Name:
E number:
Math 1530 section number

MATH 1530 CAPSTONE TECHNOLOGY PROJECT SPRING 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. CHILDREN Categorical [✓ Quantitative] Neither

State an appropriate plot for this variable: **Boxplot or Histogram**

b. RELATIONS [✓ Categorical] Quantitative Neither

State an appropriate plot for this variable: **Bar graph or Pie Chart**

c. CLASS [✓ Categorical] Quantitative Neither

State an appropriate plot for this variable: **Bar graph or Pie Chart**

Name:
E number:
Math 1530 section number

Problem 2: Sampling. In the survey data, the variable "**NUMBER**" is the favorite number between 0 and 50 for each student.

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

N	1	2	3	4	5	6	7	8	9	10
NUMBER	**7**	**13**	**41**	**22**	**39**	**3**	**31**	**12**	**7**	**8**

b. The mean favorite of the last 10 students is 18.3. (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 "**NUMBER**" 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	**267**	**53**	**441**	**582**	**75**	**432**	**768**	**139**	**774**	**710**
NUMBER	**33**	**22**	**13**	**21**	**3**	**6**	**41**	**23**	**7**	**30**

d. Calculate and report the mean favorite number for your random sample of 10 students. The sample mean favorite number is 19.9. **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* favorite number is 16.454. Discuss (two or more complete sentences) the **differences and similarities** between 16.454 and the answers you got in (b) and (d).

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

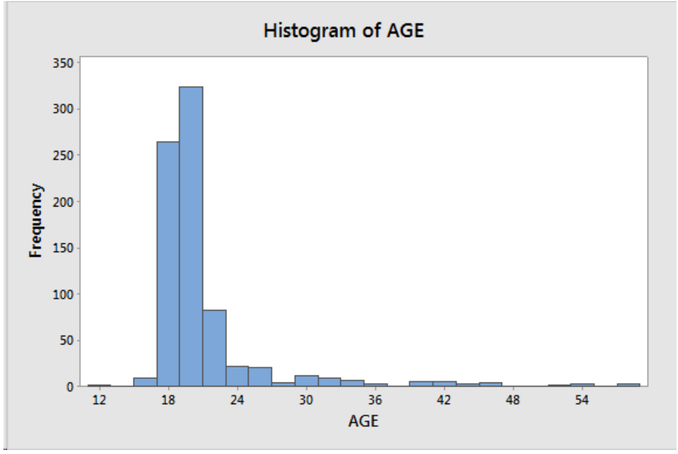
The 'convenience sample' mean found in (b.) *18.3* overestimates the 'population' mean 16.454, but that hardly matters. As it is not a random sample, there is no long-run guarantee that means from such samples would or would not come close to the population mean. The 'SRS' mean found in part (d.) of *19.9* is *above* the population mean of 16.454. 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.)

Question 1 of the SPRING 2017 survey asked students, “What is your age (in years)?”

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

Descriptive Statistics: AGE

```
Variable    N  N*    Mean  SE Mean  StDev  Variance  Minimum      Q1  Median      Q3  Maximum
AGE       775   0  20.493    0.187  5.215    27.199   12.000  18.000  19.000  20.000   58.000

Variable    IQR
AGE       2.000
```

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

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? **19**

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 = 18, Q3 = 20, and possible IQR = 2**

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.

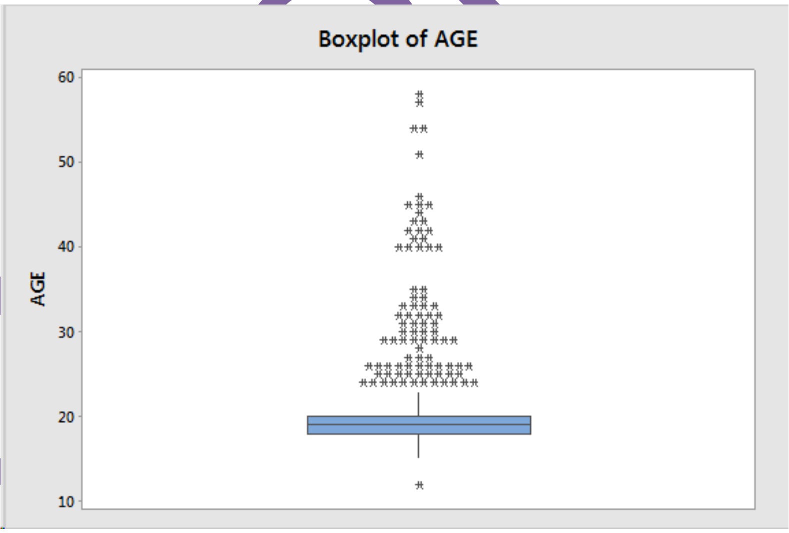
IQR = Q3 – Q1 = 20 – 18 = 2, so 1.5 * IQR = 1.5 * 2 = 3.

Q1 – 1.5 * IQR = 18 – 3 = 15 and Q3 + 1.5 * IQR = 20 + 3 = 23

Any value of 'AGE' *below* 15 or *above* 23 would be considered outliers.

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

Minitab shows outliers with * on a boxplot:

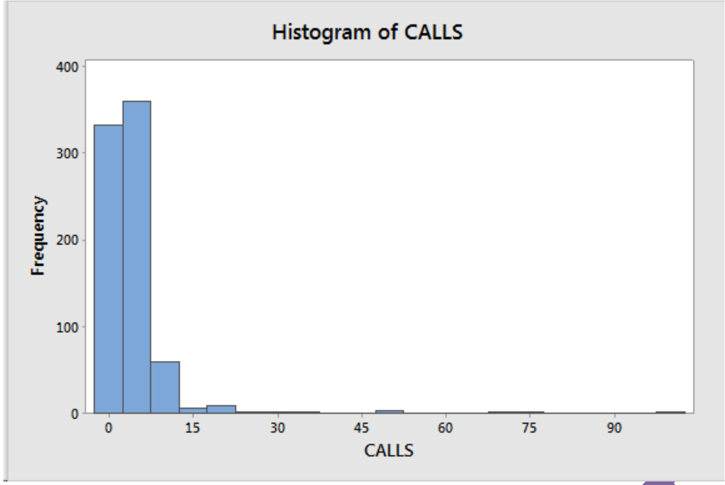


Name:
E number:
Math 1530 section number

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 4 of the SPRING 2017 survey asked students, “Approximately, how many phone calls do you make per day?”

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

Descriptive Statistics: CALLS

```
Variable     N  N*   Mean  SE Mean  StDev  Variance  Minimum     Q1  Median     Q3  Maximum
CALLS      775   0  4.208    0.243  6.755    45.635    0.000  2.000   3.000  5.000  100.000

Variable    IQR
CALLS     3.000
```

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

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 = 2, Q3 = 5, and possible IQR = 3**

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.

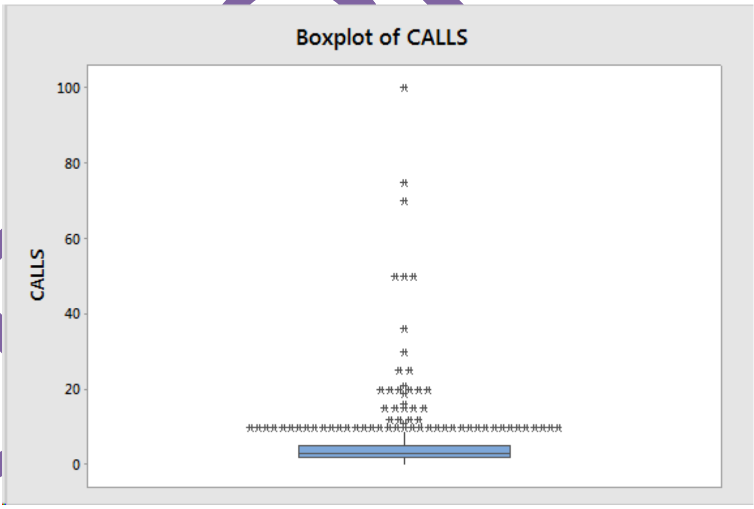
IQR = Q3 – Q1 = 5 – 2 = 3, so 1.5 * IQR = 1.5 * 3 = 4.5.

Q1 – 1.5 * IQR = 2 – 4.5 = -2.5 and Q3 + 1.5 * IQR = 5 + 4.5 = 9.5

Any value of 'CALLS' *below* -2.5 or *above* 9.5 would be considered outliers.

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

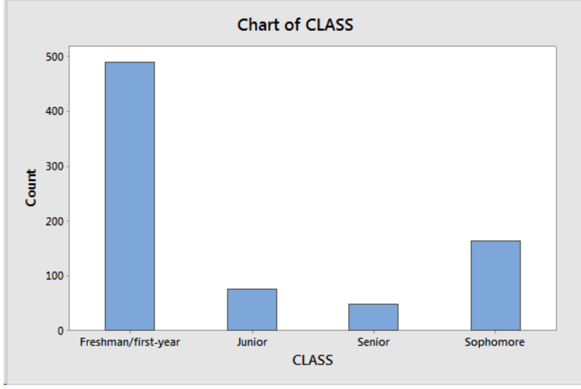
Minitab shows outliers with * on a boxplot:



Name:
E number:
Math 1530 section number

Problem 4: CLASS versus AGE. Question 1 of the survey asked students, “What is your age (in years)?” Question 2 of the survey asked students, “What is your classification in college? (Freshman/first-year, Sophomore, Junior, Senior)”

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

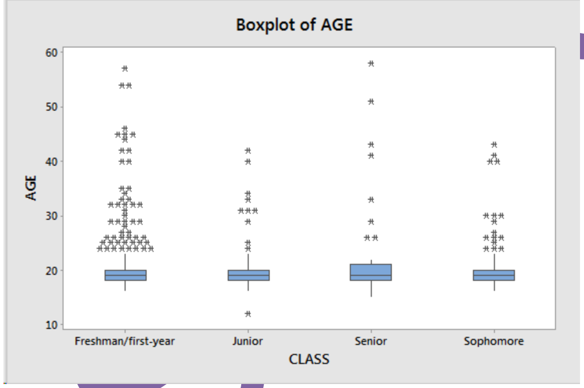


Note: Other appropriate graph(s) may have been produced.

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

[✓ Freshman/first-year] Sophomore Junior Senior

c. Create a side-by-side boxplot to display the age of students for the different levels of **CLASS**. (Go to Graph > Boxplot > One Y with Groups > OK. Select **AGE** for the “Graph variables” slot and **CLASS** 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 class has the oldest student? **Senior**

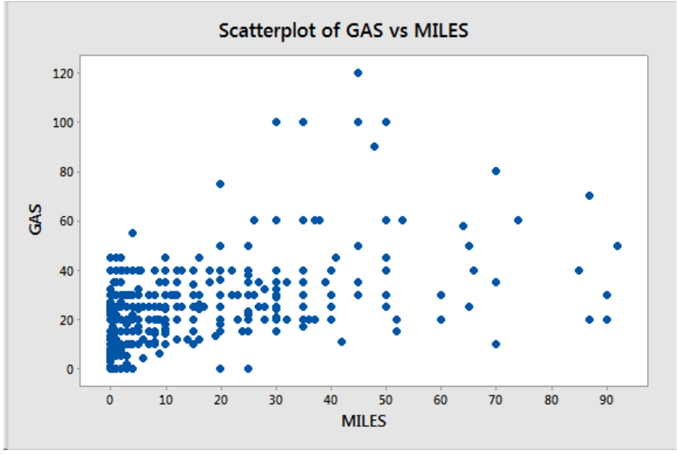
e. Which class has the youngest student? **Junior**

f. Which class has the largest IQR? **Senior**

Name:
E number:
Math 1530 section number

<u>Problem 5: MILES vs. GAS.</u> On the SPRING 2017 Math 1530 survey, question 12 asked students, “Approximately, how many miles do you live from campus? (Enter 0 if you live on campus)” and question 13 asked students, “Approximately, how do you spend on gas (in U.S. dollars) a week?” We are interested in seeing whether we can use the number of miles to predict the amount of dollars spent on gas for a week.

a. Create an appropriate graph to display the relationship between **MILES** and **GAS**. 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 miles increases, the amount you spend on gas in a week increases.

c. Describe the *form* of the relationship between **MILES** and **GAS.**

Linear

d. Report the value of the correlation between this pair of variables? **r = 0.494**

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 moderate.

f. Using Minitab, obtain the equation for the least squares regression of **GAS** against **MILES**. Copy & paste the output here.

```
The regression equation is
GAS = 16.11 + 0.4824 MILES
```

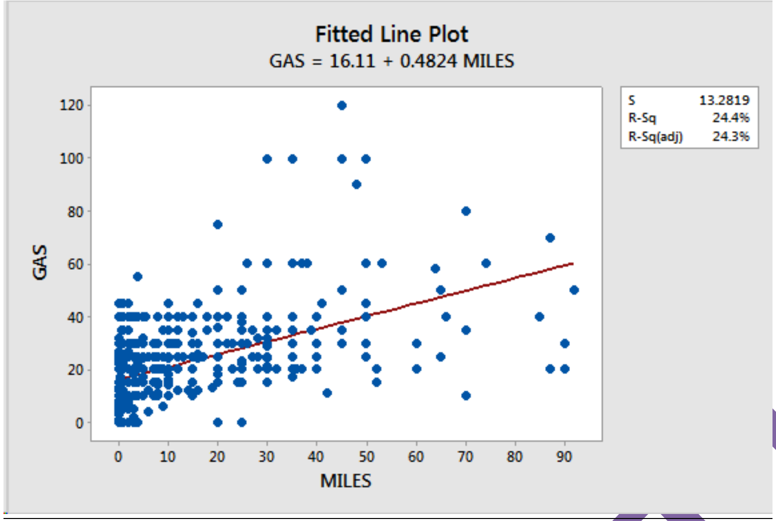
g. Interpret the value of the slope in the least squares regression equation you found in part (f).

For every additional mile lived from campus, the estimated dollars spent on gas in week increase by $0.4824.

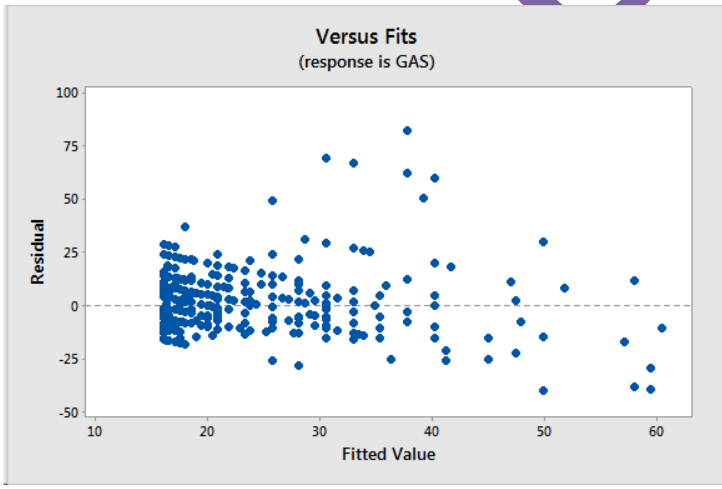
h. Use the regression equation in part (f) to predict amount of dollars spent on gas for a week for a student that lives 5 miles from campus. (Show your math.)

Predicted amount of dollars = 16.11 + 0.4824*5 = $18.52

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 24.3% of the variation we observed in amount spent on gas in a week is explained by the linear relationship with the number of miles a student lives from campus.



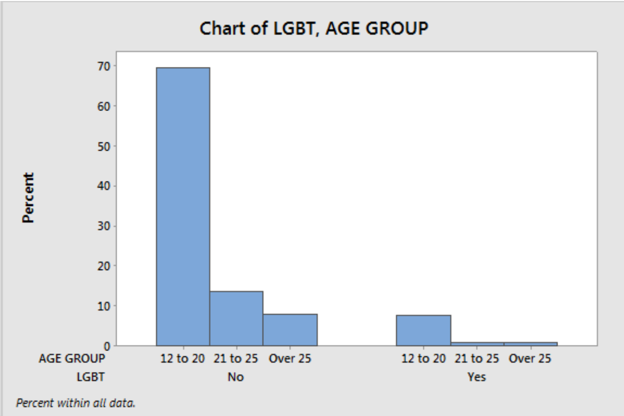
Note: Another scatterplot that is useful to see whether the model makes sense is the residual plot. This helps in determining the appropriateness of the regression model. Recall that the residuals are Residual = Observed Data – Predicted Data. The residual plot shouldn't have any interesting features,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

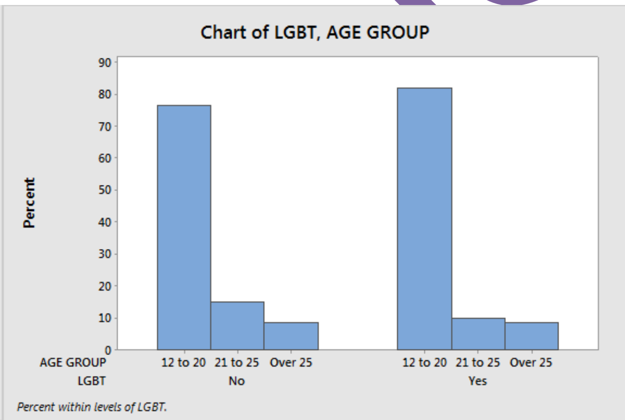
Problem 6 (T): Flip a fair coin. If it lands on tails do this problem (Omit this page/problem AND DO PROBLEM 6(H) if it lands on heads.)

LGBT AND AGE GROUP Question 9 from the SPRING 2017 Math 1530 survey asked students "In the U.S., more Americans are identifying as LGBT. Do you, personally, identify as lesbian, gay, bisexual, or transgender? (Yes, No)" and Question 1 of the survey asked students, "What is your age (in years)?" This variable was divided into three age groups: Ages "12 to 20", "21 to 25", and "Over 25". We named this variable **AGE GROUP**. We want to check if there is a relationship between **LGBT AND AGE GROUP** among ETSU students. Assume the students who took the (SPRING 2017 Math 1530) class survey are from an SRS of ETSU students.

a. Create an appropriate **graph** to display the relationship between **LGBT** and **AGE GROUP**. 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*").**

Tabulated Statistics: LGBT, AGE GROUP

```
Rows: LGBT   Columns: AGE GROUP

       12 to  21 to  Over
          20     25    25  All

No       538    105    60  703
Yes       59      7     6   72
All      597    112    66  775

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 identifies as LGBT *and* is aged 12 to 20?

P = 59/775 = 0.0761 = 7.61%

d. What is the probability that this student identifies as LGBT *or* is aged 12 to 20?

P = (72 + 597 – 59)/775 = 0.7871 = 78.71%

e. What is the probability that this student does not identify as LGBT given that the student is aged over 25?

P = 60/66= 0.9091 = 90.91%

f. What is the probability that this student is aged over 25 given that the student does not identify as LGBT?

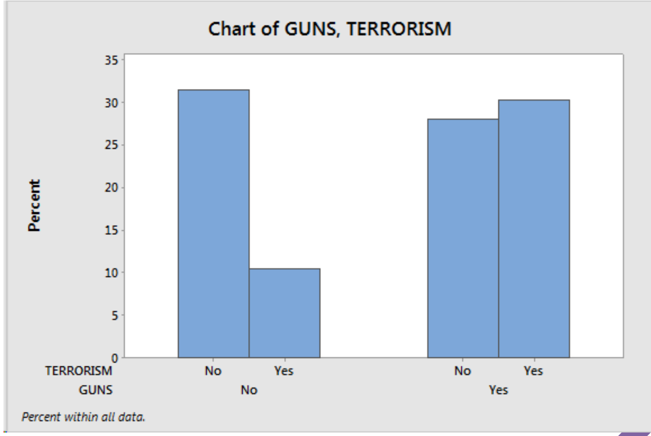
P = 60/703= 0.0853 = 8.53%

Name:
E number:
Math 1530 section number

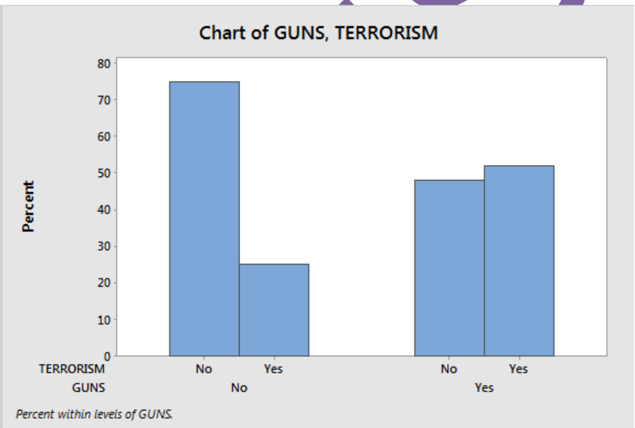
Problem 6 (H): Flip a fair coin. If it lands on heads do this problem (Omit this page/problem AND DO PROBLEM 6(T) if it lands on tails.)

GUNS AND TERRORISM Question 10 from the SPRING 2017 Math 1530 survey asked students "Are you satisfied with America's law or policies on guns? (Yes, No)" and Question 11 from the SPRING 2017 Math 1530 survey asked students "Are you satisfied with America's security with terrorism? (Yes, No)" We want to check if there is a relationship between **GUNS** and **TERRORISM** among ETSU students. Assume the students who took the (SPRING 2017 Math 1530) class survey are from an SRS of ETSU students.

a. Create an appropriate **graph** to display the relationship between **GUNS** and **TERRORISM**. 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*").**

Tabulated Statistics: GUNS, TERRORISM

```
Rows: GUNS   Columns: TERRORISM

        No  Yes  All

No     243   81  324
Yes    217  234  451
All    460  315  775

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 satisfied with America's law or policies on guns *and* the student is satisfied with America's security with terrorism?

P = 234/775 = 0.3019 = 30.19%

Name:
E number:
Math 1530 section number

d. What is the probability that this student is satisfied with America's law or policies on guns *or* the student is satisfied with America's security with terrorism?

P = (451 + 315 – 234)/775 = 0.6865 = 68.65%

e. What is the probability that this student is satisfied with America's law or policies on guns given that the student is not satisfied with America's security with terrorism?

P = 217/460 = 0.4717 = 47.17%

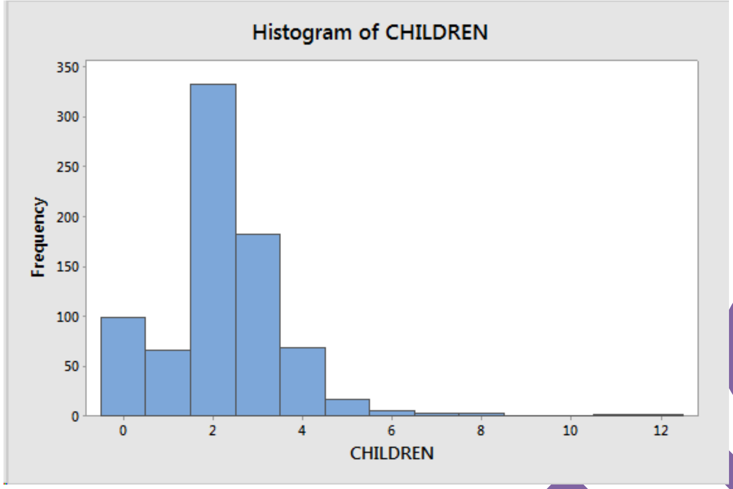
f. What is the probability that this student is not satisfied with America's security with terrorism given that this student is satisfied with America's law or policies on guns?

P = 217/451 = 0.4812 = 48.12%

Name:
E number:
Math 1530 section number

Problem 7: In 2013, Gallup found that the ideal number of children Americans want is 2.6 children per family. (http://www.gallup.com/poll/164618/desire-children-norm.aspx). Question 3 of the survey asked students, "What is your ideal number of children? " A professor feels that this number may be lower for college students. Is ETSU student's ideal number of children, on average, less than 2.6 children?

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



Perform a test of significance to see if ETSU college student's ideal number of children, on average, is lower than the 2.6 children reported by Gallop using $\alpha = 0.05$.

b. Write the correct null and alternative hypothesis for the test: **H_o: $\mu = 2.6$ children versus H_a: $\mu < 2.6$ children**

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

One-Sample T: CHILDREN

```
Test of μ = 2.6 vs < 2.6


Variable     N    Mean   StDev  SE Mean  95% Upper Bound      T      P
CHILDREN   775  2.2168  1.3600   0.0489           2.2972  -7.84  0.000
```

d. What is the name of your test statistic and what is its value? ***t* test statistic, t = -7.84**

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

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: $\mu < 2.6$ children.

Name:
E number:
Math 1530 section number

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 number of children, on average, is less than the reported 2.6 children reported by Gallup.

h. Is the P-value valid in this case?

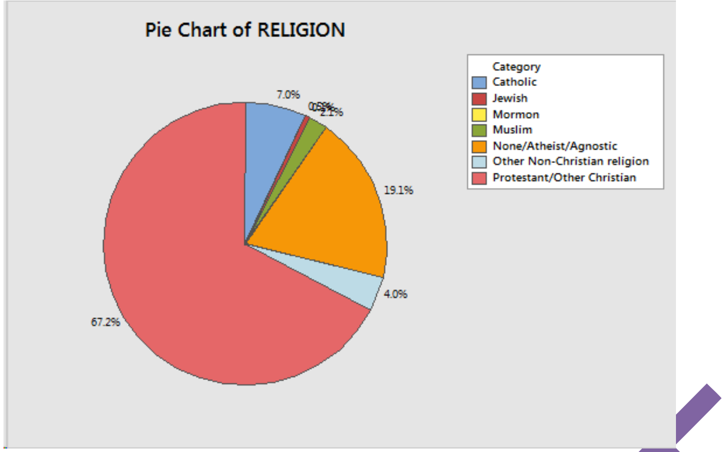
i. What assumptions are you making in order to carry out this test?

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 = 775, is large enough for the t-statistic to be valid.

Name:
E number:
Math 1530 section number

Bonus Problem: Question 13 on the SPRING 2017 Math 1530 asked, “What is your religious identification? (Protestant/Other Christian, Catholic, Mormon, Jewish, Muslim, Other Non-Christian religion, None/Atheist/Agnostic)” The Gallup took a survey of U.S. adults in December 2016 and reported that and reported that 18.2% of U.S. adults said their religion identification was None/Atheist/Agnostic (http://www.gallup.com/poll/200186/five-key-findings-religion.aspx?g_source=Religion&g_medium=newsfeed&g_campaign=tiles). Is the same true for the population of all U.S. college/university students?

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



Note: Other appropriate graph(s) may have been produced.

b. How many of the students surveyed said “None/Atheist/Agnostic?” **148**

Tally for Discrete Variables: RELIGION

```
                    RELIGION  Count  Percent
                    Catholic     54     6.97
                      Jewish      4     0.52
                      Mormon      1     0.13
                      Muslim     16     2.06
       None/Atheist/Agnostic    148    19.10
Other Non-Christian religion     31     4.00
  Protestant/Other Christian    521    67.23
                          N=    775
```

c. What proportion of our sample said “None/Atheist/Agnostic?” **19.10%**

Tally for Discrete Variables: RELIGION

```
                    RELIGION  Count  Percent
                    Catholic     54     6.97
                      Jewish      4     0.52
                      Mormon      1     0.13
                      Muslim     16     2.06
       None/Atheist/Agnostic    148    19.10
Other Non-Christian religion     31     4.00
  Protestant/Other Christian    521    67.23
                          N=    775
```

Name:
E number:
Math 1530 section number

d. Assume (for the purpose of this problem) that we may treat the SPRING 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 “None/Atheist/Agnostic” to the survey question (based on our sample data). Copy and paste the Minitab output here.

(In Minitab, go to Stat > Basic Statistics > 1-proportion, then choose “Summarized data” from the drop-down menu and put in 148 and 775 for the number of events and trials, respectively.)

Test and CI for One Proportion

```
Sample    X    N  Sample p            95% CI
1       148  775  0.190968  (0.163868, 0.220442)
```

Test and CI for One Proportion

```
Sample    X    N  Sample p            95% CI
1       148  775  0.190968  (0.163294, 0.218641)

Using the normal approximation.
```

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

With 95% confidence, the true proportion of students who would chose “None/Atheist/Agnostic” to the survey question is between 16.39% and 22.04%.

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

Because the value 18.2% is in the calculated confidence interval, our sample suggests that the proportion of US college/university students that chose “None/Atheist/Agnostic” to the survey question is 18.2%. Therefore, 18.2% is within the 95% CI so our results did appear to be in agreement with the Gallup poll.