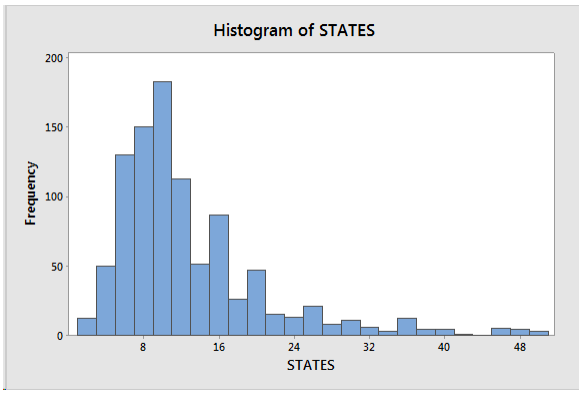
Name:
E number:
Math 1530 section number

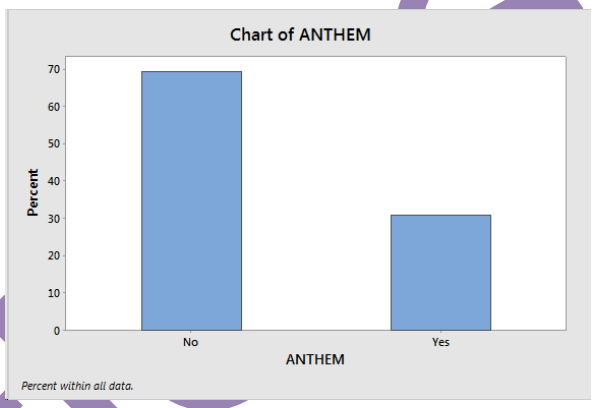
MATH 1530 CAPSTONE TECHNOLOGY PROJECT SOLUTIONS FALL 2016

Problem 1: Identify Variable Type. One of these is a variable that is categorical and one is quantitative. Consider the different graphs that correspond to each variable type. Use Minitab to create a graph appropriate for **each** variable's type. Interpret each graph.

STATES
Quantitative



ANTHEM
Categorical



Note: Other appropriate graphs may have been produced.

STATES: The shape of the distribution is right skewed with a few possible outliers. The spread of the distribution is from 1 to 50 with the center around 10 states.

ANTHEM: We see that the majority of students responded "No" to this question.

Name:
E number:
Math 1530 section number

Problem 2: Sampling. In the survey data, the variable "AGE" is the current age reported by 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 (yrs)	**20**	**25**	**19**	**18**	**29**	**18**	**20**	**19**	**19**	**18**

b. The mean age of the first 10 students is **20.5** years. (Type the value into the space provided.)

c. Identify the type of sampling method you have just used: **Convenience 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	**54**	**134**	**248**	**181**	**561**	**272**	**38**	**23**	**508**	**481**
AGE (yrs)	**18**	**19**	**18**	**18**	**32**	**21**	**20**	**27**	**18**	**20**

e. Calculate and report the mean age for your random sample of 10 students. The sample mean age is **21.5** years. **ANSWERS WILL VARY**

f. Identify the type of sampling method you have just used: **Simple Random Sampling**

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 20.35 Discuss (two or more complete sentences) the **differences and similarities** between 20.35 and the answers you got in (b) and (e).

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

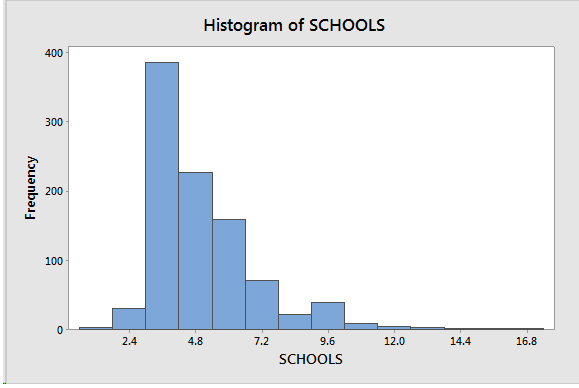
The 'convenience sample' mean *20.5 very slightly* overestimates the 'population' mean 20.35, but that hardly matters. As it is not a **random** sample, there is no long-run guarantee that means from such samples would or would not come close to the population mean. The 'SRS' mean of *21.5* is *below* the population mean of 20.35. 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

<u>Problem 3(e)</u>: 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 10 of the FALL 2016 survey asked students, "How many schools have you attended (including elementary, middle school, high school, trade school, college, etc…)?"

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

Descriptive Statistics: SCHOOLS

Variable	N	N*	Mean	StDev	Variance	Minimum	Q1	Median	Q3	Maximum	IQR
SCHOOLS	956	6	5.1339	1.9429	3.7747	1.0000	4.0000	5.0000	6.0000	17.0000	2.0000

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

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

e. What is the value of that statistic? **5**

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

g. What is (are) the value(s) of that (those) statistic(s)? **Q1 = 4, Q3 = 6, and possible IQR = 2**

Name:
E number:
Math 1530 section number

h. Look up the IQR rule on page 55 in our textbook. Are there any outliers in this distribution? If so, what are their values? How many are there? Justify your answer.

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

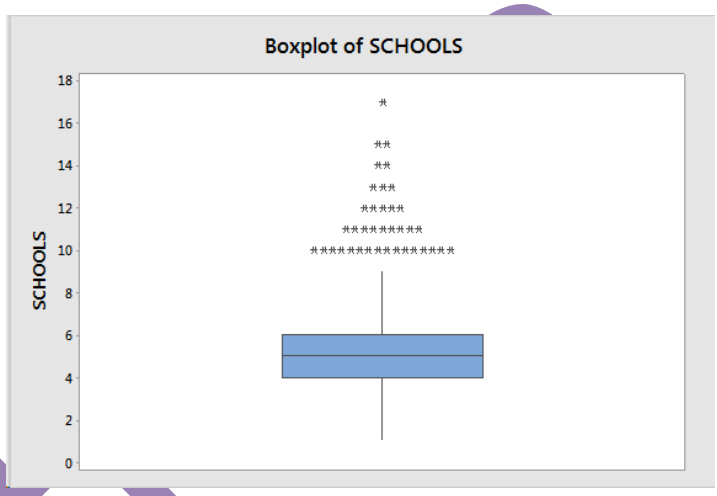
IQR = Q3 – Q1 = 6 – 4 = 2, so 1.5 * IQR = 1.5 * 2 = 3.

Q1 – 1.5 * IQR = 4 – 3 = 1 and Q3 + 1.5 * IQR = 6 + 3 = 9

Any value of 'School' *below* 1 or *above* 9 would be considered outliers. A tally table shows 38 outliers.

Yes, there are definitely outliers in the distribution of 'School'

Minitab shows outliers with * on a boxplot:

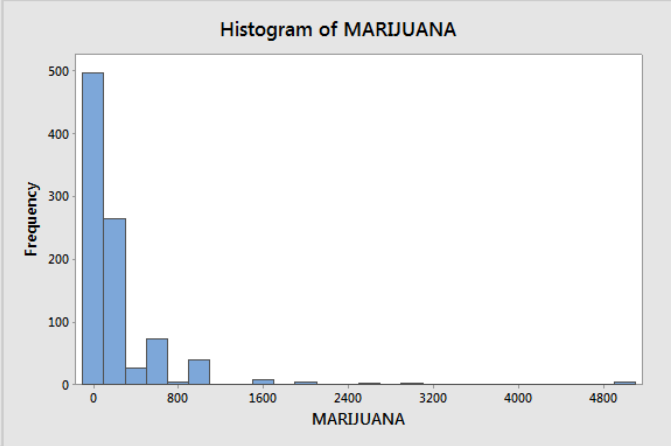


Name:
E number:
Math 1530 section number

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 8 of the FALL 2016 survey asked students, "If a student is caught on campus with marijuana, how much should that student be fined? (in US dollars)"

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

Descriptive Statistics: MARIJUANA

Variable	N	N*	Mean	StDev	Variance	Minimum	Q1	Median	Q3	Maximum	IQR
MARIJUANA	917	45	185.9	416.0	173069.4	0.0	0.0	50.0	200.0	5000.0	200.0

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

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

e. What is the value of that statistic? **50**

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

g. What is (are) the value(s) of that (those) statistic(s)? **Q1 = 0, Q3 = 200, and possible IQR = 200**

Name:
E number:
Math 1530 section number

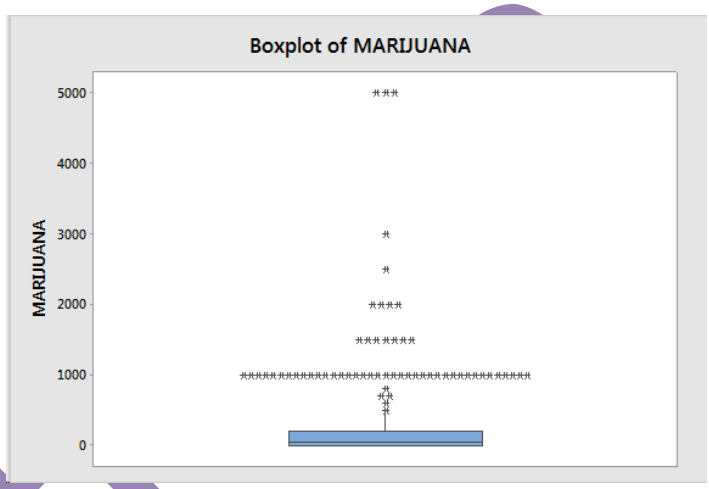
h. Look up the IQR rule on page 55 in our textbook. Are there any outliers in this distribution? If so, what are their values? How many are there? Justify your answer.

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

IQR = Q3 – Q1 = 200 – 0 = 200, so 1.5 * IQR = 1.5 * 200 = 300.

Q1 – 1.5 * IQR = 0 – 300 = -300 and Q3 + 1.5 * IQR = 200 + 300 = 500

Any value of 'MARIJUANA' *below* -300 or *above* 500 would be considered outliers. A tally table shows 60 outliers.



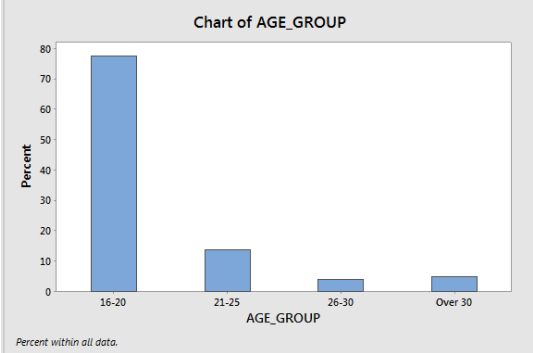
Yes, there are definitely outliers in the distribution of 'MARIJUANA.'

Minitab shows outliers with * on a boxplot:

Name:
E number:
Math 1530 section number

Problem 4: AGE_GROUP versus STATES. Question 2 of the survey asked students, "What is your age (in years)?" This variable was divided into four age groups: Ages "16 to 20", "21 to 25", "26 to 30", and "Over 30". We named this variable **AGE_GROUP**. Question 9 of the survey asked students, "How many U.S. states have you visited?"

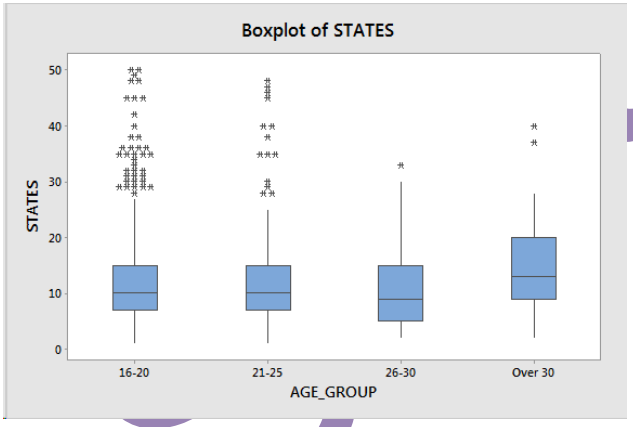
a. Create a suitable graph to display the *distribution* of **AGE_GROUP** and insert it here.



b. What is the mode of this distribution? (Please underline one option.)

Ages 16 to 20 Ages 21 to 25 Ages 26 to 30 Over 30

c. Create a side-by-side boxplot to display the number of states students have visited for the different levels of **AGE_GROUP**. (Go to Graph > Boxplot > One Y with Groups > OK. Select **STATES** for the "Graph variables" slot and **AGE_GROUP** 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 age-group has visited the most U.S. states? **Ages 16 to 20 if reading this question as which age-group had the student that visited the most U.S. States. If reading this question in the overall sense, then the age-group Over 30, which had the highest mean and median.**

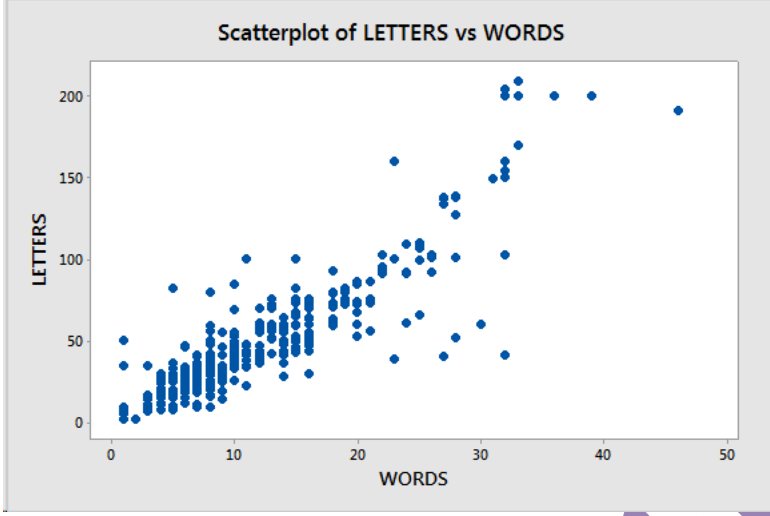
e. Which age-group has the smallest median? **Ages 26 – 30**

f. Which age-group has the largest IQR? **Over 30**

Name:
E number:
Math 1530 section number

Problem 5: WORDS vs. LETTERS. On the FALL 2016 Math 1530 survey, questions 12 and 13 asked students to either write or generate a random sentence. Question 12 asked the students to state the number of words in that sentence and question 13 asked the students to state the number of letters in that sentence. We are interested in seeing whether we can use the number of words to predict the number of letters.

a. Create an appropriate graph to display the relationship between **WORDS** and **LETTERS**. 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 words in a sentence increases, the number letters in the sentence increases.

c. Describe the *form* of the relationship between **WORDS** and **LETTERS.**

Linear

d. Report the value of the correlation between this pair of variables? **r = 0.903**

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 fairly strong.

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

```
Regression Equation

LETTERS = -1.050 + 4.1148 WORDS
```

If used Minitab 16: `LETTERS = - 1.05 + 4.11 WORDS`

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

Name:
E number:
Math 1530 section number

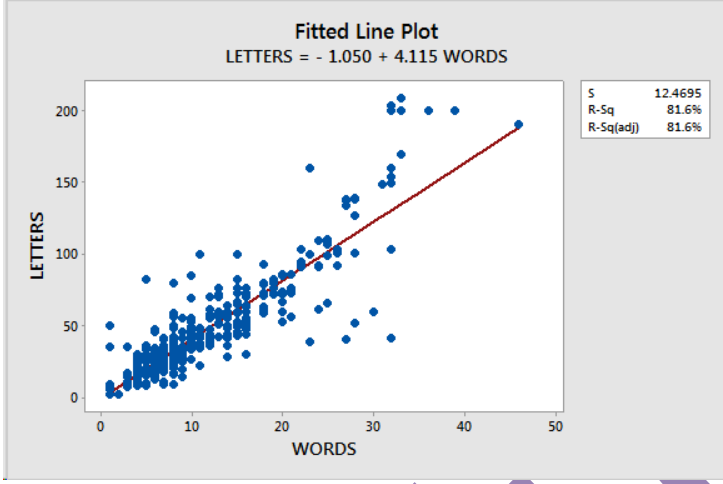
For every additional word in a sentence, the estimated number of letters increases by 4.1148 letters.

h. Use the regression equation in part (f) to predict the number of letters for a sentence that has 8 words. (Show your math.)

Predicted number of letters = -1.050 + 4.1148*8 = 31.8684 letters

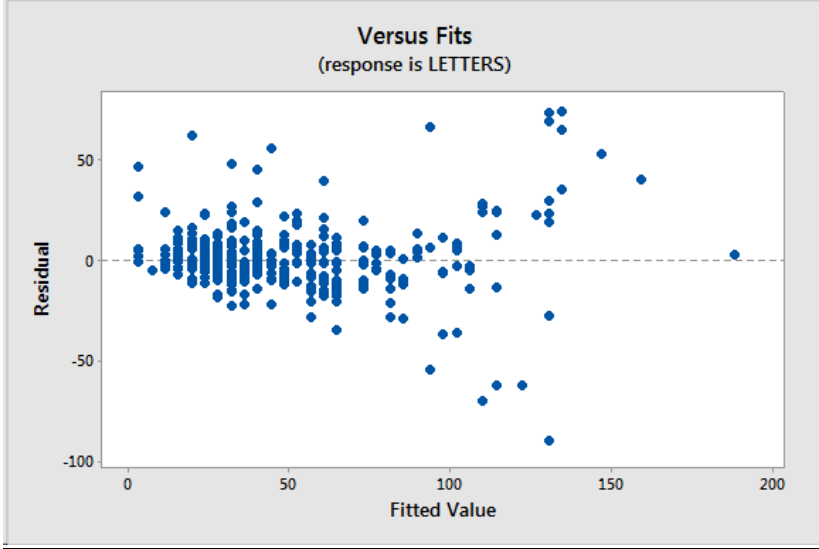
If used Minitab 16: 31.83 letters

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 moderate to strong one and can be seen clearly in the fitted line plot. There are some points that are scattered far away from the regression line. The squared correlation (R^2) indicates that 81.6% of the variation we observed in the number of letters in a sentence is explained by the linear relationship with the number of words in a sentence.

Name:
E number:
Math 1530 section number



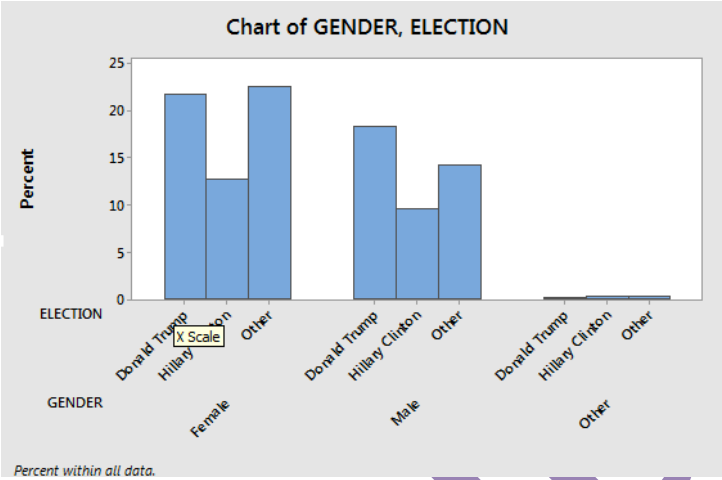
Note: Another scatterplot that is useful to see whether the model makes sense is the residual plot. This helps in determining the appropriateness of the regression model. Recall that the residuals are Residual = Observed Data – Predicted Data. The residual plot shouldn't have any interesting features,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

Name:
E number:
Math 1530 section number

<u>Problem 6 (T):</u> 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.)

GENDER AND ELECTION Question 1 from the FALL 2016 Math 1530 survey asked students "What gender do you identify with? (Female, Male, Other)" and Question 5 from the FALL 2016 Math 1530 survey asked students "In the upcoming 2016 U.S. presidential election, which presidential candidate do you plan to vote for? (Hillary Clinton, Donald Trump, Other)" We want to check if there is a relationship between **GENDER** and **ELECTION** among ETSU students. Assume the students who took the (FALL 2016 Math 1530) class survey are from an SRS of ETSU students.

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



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: ELECTION

	Donald Trump	Hillary Clinton	Other	All
Female	206	121	214	541
Male	174	90	135	399
Other	1	2	3	6
All	381	213	352	946

Cell Contents: Count

Name:
E number:
Math 1530 section number

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

c. What is the probability that this student is female *or* will vote for Hillary Clinton?

P = (541 + 213 – 121)/946 = 0.6691 = 66.91%

d. What is the probability that this student is male given that the student will vote for Donald Trump?

P = 174/381= 0.4567 = 45.67%

e. What is the probability that this student will vote for Donald Trump given that the student is male?

P = 174/399= 0.4361 = 43.61%

f. Do you think there may be an association between **GENDER** and **ELECTION**? Why or why not? Explain your reasoning based on what you see in your graph.

Although there are slight differences, the two bar graphs appear similar to each other. I don't see an association between GENDER and ELECTION (based on this graph). Exact wording will certainly vary on questions like this one. Not all sections of Math 1530 will get to the Chi-square test in chapter 23. For students who *have* seen that material, this question practically screams chi-square. In that case

```
Rows: GENDER   Columns: ELECTION

          Donald  Hillary
           Trump  Clinton   Other  All

Female       206      121     214  541
          217.89   121.81  201.30

Male         174       90     135  399
          160.70    89.84  148.47

Other          1        2       3    6
            2.42     1.35    2.23

All          381      213     352  946

Cell Contents:      Count
                    Expected count

Pearson Chi-Square = 5.184, DF = 4, P-Value = 0.269
Likelihood Ratio Chi-Square = 5.360, DF = 4, P-Value = 0.252
```

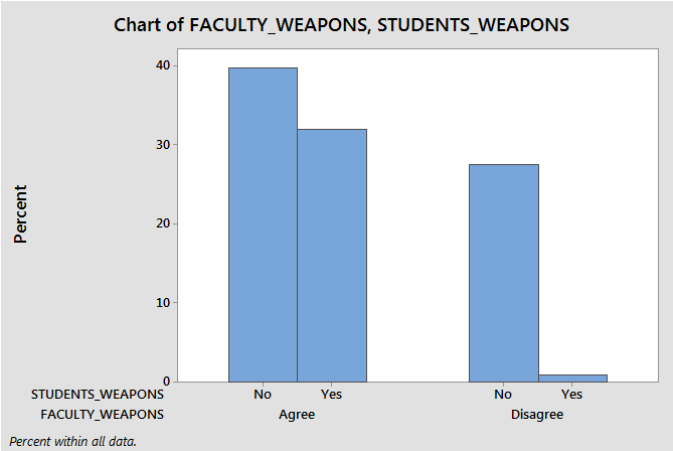
Under the assumption of 'no association' between the variables, the chi-square test, gives a P-Value = 0.252, which is larger than α = 0.05 and suggests there is no association between gender and election. However one should note, that three expected cell counts are less than 5 indicating that the assumptions are not satisfied and the results may not be valid.

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.)

FACULTY_WEAPONS AND STUDENTS_WEAPONS Question 6 from the FALL 2016 Math 1530 survey asked students "Do you agree with ETSU faculty and staff being able to carry concealed weapons on campus? (Agree, Disagree)" and Question 7 from the FALL 2016 Math 1530 survey asked students "Should students be able to carry concealed weapons on campus? (Yes, No)" We want to check if there is a relationship between **FACULTY_WEAPONS** and **STUDENTS_WEAPONS** among ETSU students. Assume the students who took the (FALL 2016 Math 1530) class survey are from an SRS of ETSU students.

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



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: FACULTY_WEAPONS Columns: STUDENTS_WEAPONS

	No	Yes	All
Agree	375	301	676
Disagree	259	8	267
All	634	309	943

Cell Contents: Count

Name:
E number:
Math 1530 section number

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

c. What is the probability that this student is agrees faculty should be able to carry concealed weapons on campus *or* students should be able to carry concealed weapons on campus?

P = (676 + 309 – 301)/943 = 0.7253 = 72.53%

d. What is the probability that this student does not agree faculty should be able to carry concealed weapons on campus given that the student thinks students should not be able to carry concealed weapons on campus?

P = 259/634= 0.4085 = 40.85%

e. What is the probability that this student thinks students should not be able to carry concealed weapons on campus given that this student does not agree faculty should be able to carry concealed weapons on campus?

P = 259/267= 0.9700 = 97%

f. Do you think there may be an association between **FACULTY_WEAPONS** and **STUDENTS_WEAPONS**? Why or why not? Explain your reasoning based on what you see in your graph.

There appears to be some differences between the two bar graphs indicating there may be an association between FACULTY_WEAPONS and STUDENTS_WEAPONS? (based on this graph). Exact wording will certainly vary on questions like this one. Not all sections of Math 1530 will get to the Chi-square test in chapter 23. For students who *have* seen that material, this question practically screams chi-square. In that case

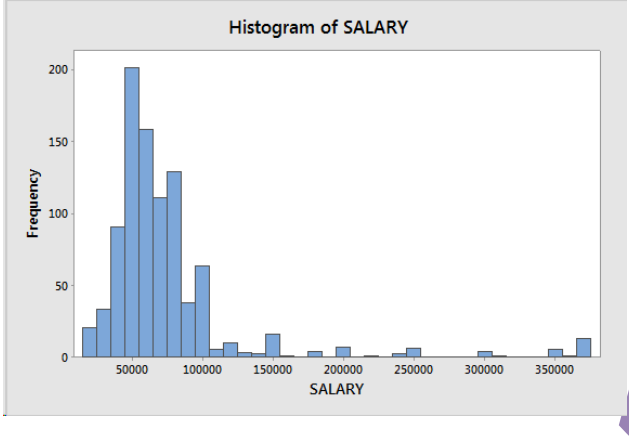
```
Rows: FACULTY_WEAPONS   Columns: STUDENTS_WEAPONS

                No     Yes  All

Agree          375     301  676
            454.5   221.5

Disagree       259       8  267
            179.5    87.5

All            634     309  943

Cell Contents:      Count
                    Expected count


Pearson Chi-Square = 149.849, DF = 1, P-Value = 0.000
Likelihood Ratio Chi-Square = 192.035, DF = 1, P-Value = 0.000
```

Under the assumption of 'no association' between the variables, the chi-square test, gives a P-Value = 0.000, which is smaller than α = 0.05 and suggests there is and association between FACULTY_WEAPONS and STUDENTS_WEAPONS.

Name:
E number:
Math 1530 section number

Problem 7: In 2015, the National Association of Colleges and Employers found that the average starting salary for a bachelor's degree graduate from the Class of 2015 is $50,651 (http://www.naceweb.org/s11182015/starting-salary-class-2015.aspx). Question 11 of the survey asked students, "What is your ideal starting salary (yearly not hourly) that you wish to make after graduating college? (in US dollars)" Is ETSU student's ideal starting salary, on average, $50,651 per year?

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



b. Perform a test of significance to see if ETSU college student's ideal starting salary, on average, is the same amount as the starting salary for a bachelor's degree graduate from the Class of 2015. If this is true, then the average **SALARY** reported by ETSU students should be $50,651. Thus,

H_o: μ = 50651 dollars

Write the correct alternative hypothesis for the test: **H_a: μ ≠ 50651 dollars**

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

```
Test of μ = 50651 vs ≠ 50651

Variable    N   Mean   StDev  SE Mean      95% CI          T      P
SALARY    924  74811   56168     1848  (71184, 78437)  13.07  0.000
```

d. What is the name of your test statistic and what is its value? ***t* test statistic, t = 13.07**

e. What is the P-value for the test? **P = 0.000**

Name:
E number:
Math 1530 section number

f. State your decision regarding the hypotheses being tested.

Because the P-value = 0.000 is small, we *reject* the null Hypothesis. We believe H_a: $\mu \neq 50651$ dollars.

g. State your conclusion. USE COMPLETE SENTENCES.

Based on the sample data provided, we did reject the null hypothesis. We believe that ETSU students' ideal starting salary, on average, is the not same amount as the starting salary for a bachelor's degree graduate from the Class of 2015.

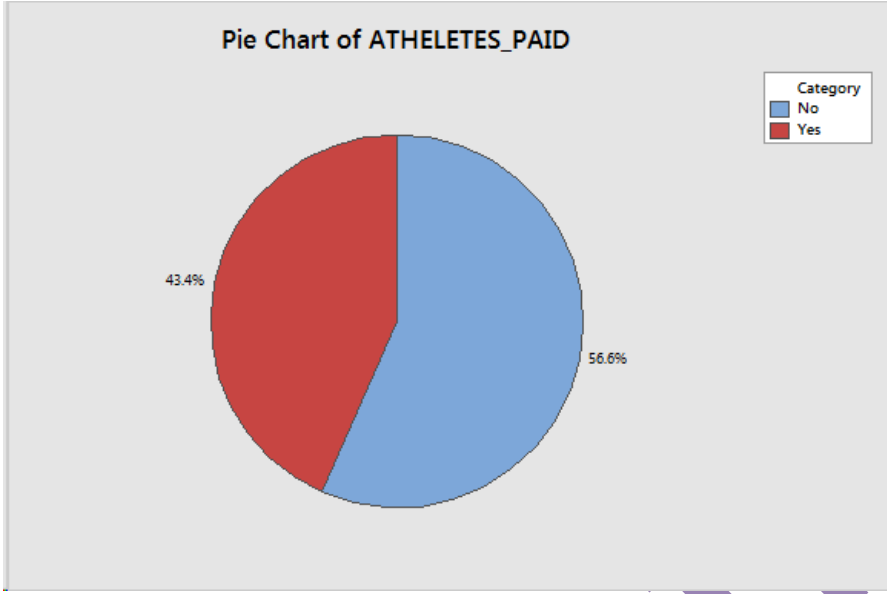
h. Is the P-value valid in this case? What assumptions are you making in order to carry out this test?

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 = 924, is large enough for the t-statistic to be valid.

Name:
E number:
Math 1530 section number

Bonus Problem: Question 3 on the FALL 2016 Math 1530 asked, "Should collegiate athletes be paid to play? (Yes, No)" The YouGov/Huffington Post took a survey of 1000 U.S. adults in October 2015 and reported that and reported that 44% of college athletes should be paid (https://today.yougov.com/news/2015/10/28/poll-results-paying-college-athletes/). Is the same true for the population of all U.S. college/university students?

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



b. How many of the students surveyed said "yes?" **411**

ATHELETES_PAID	Count	Percent
No	535	56.55
Yes	411	43.45
N=	946	
*=	16	

c. What proportion of our sample said "yes?" **43.45%**

ATHELETES_PAID	Count	Percent
No	535	56.55
Yes	411	43.45
N=	946	
*=	16	

Name:
E number:
Math 1530 section number

d. Assume (for the purpose of this problem) that we may treat the FALL 2016 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 "yes" 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       411  946  0.434461  (0.402592, 0.466739)
```

Test and CI for One Proportion

```
Sample    X    N  Sample p          95% CI
1       411  946  0.434461  (0.402874, 0.466048)

Using the normal approximation.
```

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

With 95% confidence, the true proportion of students who would choose "yes" to the survey question is between 40.26% and 46.67%.

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

Because the value 44% is in the calculated confidence interval, our sample suggests that the proportion of US college/university students that chose "yes" to the survey question is 44%. Therefore, 44% is within the 95% CI so our results did appear to be in agreement with the YouGov/Huffington Post.