Name:
E-number:
Section Number:

Math 1530 Capstone Project Solutions Fall 2014 100 points

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 the best option (or you may highlight in yellow if you are using a color printer).

a.	**CHILDREN**	Categorical	[✓ Quantitative]	Neither
b.	**PARTY**	[✓ Categorical]	Quantitative	Neither
c.	**DEATH_PENALTY**	[✓ Categorical]	Quantitative	Neither
d.	**TV**	Categorical	[✓ Quantitative]	Neither
e.	**SHOES**	Categorical	[✓ Quantitative]	Neither

Note: A categorical variable places an individual into one of several groups or categories. A quantitative variable takes numerical values for which arithmetic operations such as adding and averaging makes sense.

Name:
E-number:
Section Number:

2. **Sampling.** In the survey data, the variable "**SLEEP"** is the number of hours you sleep in a night.

a. Type in the first 10 observations from the variable "**SLEEP"** and use this as your sample data. Record the values in the table below.

n	1	2	3	4	5	6	7	8	9	10
SLEEP	**6**	**6**	**7**	**6**	**6**	**6**	**6**	**6**	**7**	**6**

Obtain the mean number of hours that one sleeps in a night for the first 10 observations.

The mean number of hours of sleep in a night is **6.2** hours.

Identify the type of sampling method you have just used: **convenience sampling**

b. Now, generate a random sample of size n=10 (Calc > Random Data > Sample from Columns). Enter 10 in the "Number of rows to Sample" box. Enter the variable "ID" and "**SLEEP**" into the "From columns" box. Enter C17-C18 into the "Store samples in" box. Record the information in the table below.

n	1	2	3	4	5	6	7	8	9	10
ID	**694**	**616**	**299**	**912**	**828**	**614**	**856**	**30**	**571**	**462**
SLEEP	**7**	**8**	**5**	**8**	**8**	**8**	**6**	**7**	**6**	**4**

Obtain the mean number of hours that one sleeps in a night for the first 10 observations.

The mean number of hours of sleep in a night is **6.7** hours.

Note: the solution above is not unique because of random sampling.

Identify the type of sampling method you have just used: **simple random sampling**

c. Obtain the mean number of hours that one sleeps in a night for all 1184 observations.

The population mean number of hours of sleep in a night is **6.5853** hours.

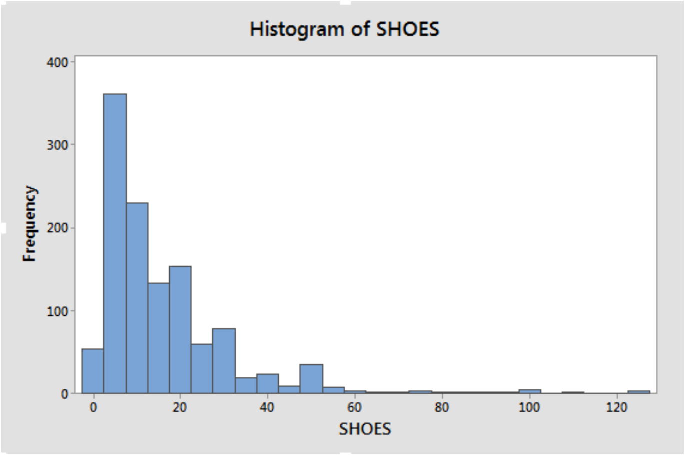
d. Compare the population mean you found in Part (c) to the means you found in Parts (a) and (b). Which sampling method provides a better estimate of the population mean number of hours one sleeps at night?

The mean hours of sleep in a night found by a SRS is closer to the population mean and provides a better estimate for the population mean. A statistic calculated from a biased sampling method such as convenience sampling tends to be biased. That is, it tends to underestimate or overestimate the true parameter.

Name:
E-number:
Section Number:

3. **If you are female then do this question.** **(Omit this page/problem if you are male.)** **SHOES.** Question 13 from the survey asked, “How many pairs of shoes do you own?”

a. Create an appropriate display for this variable and insert it here.



b. Which of the following best describes the shape of the distribution? Circle your answer.

Skewed left Symmetric [✓ Skewed right]

c. Calculate numerical measures appropriate for the shape of the distribution to describe the center and spread of **shoes**. Include appropriate output from Minitab here.

Descriptive Statistics: SHOES

```
Variable      N  N*    Mean  SE Mean   StDev  Minimum     Q1  Median      Q3  Maximum
SHOES      1184   0  16.203    0.464  15.952    0.000  5.000  11.000  20.000  125.000
```

Since the data is right skewed, the five-number summary should be used to describe the distribution.

i. Which statistic will you use to describe the center of the distribution? **median**

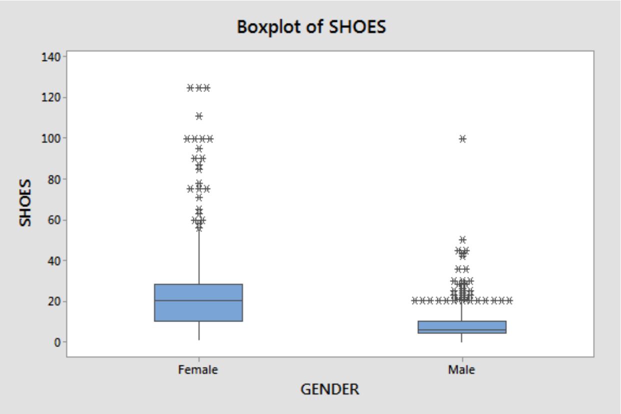
ii. What is the value of that statistic? **11**

iii. Which statistic(s) will you use to describe the spread of the distribution? **five-number summary**

iv. What is(are) the value(s) of that statistic? **min = 0, Q1 = 5, Median = 11, Q3 = 20, Max = 125**

Name:
E-number:
Section Number:

d. Create a side-by-side boxplot to compare the distributions of **shoes** for **genders**. Insert the graph below.



e. Describe the distributions of **shoes** for the two groups and compare them.

The distribution for "Females" and "Males" is right skewed with high outliers.

f. Are there any outliers in each group? Identify them and justify your answers.

The boxplot of the variable shows there are some outliers (*) in the two groups, "Female" and "Male". To verify this, first obtain the statistics for the two groups.

```
Variable   GENDER     N    N*     Mean  SE Mean    StDev  Minimum      Q1   Median       Q3  Maximum
 SHOES     Female   709    0    21.499    0.660   17.575    1.000  10.000   20.000   28.000  125.000
             Male   475    0     8.297    0.382    8.315    0.000   4.000    6.000   10.000  100.000
```

Then use the quartiles to compute the boundaries: Q1 - 1.5*IQR and Q3 + 1.5 *IQR as shown below.

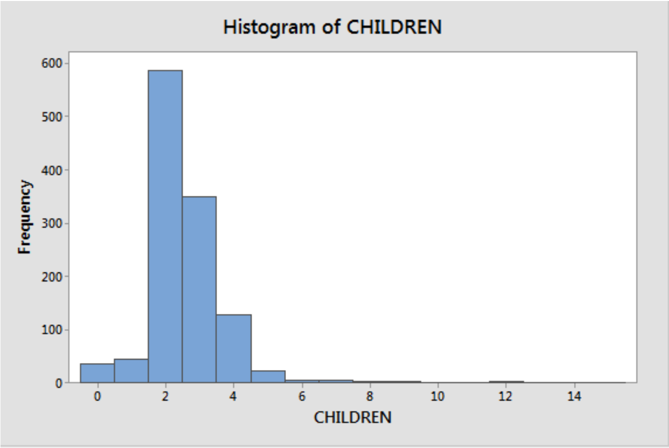
Gender	Q1	Q3	IQR	Q1 – 1.5*IQR	Q3 + 1.5*IQR
Female	10	28	18	-17	55
Male	4	10	6	-5	19

Any value below the lower fence or above the upper fence is an outlier.

Name:
E-number:
Section Number:

4. **If you are male then do this question.** **(Omit this page/problem if you are female.)** **CHILDREN.** Question 2 from the survey asked, “What do you think is the ideal number of children for a family to have?”

a. Create an appropriate display for this variable and insert it here.



b. Which of the following best describes the shape of the distribution? Circle your answer.

Skewed left Symmetric [✓ Skewed right]

c. Calculate numerical measures appropriate for the shape of the distribution to describe the center and spread of **children**. Include appropriate output from Minitab here.

Descriptive Statistics: CHILDREN

```
Variable     N  N*    Mean  SE Mean  StDev  Minimum      Q1  Median      Q3  Maximum
CHILDREN  1184   0  2.5752   0.0359  1.2360   0.0000  2.0000  2.0000  3.0000  15.0000
```

Since the data is right skewed, the five-number summary should be used to describe the distribution.

i. Which statistic will you use to describe the center of the distribution? **median**

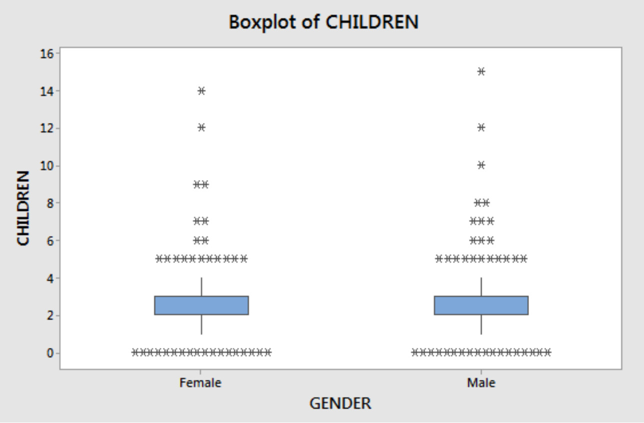
ii. What is the value of that statistic? **2**

iii. Which statistic(s) will you use to describe the spread of the distribution? **five-number summary**

iv. What is(are) the value(s) of that statistic? **min = 0, Q1 = 2, Median = 2, Q3 = 3, Max = 15**

Name:
E-number:
Section Number:

d. Create a side-by-side boxplot to compare the distributions of **children** for **gender**. Insert the graph below.



e. Describe the distributions of **children** for the two groups and compare them.

The distribution for "Females" and "Males" is right skewed with high outliers.

f. Are there any outliers in each group? Identify them and justify your answers.

The boxplot of the variable shows there are some outliers (*) in the two groups, "Female" and "Male". To verify this, first obtain the statistics for the two groups.

```
Variable  GENDER    N   N*    Mean  SE Mean  StDev  Minimum      Q1  Median      Q3  Maximum
CHILDREN  Female  709    0  2.6107   0.0431 1.1477   0.0000  2.0000  2.0000  3.0000  14.0000
            Male  475    0  2.5221   0.0623 1.3567   0.0000  2.0000  2.0000  3.0000  15.0000
```

Then use the quartiles to compute the boundaries: Q1 - 1.5*IQR and Q3 + 1.5 *IQR as shown below.

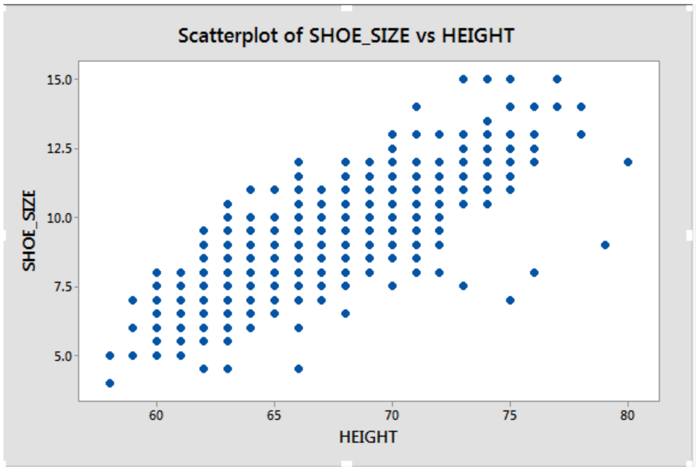
Gender	Q1	Q3	IQR	Q1 – 1.5*IQR	Q3 + 1.5*IQR
Female	2	3	1	0.5	4.5
Male	2	3	1	0.5	4.5

Any value below the lower fence or above the upper fence is an outlier.

Name:
E-number:
Section Number:

5. **Shoe size versus height.** Someone's shoe size depends on several variables and one of them could be someone's height. A previous MATH1530 class survey asked students to state their shoe size and height. Questions 14 and 15 asked students to input their U.S. shoe size (SHOE_SIZE) and their height (HEIGHT) in inches. Assume the respondents are an SRS of all ETSU students. We are interested in studying the relationship between a student's height and their shoe size. That is, we are interested in seeing whether knowing one's height can explain one's shoe size.

   a. Create an appropriate plot to display the relationship between **SHOE_SIZE** and **HEIGHT**. Insert the plot here.



   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.

   This plot shows a positive association between these two variables. As the height of a student increases, the shoe size increases.

   b. What is the correlation between the pair of variables? **r = 0.817**

   c. Obtain the least squares regression equation for the pair of variables. Insert it here.

   The regression equation is SHOE_SIZE = -16.946 + 0.39052 HEIGHT

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

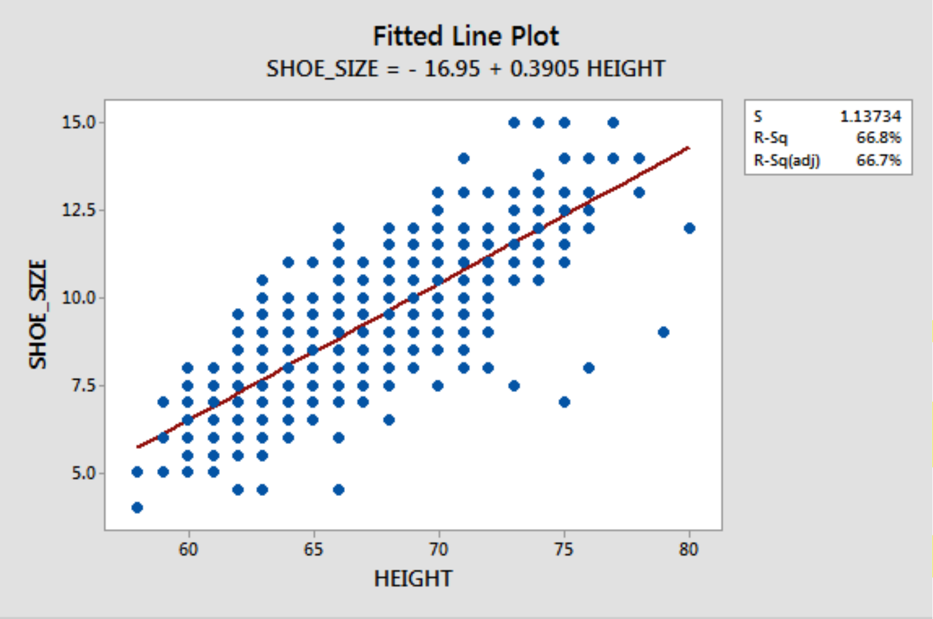
   As the height increases by 1 inch, the shoe size increases by 0.39052.

   e. Use the regression equation in part (c) to predict a student's shoe size for someone that is 73 inches tall.

   Estimated shoe size = -16.946 + 0.39052*73 = 11.56196 so approximately 11.5

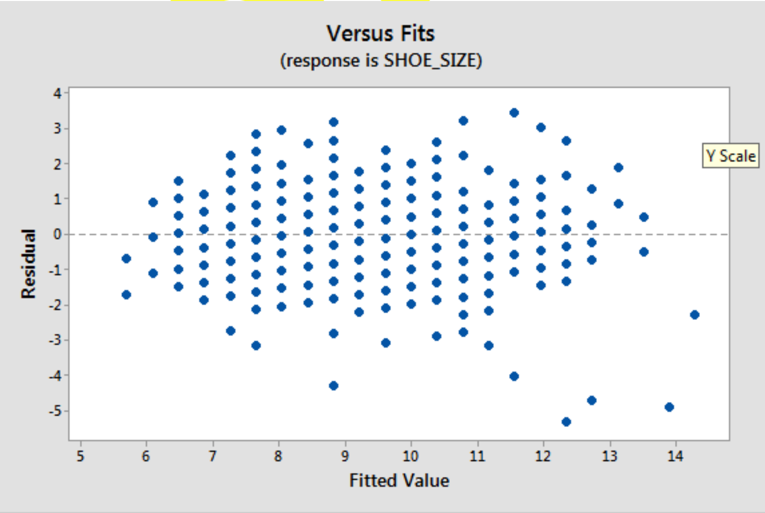
Name:
E-number:
Section Number:

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



The fitted line plot shows that the regression model fits the data moderately well. R^2 is useful in describing the linear association between X and Y. Here R^2 = 66.8%. Therefore, 66.8% of the variation in SHOE_SIZE can be explained by the linear relationship with HEIGHT.

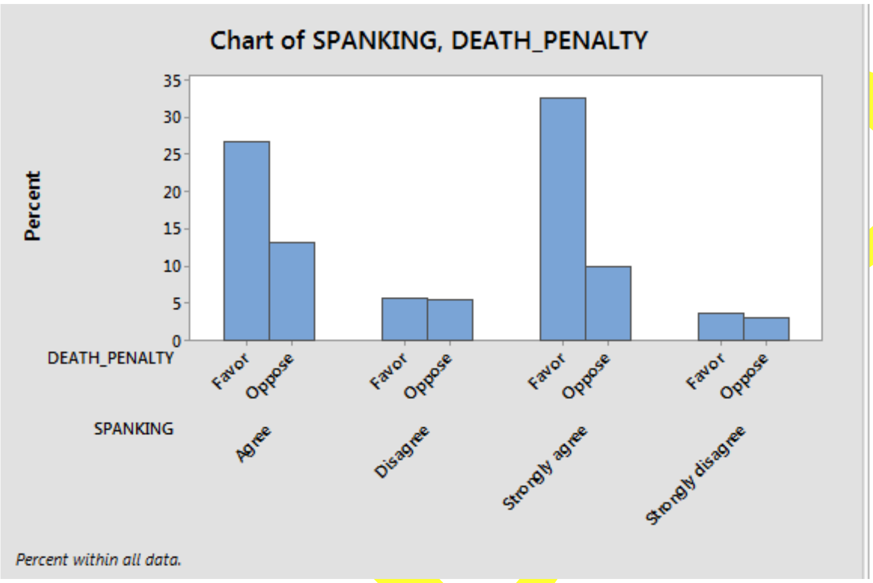
Note: Another scatterplot that is useful to see whether the model makes sense is the residual plot. This helps in determining the appropriateness of the regression model. Recall that the residuals are Residual = Observed Data – Predicted Data. The residual plot shouldn't have any interesting features,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



Name:
E-number:
Section Number:

6. **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.)** **SPANKING AND THE DEATH PENALTY** Question 7 from the survey asked students "Do you strongly agree, agree, disagree, or strongly disagree that it is sometimes necessary to discipline a child with a good, hand spanking?" and Question 8 from the survey asked students "Do you favor or oppose the death penalty for persons convicted of murder?" We want to check if there is a relationship between the belief about spanking and the belief about the death penalty. Assume the students who took the class survey are from an SRS of ETSU students.

   a. Create an appropriate graph to display the data and insert it here.
   Use Graph ➔ Bar graph ➔ Cluster or Graph ➔ Bar graph ➔ Stack to create a bar graph. Below is the cluster bar graph



   It is important to use percentages to compare when the number of individuals in the groups is not the same. If the number of individuals in the groups is the same, then comparing counts and comparing percentages are comparable. However, percents are preferred over counts.

   b. Create an appropriate two-way table to summarize the data and insert it here.

   Use Stat ➔ Tables ➔ Cross Tabulation and Chi-square.

Tabulated Statistics: SPANKING, DEATH_PENALTY

```
Rows: SPANKING   Columns: DEATH_PENALTY

                     Favor  Oppose   All

Agree                  316     155   471
Disagree                68      64   132
Strongly agree         385     117   502
Strongly disagree       42      37    79
All                    811     373  1184

Cell Contents:      Count
```

Name:
E-number:
Section Number:

SUPPOSE WE SELECT ONE STUDENT AT RANDOM:

c. Find the probability that the student strongly believes in spanking a child and favors the death penalty for persons convicted of murder.

385/1184 = 0.3252 = 32.52%

d. Find the probability that a student strongly disagrees in spanking a child or they oppose the death penalty for persons convicted of murder.

79/1184 + 373/1184 – 37/1184 = 415/1184 = 0.3505 = 35.05%

e. Find the probability that a student agrees in spanking a child given they favor the death penalty for persons convicted of murder.

316/811 = 0.3896 = 38.96%

701/811 = 0.8644 = 86.44% (In case they combine agree and strongly agree)

f. Find the probability that a student favors the death penalty for persons convicted of murder given they agree in spanking a child.

316/471 = 0.6709 = 67.09%

701/973 = 0.7205 = 72.05% (In case they combine agree and strongly agree)

g. **BONUS**: Carry out a test for the hypothesis that there is no relationship between the belief about spanking a child and the belief about the death penalty of ETSU students. Use a significance level of α = 0.05.

i. State the null and alternative hypothesis.

H_0: There is no relationship between the belief about spanking a child and the belief about the death penalty
H_A: There is a relationship between the belief about spanking a child and the belief about the death penalty

Name:
E-number:
Section Number:

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

Tabulated Statistics: SPANKING, DEATH_PENALTY

```
      Rows: SPANKING   Columns: DEATH_PENALTY

                            Favor  Oppose   All
             Agree            316     155   471
                           322.62  148.38

          Disagree             68      64   132
                            90.42   41.58

    Strongly agree            385     117   502
                           343.85  158.15

 Strongly disagree             42      37    79
                            54.11   24.89

               All            811     373  1184

    Cell Contents:          Count
                            Expected count


     Pearson Chi-Square = 42.307, DF = 3, P-Value = 0.000
```

iii. Which test statistic are you using and what is its value?

A chi-square test statistic and its value is 42.307

iv. State your decision and conclusion for the test.

Based on the chi-square test results, the p-value is 0.000. Therefore, we reject the null hypothesis and conclude there is a significant relationship between the belief about spanking a child and the belief about the death penalty.

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 1 or greater.**
All the expected counts are greater than 5.

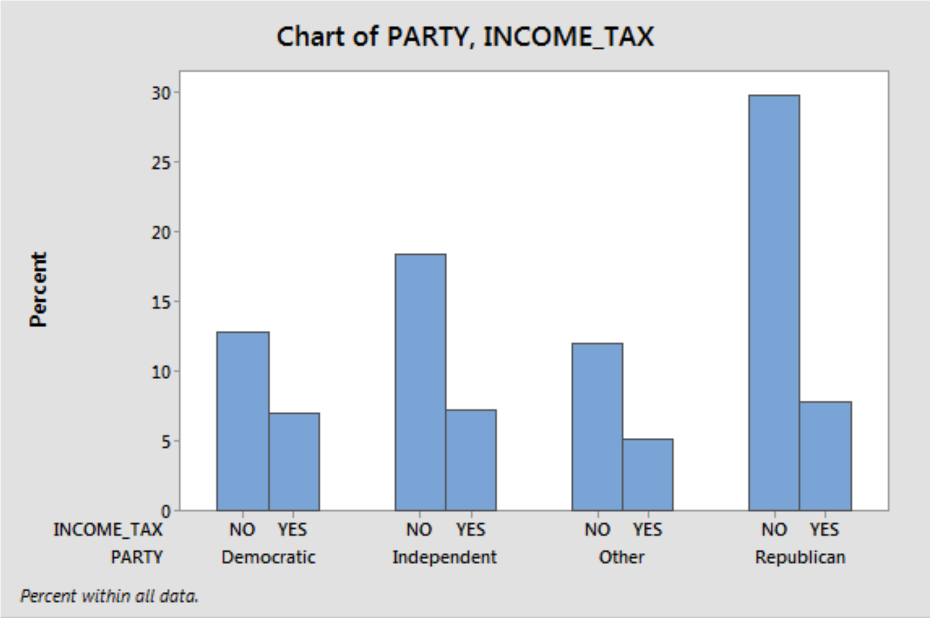
* **The data is a random sample from the population.** Here the problems states to assume the students who took the class survey are from an SRS of ETSU students.

Note: The original data were obtained from part of the population (current ETSU students.) It was from a voluntary response sampling. We selected a SRS from this part of the population**. If we feel that the students who took the survey represent the population then we can trust the above conclusions.**

Name:
E-number:
Section Number:

7. **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.)** **POLITICAL PARTY AND INCOME TAX** Question 5 from the survey asked students "What political party do you identify with?" and Question 6 from the survey asked students "Should Tennessee implement a state income tax?" We want to check if there is a relationship between political party and the belief about an income tax for Tennessee. Assume the students who took the class survey are from an SRS of ETSU students.

   a. Create an appropriate graph to display the data and insert it here.
   Use Graph ➔ Bar graph ➔ Cluster or Graph ➔ Bar graph ➔ Stack to create a bar graph. Below is the cluster bar graph



It is important to use percentages to compare when the number of individuals in the groups is not the same. If the number of individuals in the groups is the same, then comparing counts and comparing percentages are comparable. However, percents are preferred over counts.

   b. Create an appropriate two-way table to summarize the data and insert it here.

   Use Stat ➔ Tables ➔ Cross Tabulation and Chi-square.

Tabulated Statistics: PARTY, INCOME_TAX

```
Rows: PARTY   Columns: INCOME_TAX

              NO  YES   All

Democratic   152   82   234
Independent  218   85   303
Other        142   61   203
Republican   352   92   444
All          864  320  1184

Cell Contents:      Count
```

Name:
E-number:
Section Number:

SUPPOSE WE SELECT ONE STUDENT AT RANDOM:

c. Find the probability that the student identifies with the Republican Party and agrees Tennessee should implement an income tax.

92/1184 = 0.0777 = 7.77%

d. Find the probability that a student identifies with the Independent Party or they disagree Tennessee should implement an income tax.

303/1184 + 864/1184 – 218/1184 = 949/1184 = 0.8015 = 80.15%

e. Find the probability that a student identifies with the Democratic Party given they disagree Tennessee should implement an income tax.

152/864 = 0.1759 = 17.59%

f. Find the probability that a student disagrees Tennessee should implement an income tax given they identify with the Democratic Party.

152/234 = 0.6496 = 64.96%

g. **BONUS**: Carry out a test for the hypothesis that there is no relationship between one's political party and the belief about an income tax in Tennessee. Use a significance level of α = 0.05.

i. State the null and alternative hypothesis.

H_0: There is no relationship between one's political party and the belief about an income tax in Tennessee
H_A: There is a relationship between one's political party and the belief about an income tax in Tennessee

Name:
E-number:
Section Number:

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

Tabulated Statistics: PARTY, INCOME_TAX

```
Rows: PARTY   Columns: INCOME_TAX

                   NO    YES   All

Democratic        152     82   234
                170.8   63.2

Independent       218     85   303
                221.1   81.9

Other             142     61   203
                148.1   54.9

Republican        352     92   444
                324.0  120.0

All               864    320  1184

Cell Contents:       Count
                     Expected count


Pearson Chi-Square = 17.678, DF = 3, P-Value = 0.001
```

iii. Which test statistic are you using and what is its value?

A chi-square test statistic and its value is 17.678

iv. State your decision and conclusion for the test.

Based on the chi-square test results, the p-value is 0.001. Therefore, we reject the null hypothesis and conclude there is a significant relationship between one's political party and the belief about an income tax in Tennessee.

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 1 or greater.**
All the expected counts are greater than 5.

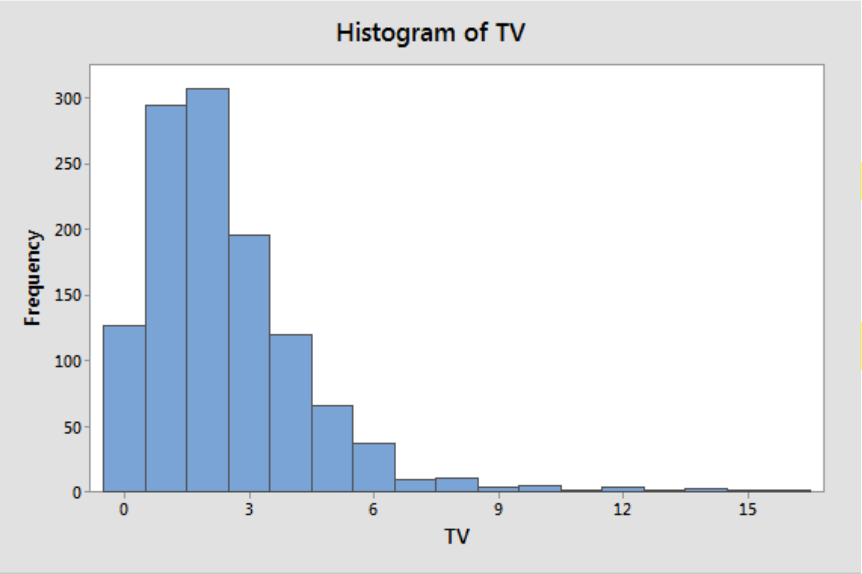
*** The data is a random sample from the population.** Here the problems states to assume the students who took the class survey are from an SRS of ETSU students.

Note: The original data were obtained from part of the population (current ETSU students.) It was from a voluntary response sampling. We selected a SRS from this part of the population**. If we feel that the students who took the survey represent the population then we can trust the above conclusions.**

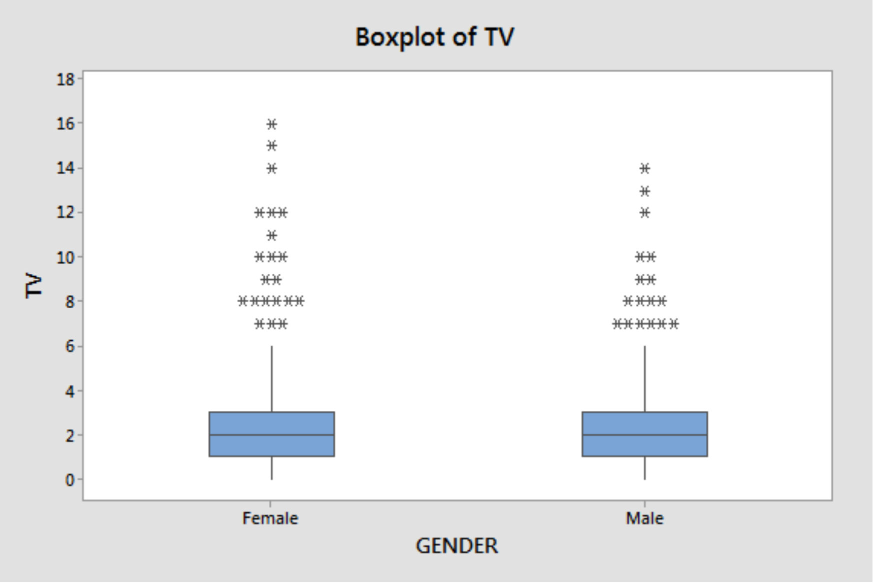
Name:
E-number:
Section Number:

8. **TV.** A marketing analyst for a cable provider wishes to know if males or females watch more TV in order for the company's advertisements to target that specific gender. After talking to the company's current sales representatives across the US, he concludes that males watch more TV. Questions 9 from the survey asked students "How many hours do you personally watch television including Netflix, Hulu Plus, Amazon Prime, etc... in a day?" Assume that the students who responded the survey are a SRS of all ETSU students. Is there good evidence to support the idea that male students at ETSU watch more TV, on average, than female students?

a. Create an appropriate graph to display the distribution of number of hours of TV watched in day and insert it here.



You may have chosen to do a side-by-side boxplot with the grouping variable gender.



Name:
E-number:
Section Number:

b. Use Minitab to calculate a 95% confidence interval for the difference in the mean number of hours of TV watched between male and female students. Interpret the confidence interval.

Use Stat ➔ Basics Statistics ➔ 2 sample t

```
Two-sample T for TV

GENDER     N  Mean  StDev  SE Mean
Female   709  2.43   2.04    0.077
Male     475  2.45   1.97    0.090


Difference = μ (Female) - μ (Male)
Estimate for difference:  -0.022
95% CI for difference:  (-0.255, 0.210)
```

We are 95% confident that the true difference in the means of the number of hours of TV watched in a day for females and for male students is between -0.255 and 0.210 hours.

c. Perform an appropriate hypothesis test and include the output from Minitab here.

Use Stat ➔ Basics Statistics ➔ 2 sample t

(Here we are assuming μ_1 = population mean for females and μ_2 = population mean for males)

```
Two-sample T for TV

GENDER     N  Mean  StDev  SE Mean
Female   709  2.43   2.04    0.077
Male     475  2.45   1.97    0.090


Difference = μ (Female) - μ (Male)
Estimate for difference:  -0.022
95% upper bound for difference:  0.173
T-Test of difference = 0 (vs <): T-Value = -0.19  P-Value = 0.425  DF = 1039
```

H_0: $\mu_1 - \mu_2 = 0$ **versus** **H_a:** $\mu_1 - \mu_2 < 0$

d. What is the value of the test statistic? **t = -0.19**

e. What is the P-value for this test? **p-value = 0.425**

f. State your decision and conclusion for the test using a significance level of α = 0.05

Since the p-value is greater than α, we conclude that there is not enough statistical evidence to claim that ETSU male students watch more TV in a day on average than female students.

g. What assumptions are we making about the samples for our interpretation to be valid?

We assume that the data presented is a representative/random sample of all ETSU students, as well as separately for female and male students.