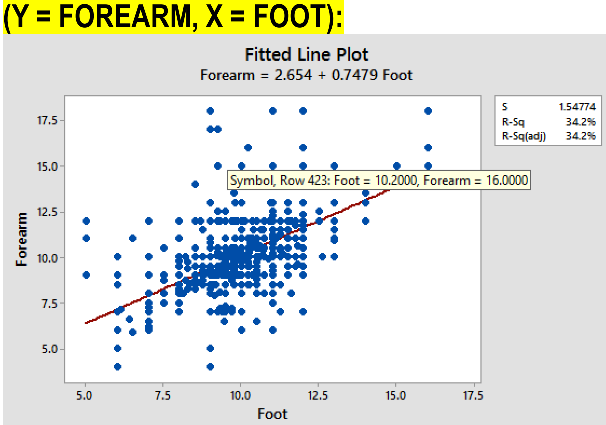
h. Interpret the value of the slope in the least squares regression equation you found in part (f).

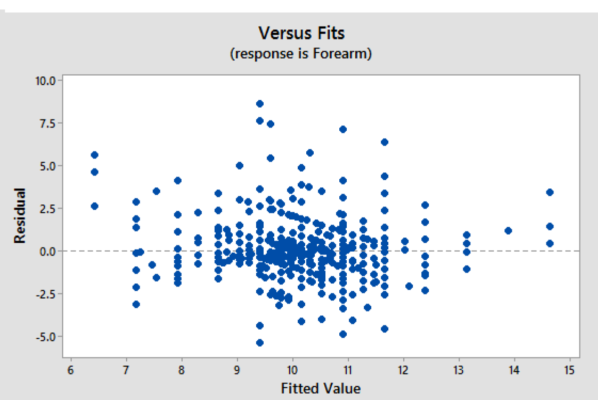
(Y = FOREARM, X = FOOT): As foot size increases by one inch, the estimated forearm length increases by 0.7479 inches.

(Y = FOOT, X = FOREARM): As forearm length increases by one inch, the estimated foot size increases by 0.4579 inches.

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

(Y = FOREARM, X = FOOT):



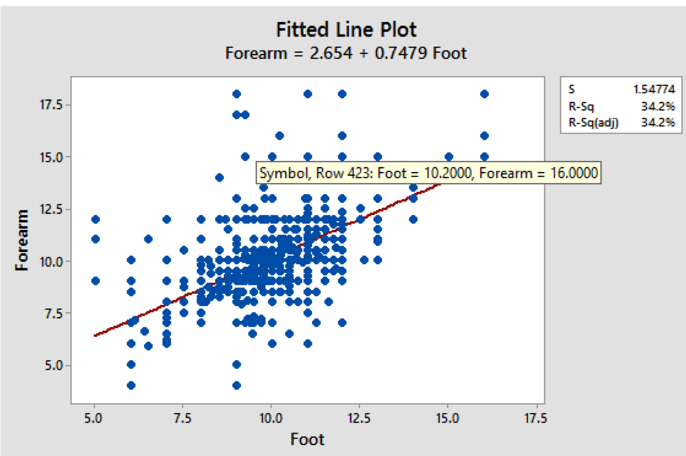


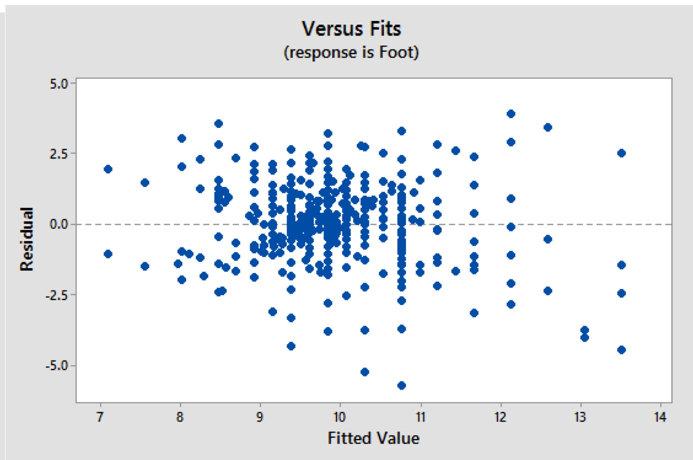
The association is somewhat weak and can be seen clearly in the fitted line plot. There are a few points that are scattered far away from the regression line. The squared correlation (R^2) indicates that 34.2% of the variation we observed in forearm length is explained by the linear relationship with foot size.

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

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 does not appear to have any cause to worry.

(Y = FOOT, X = FOREARM):





The association is somewhat 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 34.2% of the variation we observed in foot size is explained by the linear relationship with forearm length.

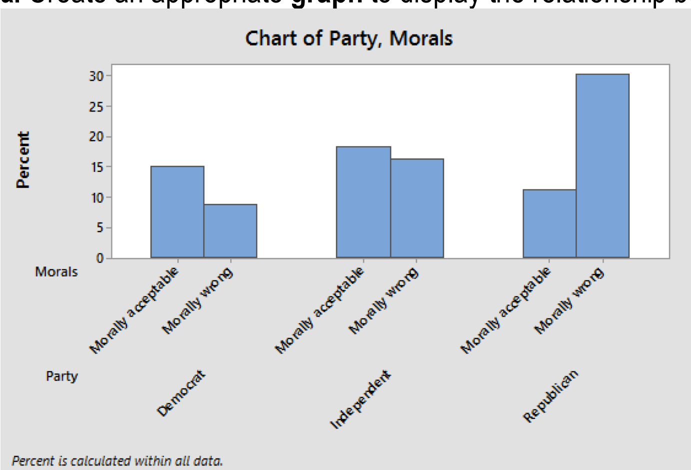
Note: Another scatterplot that is useful to see whether the model makes sense is the residual plot. This helps in determining the appropriateness of the regression model. Recall that the residuals are Residual = Observed Data – Predicted Data. The residual plot shouldn't have any interesting features,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

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

Problem 5 (T): **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.)**

Party and Morals: Question 10 from the FALL 2018 Math 1530 survey asked students, "In politics, as of today, do you consider yourself a Republican, a Democrat, or an Independent? (Republican, Democrat, Independent)" and Question 13 of the survey asked students, "Do you personally believe that in general pornography is morally acceptable or morally wrong? (Morally acceptable, Morally wrong)" We want to check if there is a relationship between Party and Morals among ETSU students. Assume the students who took the (FALL 2018 Math 1530) class survey are from an SRS of ETSU students.

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



A stacked bar graph could also have been constructed.

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: Party   Columns: Morals

	Morally acceptable	Morally wrong	All
Democrat	142	83	225
Independent	172	152	324
Republican	106	283	389
All	420	518	938

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 is a Republican *and* feels pornography is morally acceptable?

P = 106/938 = 0.113 = 11.3%

d. What is the probability that this student is a Republican *or* feels pornography is morally acceptable?

P = (389 + 429 – 106)/938 = 0.7591 = 75.91%

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

e. What is the probability that this student is a Democrat *given* that the student feels pornography is morally acceptable?

P = 142/420 = 0.3381 = 33.81%

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

i. State the null and alternative hypotheses.

H_0: There is no relationship between PARTY and MORALS
H_a: There is a relationship between PARTY and MORALS

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

	Morally acceptable	Morally wrong	All
Democrat	142 100.7	83 124.3	225
Independent	172 145.1	152 178.9	324
Republican	106 174.2	283 214.8	389
All	420	518	938

Cell Contents
Count
Expected count

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	87.964	2	0.000

iii. Which test statistic are you using and what is its value? **A chi-square test statistic and its value is 87.964**

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 party and morals.**

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.**

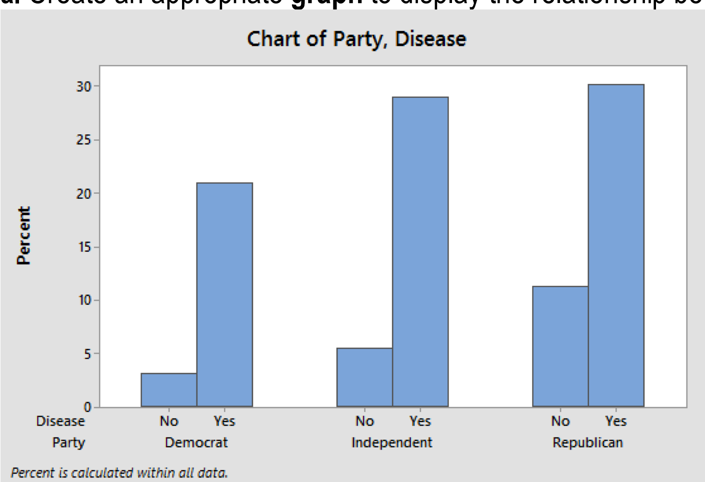
*** The data is a random sample from the population. Here the problem states to assume the students who took the class survey are from an SRS of ETSU Students.**

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.)

Party and Disease: Question 10 from the FALL 2018 Math 1530 survey asked students, "In politics, as of today, do you consider yourself a Republican, a Democrat, or an Independent? (Republican, Democrat, Independent)" and Question 14 of the survey asked students, "When a person has a disease that cannot be cured, do you think doctors should be allowed by law to end the patient's life by some painless means if the patient and his or her family request it? (Yes, No)" We want to check if there is a relationship between Party and Disease among ETSU students. Assume the students who took the (FALL 2018 Math 1530) class survey are from an SRS of ETSU students.

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



A stacked bar graph could also have been constructed.

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: Party   Columns: Disease

	No	Yes	All
Democrat	29	196	225
Independent	52	272	324
Republican	106	283	389
All	187	751	938

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 is an Independent *and* says yes that when a person has a disease that cannot be cured the doctors should be allowed by law to end the patient's life by some painless means if the patient and his or her family request it?

P = 272 /938 = 0.29 = 29%

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

d. What is the probability that this student is an Independent *or* says yes that when a person has a disease that cannot be cured the doctors should be allowed by law to end the patient's life by some painless means if the patient and his or her family request it?

P = (324 + 751 – 272)/938 = 0.8561= 85.61%

e. What is the probability that this student is a Republican *given* the student says no that when a person has a disease that cannot be cured the doctors should be allowed by law to end the patient's life by some painless means if the patient and his or her family request it?

P = 106/187 = 0.5668 = 56.68%

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

i. State the null and alternative hypotheses.

H_0: There is no relationship between PARTY and DISEASE
H_a: There is a relationship between PARTY and DISEASE

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

	No	Yes	All
Democrat	29	196	225
	44.86	180.14	
Independent	52	272	324
	64.59	259.41	
Republican	106	283	389
	77.55	311.45	
All	187	751	938

Cell Contents
Count
Expected count

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	23.102	2	0.000

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

iii. Which test statistic are you using and what is its value? **A chi-square test statistic and its value is 23.102**

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 party and disease.**

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

Conditions for inference about a chi-square test:

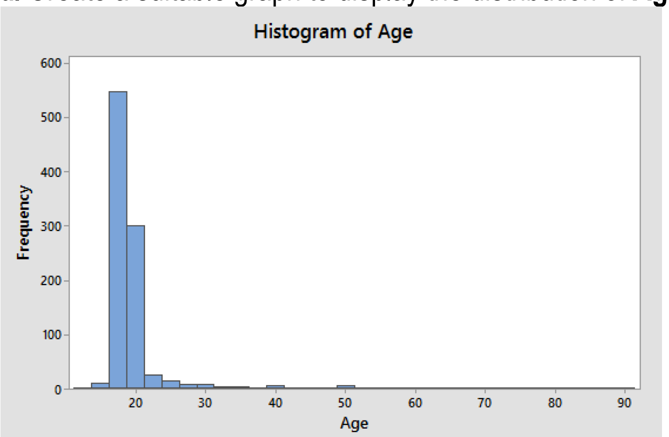
*** No more than 20% of the expected counts are less than 5 and all individual expected counts are greater than 0.** **Less than 20% of the expected counts are less than 5. In this case, there are two less than 5. Thus, violated the condition for this inference and thus, the results may not be valid.**

*** 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.**

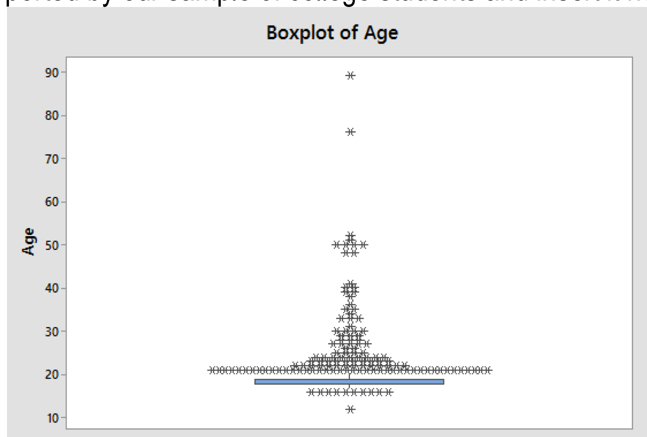
Name:
E-number:
Sum of E-number:
Math 1530 section number

Problem 6: Question 1 of the survey asked students, "What is your age?" and Question 11 of the survey asked students, "Do you think the use of marijuana should be made legal? (Yes, No)" On average, is there a difference in the mean age for those who said marijuana should be made legal and those who said marijuana should not be made legal?

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



OR



Using $\alpha = 0.05$, perform a test of significance to see if, on average, there is a difference in the mean age for those who said marijuana should be made legal and those who said marijuana should not be made legal.

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

H_o: $\mu_{yes} = \mu_{no}$

H_a: $\mu_{yes} \neq \mu_{no}$

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

Assuming Unequal Variances:

Method

μ_1: mean of Age when Marijuana = No
μ_2: mean of Age when Marijuana = Yes
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Test

Null hypothesis	H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis	H_1: $\mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
-0.94	682	0.348

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

Assuming Equal Variances:

Method

μ_1: mean of Age when Marijuana = No
μ_2: mean of Age when Marijuana = Yes
Difference: $\mu_1 - \mu_2$

Equal variances are assumed for this analysis.

Test

Null hypothesis	$H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis	$H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
-0.84	936	0.403

d. What is the name of your test statistic and what is its value?

***t* test statistic, t = -0.94 assuming unequal variance and t = -0.84 assuming equal variance**

e. What is the P-value for the test? **P-value = 0.348 assuming unequal variance and P-value = 0.403 assuming equal variance**

f. State your decision regarding the hypotheses being tested.

Because the P-value is large (greater than $\alpha = 0.05$) in both situations (unequal and equal variances), we *fail to reject* the null hypothesis.

g. State your conclusion. USE COMPLETE SENTENCES.

Based on the sample data provided, we fail to reject the null hypothesis. We believe, on average, there is a not a difference in the mean age for those who said marijuana should be made legal and those who said marijuana should not be made legal.

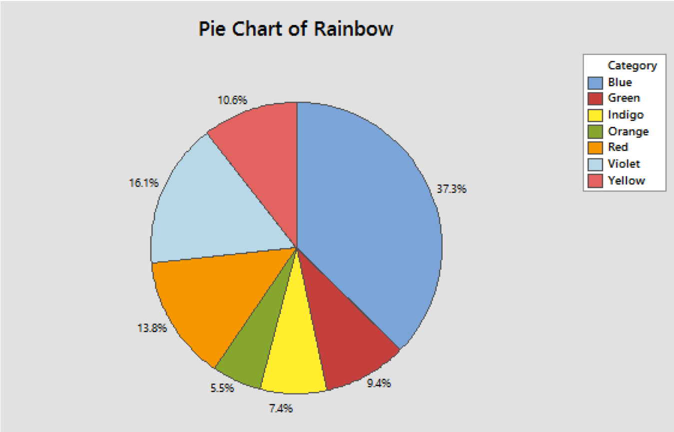
h. What assumptions are you making in order to carry out this test? Is the P-value valid in this case?

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 = 978, is large enough for the t-statistic to be valid.

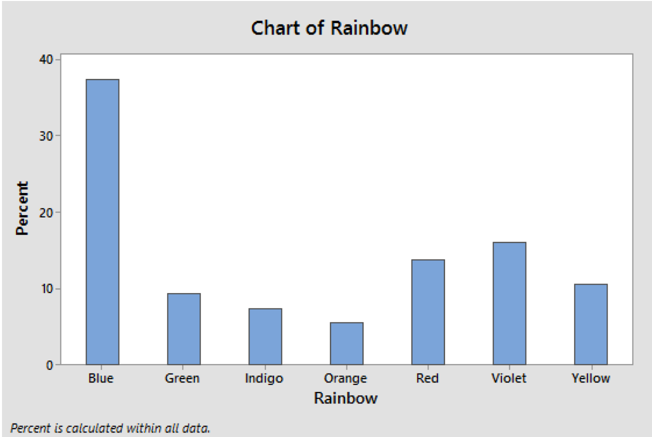
Name:
E-number:
Sum of E-number:
Math 1530 section number

Problem 7 Question 9 from the FALL 2018 Math 1530 survey asked students "What is your favorite color in a rainbow? (Red, Orange, Yellow, Green, Blue, Indigo, Violet)." Many surveys have been conducted on a person's favorite color of the rainbow is blue. One studied stated that 44% of people, stated blue was their favorite color of the rainbow (http://awp.diaart.org/km/usa/survey.html). Is the same true for the population of all U.S. college/university students?

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



OR



b. How many of the students surveyed said "Blue" was their favorite color in the rainbow? **350**

Rainbow	Count
Blue	350
Green	88
Indigo	69
Orange	52
Red	129
Violet	151
Yellow	99
N=	938

c. What proportion of our sample said "Blue" was their favorite color in the rainbow? **350/938 = 37.31%**

Tally

Rainbow	Count	Percent
Blue	350	37.31
Green	88	9.38
Indigo	69	7.36
Orange	52	5.54
Red	129	13.75
Violet	151	16.10
Yellow	99	10.55
N=	938	

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

d. Assume (for the purpose of this problem) that we may treat the FALL 2018 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 90% confidence interval for the proportion of students in the population who chose "Blue" to the survey question (based on our sample data). Copy and paste the Minitab output here.

Exact:

Method

p: event proportion
Exact method is used for this analysis.

Descriptive Statistics

N	Event	Sample p	90% CI for p
938	350	0.373134	(0.346949, 0.399912)

Normal Approximation:

Method

p: event proportion
Normal approximation method is used for this analys

Descriptive Statistics

N	Event	Sample p	90% CI for p
938	350	0.373134	(0.347160, 0.399109)

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

With 90% confidence, the true proportion of students who would chose "Blue" to the survey question is between 34.69% and 40%.

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 44% is not within the calculated confidence interval, <u>our sample suggests that the proportion of US college/university students that chose "Blue" to the survey question is not 44%.</u> Therefore, 44% is not within the 90% 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 will be 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.