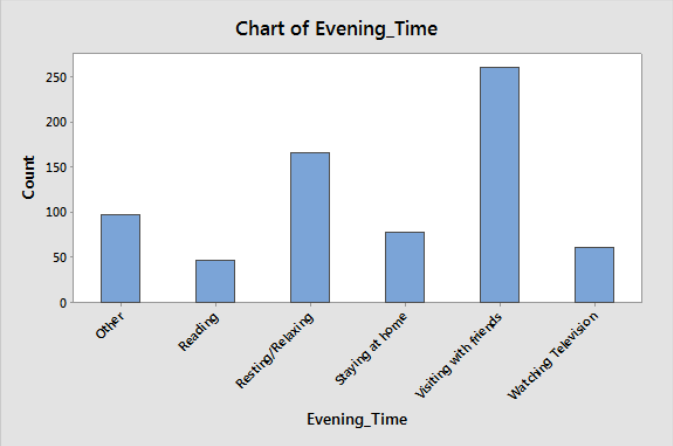
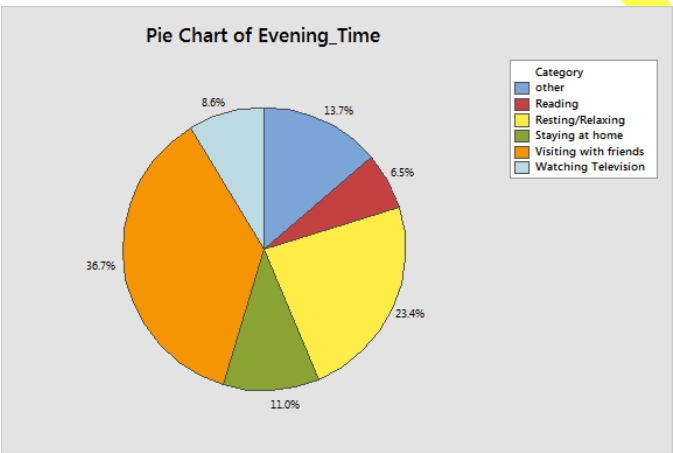
Name:
E number:
Math 1530 section number

MATH 1530 CAPSTONE TECHNOLOGY PROJECT SOLUTIONS SPRING 2016

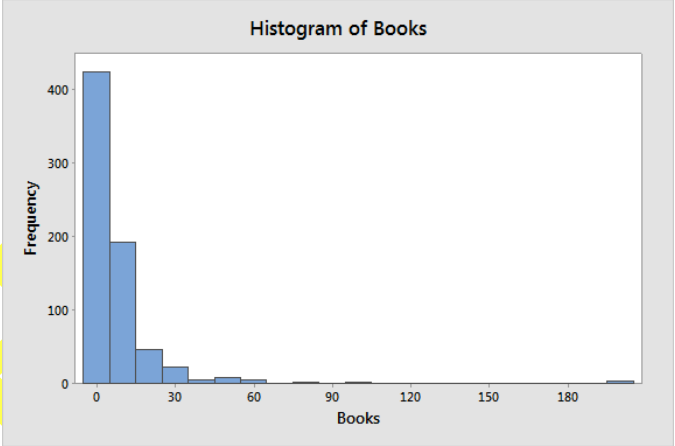
Problem 1: Identify Variable Type. One of these is a variable that is categorical and one is quantitative. Consider the different graphs that correspond to each variable type. Use Minitab to create ***two*** *different* graphs appropriate for **each** variable's type. That is, there is a total of four graphs. EXTRA CREDIT if you can resize to fit all of the graphs on one page.

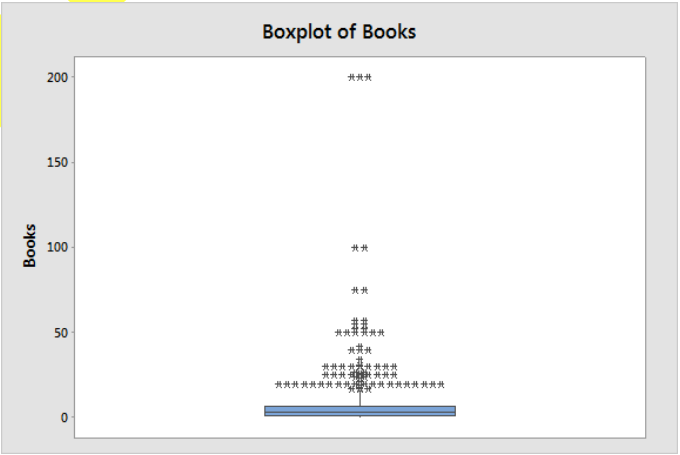
Evening_Time
Categorical





Books
Quantitative





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)	**22**	**19**	**18**	**20**	**17**	**20**	**20**	**18**	**26**	**19**

b. The mean age of the first 10 students is 19.9 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	**214**	**574**	**27**	**192**	**266**	**428**	**128**	**421**	**401**	**57**
AGE (yrs)	**18**	**20**	**20**	**19**	**18**	**18**	**18**	**19**	**44**	**19**

e. Calculate and report the mean age for your random sample of 10 students. The sample mean age is 21.30 years. **ANSWERS WILL VARY**

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

g. REPEAT the random sample selection process three more times. Calculate and report the mean age for each random sample of 10 students. **ANSWERS WILL VARY**

N	1	2	3	4	5	6	7	8	9	10
ID	**325**	**1**	**506**	**385**	**97**	**132**	**161**	**53**	**534**	**656**
AGE (yrs)	**19**	**22**	**39**	**23**	**18**	**19**	**18**	**19**	**19**	**20**

ii) The sample mean age is 21.60 years.

N	1	2	3	4	5	6	7	8	9	10
ID	**273**	**85**	**640**	**489**	**466**	**269**	**377**	**79**	**32**	**554**
AGE (yrs)	**20**	**22**	**24**	**37**	**18**	**19**	**39**	**22**	**19**	**18**

iii) The sample mean age is 23.80 years.

N	1	2	3	4	5	6	7	8	9	10
ID	**236**	**418**	**634**	**699**	**316**	**558**	**455**	**26**	**106**	**667**
AGE (yrs)	**19**	**19**	**19**	**19**	**20**	**20**	**19**	**21**	**23**	**19**

iv) The sample mean age is 19.80 years.

Name:
E number:
Math 1530 section number

h. 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 21.25. Discuss (two or more complete sentences) the **differences and similarities** between 21.25 and the answers you got in (b), (e), and ii), iii), and iv).

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

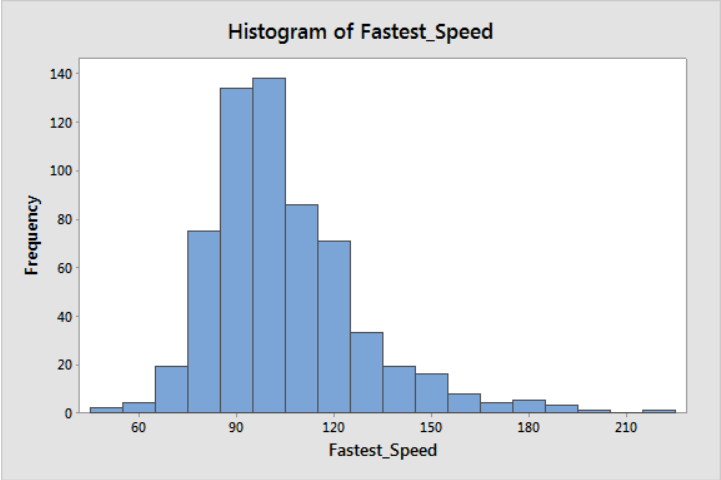
The 'convenience sample' mean *19.9* underestimates the 'population' mean 21.25, but that hardly matters. As it is not a **random** sample, there is no long-run guarantee that means from such samples would or would not come close to the population mean. The 'SRS' mean of *19.80* are *below* the population mean 21.25, yet the 'SRS' means of *23.90, 21.60,* and *21.30* are *above* the population mean. This is consistent with the 'long run' 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 11 of the SPRING 2016 survey asked students, "What is the fastest you have ever driven a car? (in mph)"

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



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

Skewed left Symmetric Skewed right

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

Descriptive Statistics: Fastest_Speed

Variable	N	N*	Mean	StDev	Minimum	Q1	Median	Q3	Maximum	IQR
Fastest_Speed	619	0	103.16	22.97	45.00	90.00	100.00	115.00	215.00	25.00

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

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

e. What is the value of that statistic? **100**

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 = 90, Q3 = 115, and possible IQR = 25**

Name:
E number:
Math 1530 section number

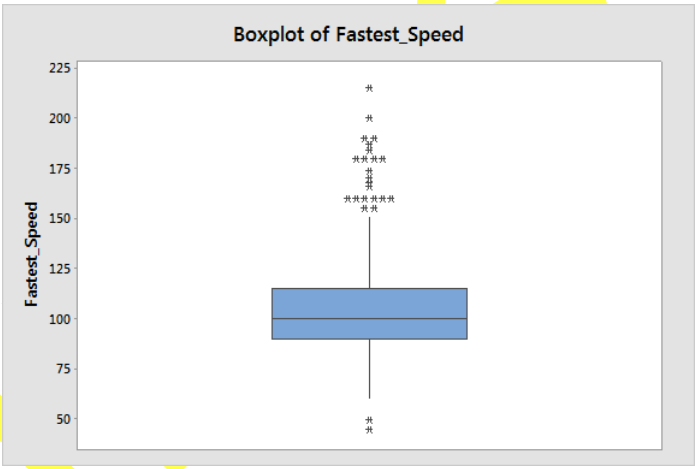
h. Look up the IQR rule on page 55 in our textbook. Are there any outliers in this distribution? If so, what are their values? How many are there? Justify your answer.

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

IQR = Q3 – Q1 = 115 – 90 = 25, so 1.5 * IQR = 1.5 * 25 = 37.5.

Q1 – 1.5 * IQR = 90 – 37.5 = 52.5 and Q3 + 1.5 * IQR = 115 + 37.5 = 152.5

Any value of 'Fastest_Speed' *below* 52.5 or *above* 152.5 would be considered outliers. A tally table shows 24 outliers.



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

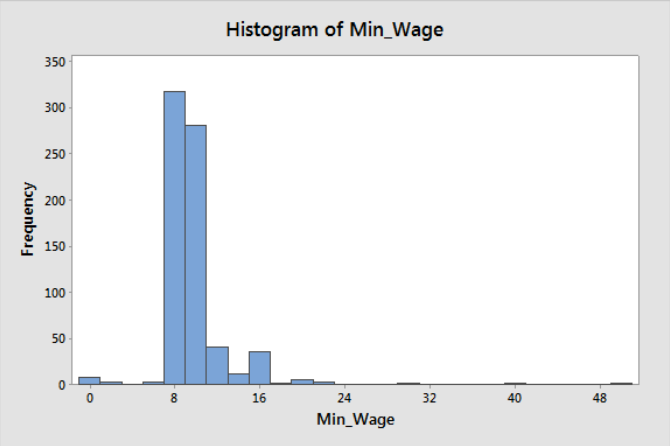
Minitab shows outliers with * on a boxplot:

Name:
E number:
Math 1530 section number

<u>Problem 3(o)</u>: 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 SPRING 2016 survey asked students, "How much do you believe minimum wage should be? (in US dollars)"

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



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

Skewed left Symmetric Skewed right

c. Using Minitab, calculate the basic statistics for the data collected on **Min_Wage** and copy & paste the Minitab output here.

Descriptive Statistics: Min_Wage

Variable	N	N*	Mean	StDev	Minimum	Q1	Median	Q3	Maximum	IQR
Min_Wage	708	0	9.426	3.204	0.000	8.000	9.000	10.000	50.000	2.000

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

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

ii) What is the value of that statistic? **9**

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

iv) What is (are) the value(s) of that (those) statistic(s)? **Q1 = 8, Q3 = 10, and possible IQR = 2**

Name:
E number:
Math 1530 section number

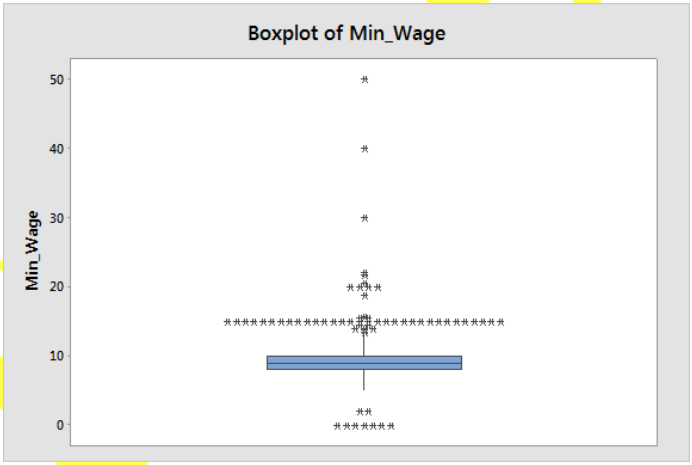
v) Look up the IQR rule on page 55 in our textbook. Are there any outliers in this distribution? If so, what are their values? How many are there? Justify your answer.

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

IQR = Q3 – Q1 = 10 – 8 = 2, so 1.5 * IQR = 1.5 * 2 = 3.

Q1 – 1.5 * IQR = 8 – 3 = 5 and Q3 + 1.5 * IQR = 10 + 3 = 13

Any value of 'Min_Wage' *below* 5 or *above* 13 would be considered outliers. A tally table shows 64 outliers.



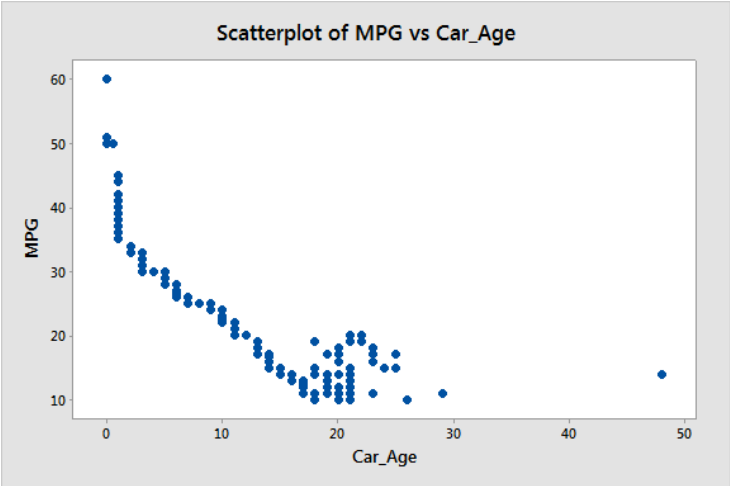
Yes, there are definitely outliers in the distribution of 'Min_Wage.'

Minitab shows outliers with * on a boxplot:

Name:
E number:
Math 1530 section number

Problem 4: CAR_AGE vs. MPG. On the SPRING 2016 Math 1530 survey, questions 13 and 14 asked students to state the age of their car (in years) and the miles per gallon (mpg) for that car. We are interested in seeing whether we can use the age of the car to predict the mpg of the car.

a. Create an appropriate graph to display the relationship between **CAR_AGE** and **MPG**. 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.

Negative association: As car age increases, the MPG decreases

c. Describe the *form* of the relationship between **CAR_AGE** and **MPG.**

I would describe the form as curved but arguably one could describe the form as linear with possible outliers. Thus, accept either answer.

d. Report the value of the correlation between this pair of variables? **r = -0.906**

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 MPG on CAR_AGE. Copy & paste the output here.

```
Regression Equation

MPG = 35.651 - 1.2260Car_Age
```

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

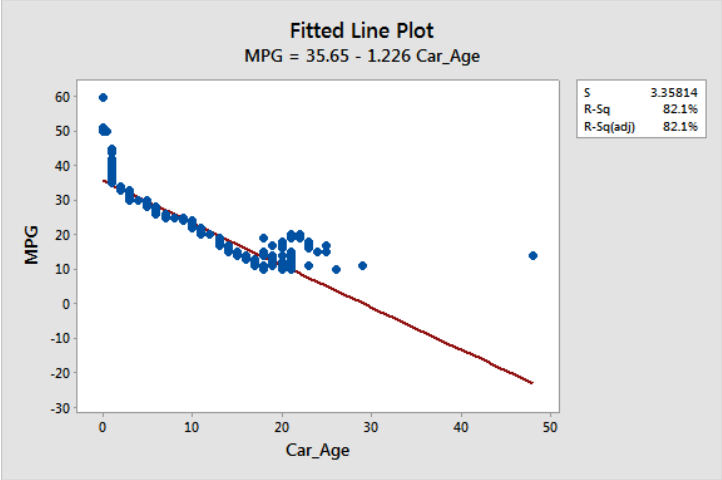
For every 1 year older of the car, the estimated MPG decreases by 1.2260 MPG.

Name:
E number:
Math 1530 section number

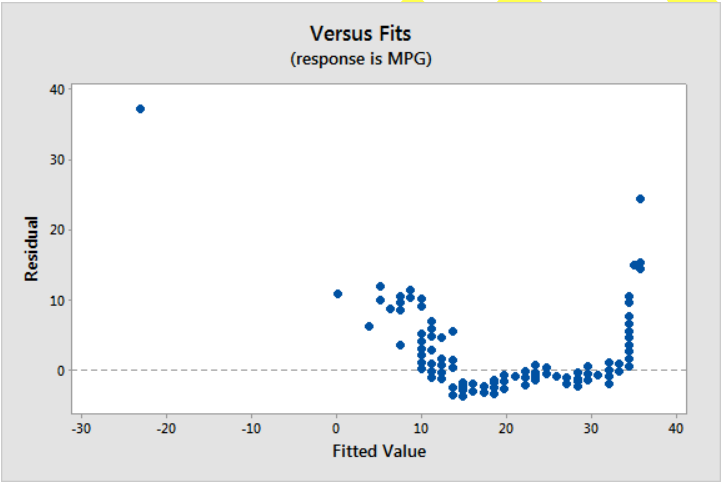
h. Use the regression equation in part (f) to predict the mpg for a car that is 10 years old. (Show your math.)

Predicted mpg = 35.651 - 1.2260*10 = 23.391 MPG

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 82.1% of the variation we observed in the MPG is explained by the linear relationship with the age of the care.



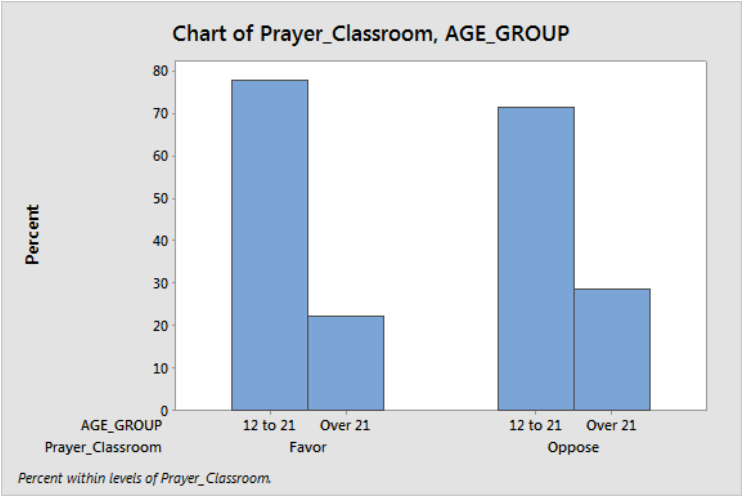
Note: Another scatterplot that is useful to see whether the model makes sense is the residual plot. This helps in determining the appropriateness of the regression model. Recall that the residuals are Residual = Observed Data – Predicted Data. The residual plot shouldn't have any interesting features,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

Name:
E number:
Math 1530 section number

Problem 5 (T): Flip a fair coin. If it lands on tails do this problem (Omit this page/problem AND DO PROBLEM 5(H) if it lands on heads.)

PRAYER_CLASSROOM AND AGE_GROUP Question 5 from the SPRING 2016 Math 1530 survey asked students "Do you favor or oppose daily prayer in the classroom?" and Question 2 from the SPRING 2016 Math 1530 survey asked students "What is your age?" The variable **AGE** was split into two groups: ages "12 to 21" and "Over 21." These two groups formed the variable **AGE_GROUP.** We want to check if there is a relationship between PRAYER_CLASSROOM and AGE_GROUP among ETSU students. Assume the students who took the (SPRING 2016 Math 1530) class survey are from an SRS of ETSU students.

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



b. Create an appropriate two-way table to summarize the data. Insert your table here.

Rows: Prayer_Classroom Columns: AGE_GROUP

	12 to 21	Over 21	All
Favor	355	102	457
Oppose	179	72	251
All	534	174	708

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 either "Over 21" *or* favors prayer in the classroom?

P = (174 + 457 – 102)/708 = 0.7471751 = 74.72%

Name:
E number:
Math 1530 section number

d. What is the probability that this student is aged "12 to 21" given that the student opposes prayer in the classroom?

P = 179/251= 0.7131474 = 71.31%

e. What is the probability that this student opposes prayer in the classroom given that the student is aged "12 to 21"?

P = 179/534 = 0.335206 = 33.52%

f. Do you think there may be an association between PRAYER_CLASSROOM and AGE_GROUP? 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 AGE_GROUP and PRAYER_CLASSROOM (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: Prayer_Classroom   Columns: AGE_GROUP

          12 to
             21  Over 21  All

Favor        355      102  457
Oppose       179       72  251
All          534      174  708

Cell Contents:      Count


Pearson Chi-Square = 3.542, DF = 1, P-Value = 0.060
Likelihood Ratio Chi-Square = 3.491, DF = 1, P-Value = 0.062
```

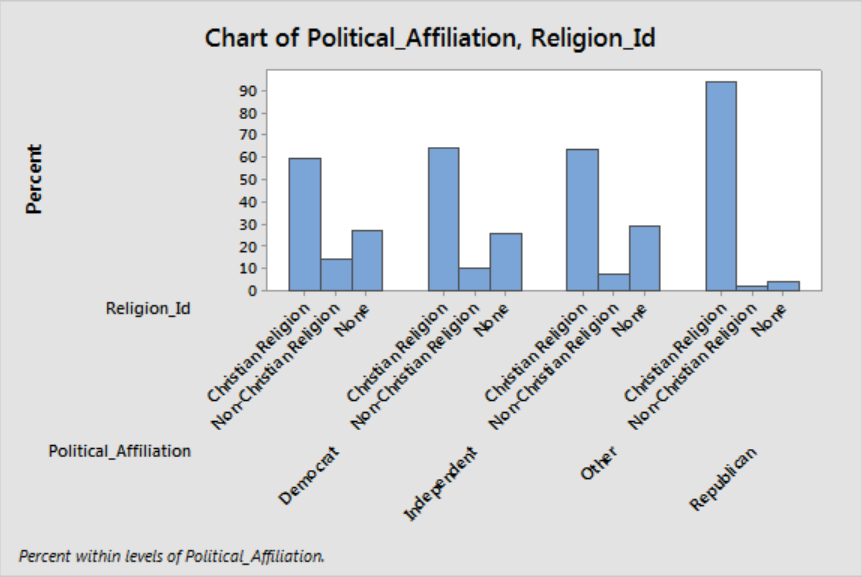
Under the assumption of 'no association' between the variables, the chi-square test, gives a P-Value = 0.062, which is larger than $\alpha = 0.05$ and suggests there is no association between age group and prayer in classroom.

Name:
E number:
Math 1530 section number

Problem 5 (H): Flip a fair coin. If it lands on heads do this problem (Omit this page/problem AND DO PROBLEM 5(T) if it lands on tails.)

POLITICAL_AFFLILATION AND RELIGION_ID Question 6 from the SPRING 2016 Math 1530 survey asked students "What political party do you identify with?" and Question 10 from the SPRING 2016 Math 1530 survey asked students "What is your religious identification?" We want to check if there is a relationship between POLITICAL_AFFLILATION AND RELIGION_ID among ETSU students. Assume the students who took the (SPRING 2016 Math 1530) class survey are from an SRS of ETSU students.

a. Create an appropriate **graph** to display the relationship between POLITICAL_AFFLILATION AND RELIGION_ID. Insert your graph here.



b. Create an appropriate two-way table to summarize the data. Insert your table here.

Rows: Political_Affiliation Columns: Religion_Id

	Christian Religion	Non-Christian Religion	None	All
Democrat	98	23	45	166
Independent	115	18	46	179
Other	83	10	38	131
Republican	217	5	10	232
All	513	56	139	708

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's political affiliation is Independent or identifies with a Non-Christian religion?

P = (179+ 56 – 18)/708 = 0.3064972 = 30.65%

d. What is the probability that this student's political affiliation is Republican given the student identifies with a Christian religion?

P = 217/513 = 0.4230019 = 42.30%

e. What is the probability that this student identifies with a Christian religion given the student's political affiliation is Republican?

P = 217/232 = 0.9353448 = 93.53%

f. Do you think there may be an association between POLITICAL_AFFLILATION AND RELIGION_ID? 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 POLITICAL_AFFLILATION and RELIGION_ID (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: Political_Affiliation Columns: Religion_Id

	Christian Religion	Non-Christian Religion	None	All
Democrat	98	23	45	166
Independent	115	18	46	179
Other	83	10	38	131
Republican	217	5	10	232
All	513	56	139	708

Cell Contents: Count

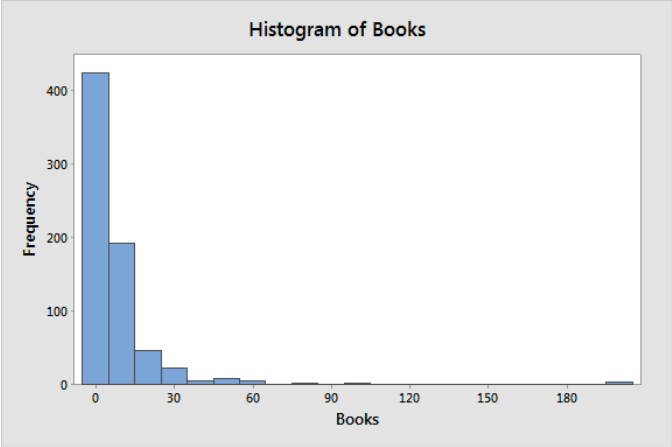
Pearson Chi-Square = 81.388, DF = 6, P-Value = 0.000
Likelihood Ratio Chi-Square = 94.589, DF = 6, P-Value = 0.000

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 political affiliation and religion identification.

Name:
E number:
Math 1530 section number

Problem 6: In 2015, the Pew Research Center found that American adults (aged 18+) read an average of 12 books per year (http://www.pewresearch.org/fact-tank/2015/10/19/slightly-fewer-americans-are-reading-print-books-new-survey-finds/ft_15-10-09_books_averageread/). Do ETSU students read, on average, 12 books per year?

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



b. Describe the distribution shown in your graph.

The graph is right skewed.

c. Perform a test of significance to see if ETSU college students read, on average, the same amount of books in a year as American adults. If this is true, then the average BOOKS reported by ETSU students should be 12. Thus,

H_0: $\mu = 12$ books

Write the correct alternative hypothesis for the test.

H_a: $\mu \neq 12$ books

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

One-Sample T: Books

```
Test of μ = 12 vs ≠ 12

Variable    N   Mean   StDev  SE Mean      95% CI          T      P
Books     708  7.571  16.717    0.628  (6.337, 8.804)  -7.05  0.000
```

e. What is the name of your test statistic and what is its value? ***t* test statistic, t = -7.05**

Name:
E number:
Math 1530 section number

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

g. 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: $\mu \neq 12$ books.

h. State your conclusion. USE COMPLETE SENTENCES.

Based on the sample data provided, we did reject the null hypothesis. We believe that ETSU students not read the same amount of books as the average American adult.

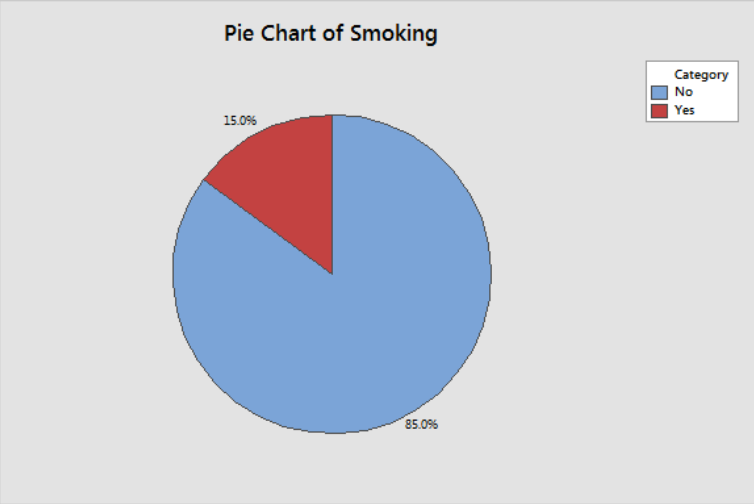
i. Is the P-value valid in this case? What assumptions are you making in order to carry out this test?

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 = 708, is large enough for the t-statistic to be valid.

Name:
E number:
Math 1530 section number

Bonus Problem: Question 4 on the SPRING 2016 Math 1530 asked, "Have you, yourself, ever smoked cigarettes in the past week?" Gallup reported that the national proportion of adults ages 18 to 29 that have smoked cigarettes in the past week is 20% (http://www.gallup.com/poll/187592/young-adults-cigarette-down-sharply.aspx?g_source=CATEGORY_WELLBEING&g_medium=topic&g_campaign=tiles). Note the results are similar for ages 30 to 64. Is the same true for the population of all U.S. college/university students?

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



b. How many of the students surveyed said "yes?" **106**

Tally for Discrete Variables: Smoking

Smoking	Count	Percent
No	602	85.03
Yes	106	14.97
N=	708	

c. What proportion of our sample said "yes?" **14.97%**

Tally for Discrete Variables: Smoking

Smoking	Count	Percent
No	602	85.03
Yes	106	14.97
N=	708	

Name:
E number:
Math 1530 section number

d. Assume (for the purpose of this problem) that we may treat the SPRING 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       106  708  0.149718  (0.124234, 0.178160)
```

Test and CI for One Proportion

```
Sample    X    N  Sample p          95% CI
1       106  708  0.149718  (0.123436, 0.175999)

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 12.42% and 17.82%.

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

Because the value 20% is not in the calculated confidence interval, our sample suggests that the proportion of US college/university students that chose "yes" to the survey question is NOT 20%.