Name:
E-number:
Section Number:

Math 1530 Capstone Project **SOLUTIONS** Spring 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.	**AREA_CODE**	[✓ Categorical]	Quantitative	Neither
b.	**HOGWARTS**	[✓ Categorical]	Quantitative	Neither
c.	**FIRST_PHONE**	Categorical	[✓ Quantitative]	Neither
d.	**DAYS**	Categorical	[✓ Quantitative]	Neither
e.	**COMPANION**	[✓ 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 "**FIRST_PHONE"** is the age that one received their first cell phone.

a. Type in the first 10 observations from the variable "**FIRST_PHONE"** 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
FIRST_PHONE	11	11	23	15	15	16	21	13	16	12

Obtain the mean age that one received their first cell phone for the first 10 observations.

The mean age is **15.30** years.

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 "**FIRST_PHONE**" into the "From columns" box. Enter C18-C19 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	488	525	548	65	474	448	171	373	495	86
FIRST_PHONE	14	13	23	13	12	17	14	15	15	13

Obtain the mean age that one received their first cell phone of your random sample.

The mean age is **14.90** years.

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 age that one received their first cell phone for all 610 observations.

The population mean age is **14.616** years.

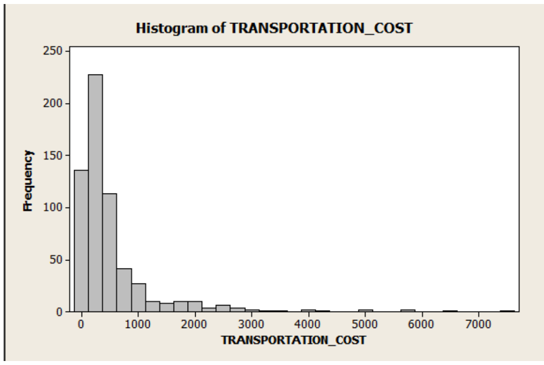
d. Compare the population mean you found in Part (c) to the means you found in Parts (a) and (b). Which sampling method provides a better estimate of the population mean age?

The mean age found by SRS is closer to the population mean and provides at 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.) TRANSPORTATION_COST.** Question 24 from the survey asked, “What is the cost of the air tickets(s), fares, or estimated cost of gas to arrive at the port of departure?”

   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 **transportation cost**. Include appropriate output from Minitab here.

Descriptive Statistics: TRANSPORTATION_COST

```
Variable             Minimum     Q1   Median      Q3   Maximum    IQR
TRANSPORTATION_COST     20.0  134.0    300.0   516.3    7430.0  382.3
```

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? **300**

iii. Which statistic(s) will you use to describe the spread of the distribution?

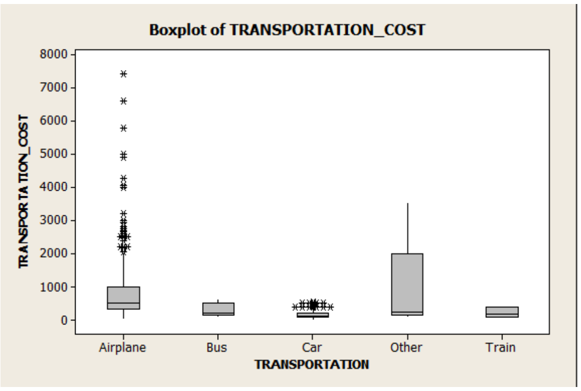
five-number summary

iv. What is(are) the value(s) of that statistic?

min = 20, Q1 = 134, Median = 300, Q3 = 516. 3, Max = 7430

Name:
E-number:
Section Number:

d. Create a side-by-side boxplot to compare the distributions of **transportation cost** for different modes of **transportation**. Insert the graph below.



e. Describe the distributions of **transportation cost** for the five groups and compare them.

The distribution is roughly symmetric (or slightly right skewed) for "Bus" and "train." The distribution is for "Airplane" and "Car" is right skewed with high outliers. The distribution for "Other" is right skewed with no apparent outliers.

f. Are there any outliers in each group? Identify them and justify your answers.
The boxplot of the variable shows that there are some outliers (*) for the group "Airplane" and "Car". To verify this, first obtain the statistics for the five groups.

Descriptive Statistics: TRANSPORTATION_COST

```
Variable            TRANSPORTATION     Mean   StDev  Minimum     Q1  Median
TRANSPORTATION_COST Airplane          853.2   993.9     50.0  319.5   500.0
                    Bus               295.9   191.5    113.0  150.0   200.0
                    Car              166.84  107.51    20.00  94.25  129.00
                    Other              1017    1288      120    148     254
                    Train             229.4   159.0     80.0   90.0   167.0

Variable            TRANSPORTATION       Q3  Maximum     IQR
TRANSPORTATION_COST Airplane         1000.0   7430.0   680.5
                    Bus               500.0    600.0   350.0
                    Car              206.00   535.00  111.75
                    Other              2000     3500    1852
                    Train             400.0    400.0   310.0
```

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

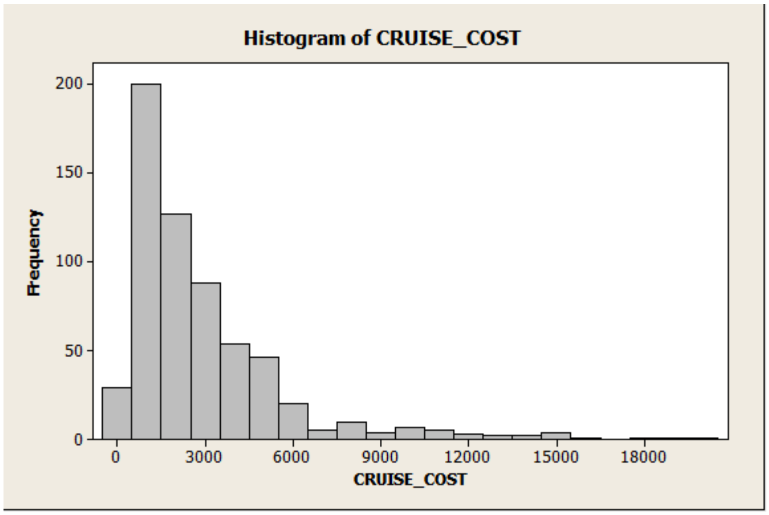
Transportation Mode	Q1	Q3	IQR	Q1 – 1.5*IQR	Q3 + 1.5*IQR
Airplane	319.5	1000	680.5	-701.25	2020.75
Bus	150	500	350	-375	1025
Car	94.25	206	111.75	-73.375	373.625
Other	148	2000	1852	-2630	4778
Train	90	400	310	-375	865

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.)** **CRUISE_COST.** Question 31 from the survey asked, "What is the total cost of the cruise for your entire party, not including shore excursions and travel to port of departure?"

   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 **cruise cost**. Include appropriate output from Minitab here.

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

Descriptive Statistics: CRUISE_COST

```
  Variable      Minimum     Q1    Median      Q3    Maximum    IQR
CRUISE_COST         143   1019      2000    3599      20464   2580
```

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

   ii. What is the value of that statistic? **2000**

   iii. Which statistic(s) will you use to describe the spread of the distribution?

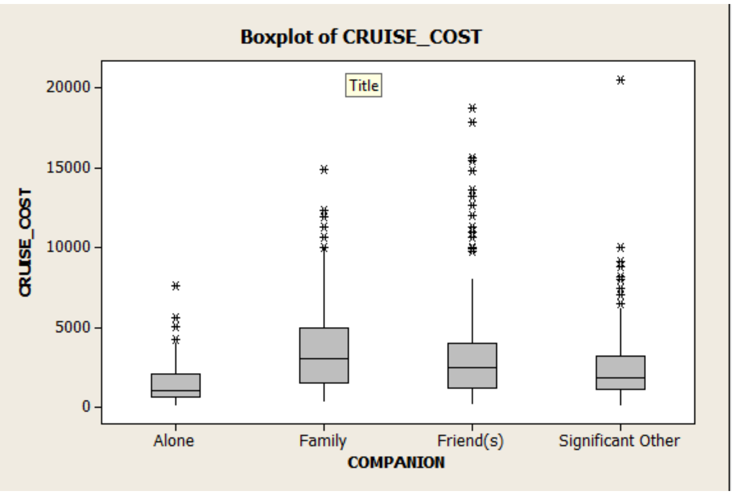
   five-number summary

   iv. What is(are) the value(s) of that statistic?

   min = 143, Q1 = 1019, Median = 2000, Q3 = 3599, Max = 20464

Name:
E-number:
Section Number:

d. Create a side-by-side boxplot to compare the distributions of **cruise cost** for **companion**. Insert the graph below.



e. Describe the distributions of **cruise cost** for the four groups and compare them.

Each distribution is right skewed with outliers.

f. Are there any outliers in each group? Justify your answer.

The boxplot of the variable shows that there are some outlier (*) in each group. To verify this, first obtain the statistics for the four groups.

Descriptive Statistics: CRUISE_COST

```
Variable      COMPANION           Minimum     Q1  Median     Q3  Maximum   IQR
CRUISE_COST   Alone                   159    600    1000   2026     7548  1426
              Family                  359   1499    3000   4899    14893  3400
              Friend(s)               200   1200    2450   4000    18706  2800
              Significant Other       143   1085    1800   3171    20464  2086
```

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

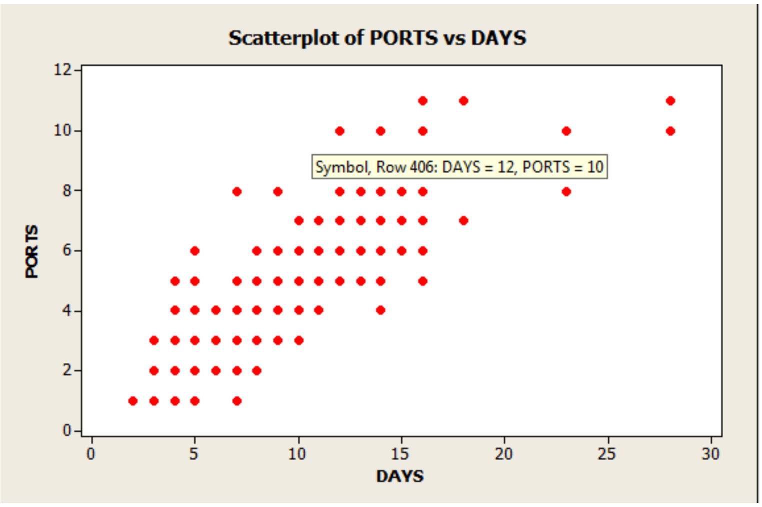
Transportation Mode	**Q1**	**Q3**	**IQR**	**Q1 – 1.5*IQR**	**Q3 + 1.5*IQR**
Alone	600	2026	1426	-1539	4165
Family	1499	4899	3400	-3601	9999
Friend(s)	1200	4000	2800	-3000	8200
Significant Other	1085	3171	2086	-2044	6300

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

Name:
E-number:
Section Number:

5. **Number of Ports versus Length of Cruise.** The number of ports visited during a sea cruise depends on several variables and one of them could be the length of the cruise. The MATH1530 class survey asked students to plan a sea cruise for the summer of 2014. Questions 29 and 33 asked students to input the number of days that the cruise would last (DAYS) and the number of ports that are included on the cruise ship's itinerary (PORTS). Assume the respondents are an SRS of all ETSU students. We are interested in studying the relationship between the days of the cruise and the number of ports visited and whether knowing the length of the cruise would explain the number of ports visited on the cruise.

a. Create an appropriate plot to display the relationship between **DAYS** and **PORTS**. 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 length of the cruise increases, the number of ports visited increases.

b. What is the correlation between the pair of variables? **r = 0.839**

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

The regression equation is PORTS = 0.468 + 0.437 DAYS

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

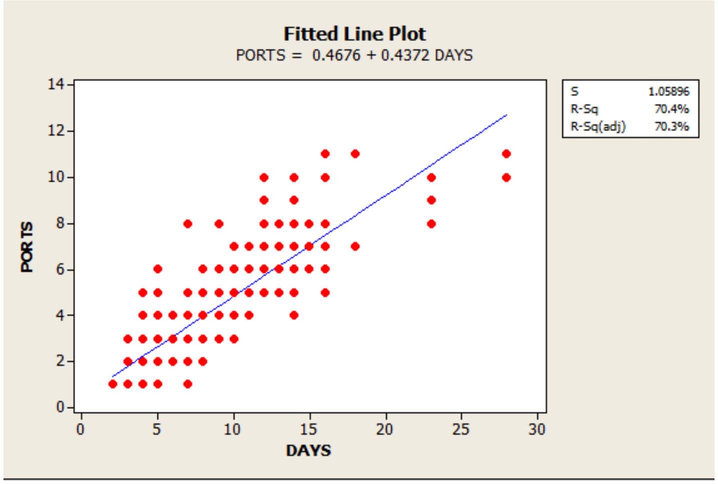
As the length of the cruise increases by 1 day, the number of ports visited increases by 0.437 ports.

e. Use the regression equation in part (c) to predict the number of ports that will be visited if a student would like to take a 14 day cruise.

Estimated number of ports = 0.468 + 0.437*14 = 6.586 ports

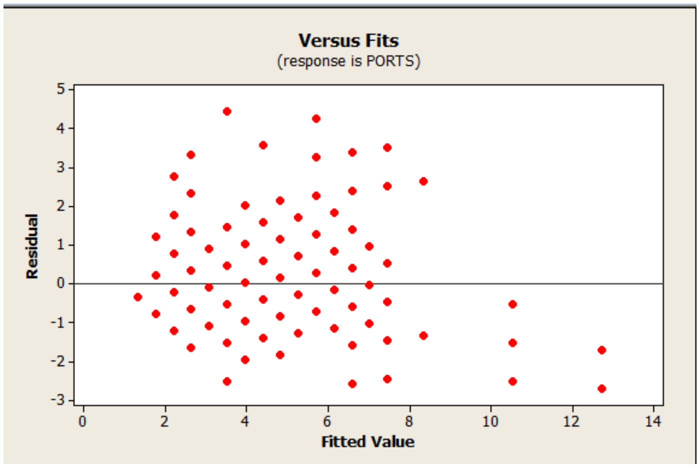
Name:
E-number:
Section Number:

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 fairly well. R^2 is useful in describing the linear association between X and Y. Here R_2 = 70.4%. Therefore, 70.4% of the variation in PORTS can be explained by the linear relationship with DAYS.

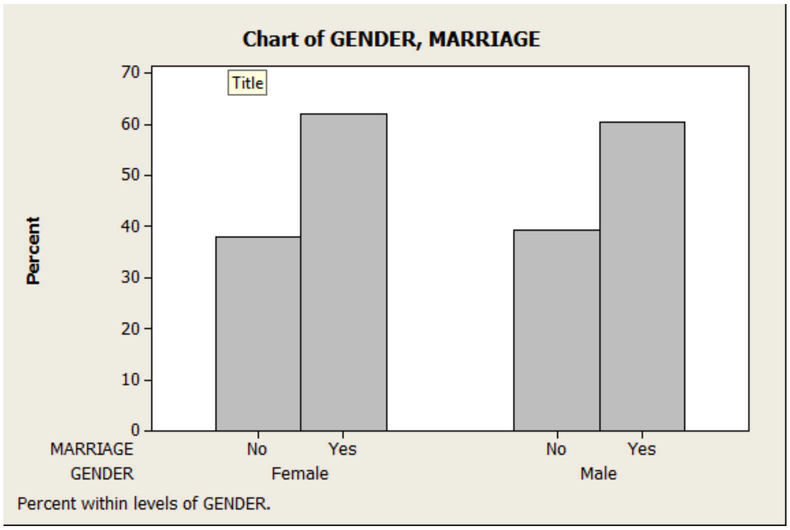
Note: Another scatterplot that is useful to see whether the model makes sense is the residual plot. This helps in determining the appropriateness of the regression model. Recall that the residuals are Residual = Observed Data – Predicted Data. The residual plot shouldn't have any interesting features, 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.) Do males and females believe that same-sex marriage should be legal in the USA?** Question 19 from the survey asked students "Do you think same-sex marriage should be legal in the USA?" We want to check if there is a relationship between gender and the belief about same-sex marriage in the USA. 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



   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: GENDER, MARRIAGE

```
Rows: GENDER   Columns: MARRIAGE

           No  Yes  All

Female    141  231  372
Male       94  144  238
All       235  375  610

Cell Contents:      Count
```

Name:
E-number:
Section Number:

SUPPOSE WE SELECT ONE STUDENT AT RANDOM:

c. Find the probability that the student is a male and believes that same-sex marriage should be legal in the USA.

144/610 = 0.236 = 23.6%

d. Find the probability that a student is female or they believe that same-sex marriage should be legal in the USA.

372/610 + 375/610 – 231/610 = 516/610 = 0.8459 = 84.59%

e. Find the probability that a student does not believe that same-sex marriage should be legal in the USA given that the student was a female.

141/372 = 0.379 = 37.9%

f. Find the probability that a student was a female given that the student does not believe that same-sex marriage should be legal in the USA.

141/235 = 0.6 = 60%

Carry out a test for the hypothesis that there is no relationship between gender and the belief about same-sex marriage in the USA 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 gender and belief of same-sex marriage
H_A: There is a relationship between gender and belief of same-sex marriage

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

Tabulated statistics: GENDER, MARRIAGE

```
Rows: GENDER   Columns: MARRIAGE

             No    Yes    All

Female      141    231    372
          143.3  228.7  372.0

Male         94    144    238
           91.7  146.3  238.0

All         235    375    610
          235.0  375.0  610.0

Cell Contents:      Count
                    Expected count
```

```
Pearson Chi-Square = 0.155, DF = 1, P-Value = 0.693
```

Name:
E-number:
Section Number:

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

A chi-square test statistic and its value is 0.155

iv. State your decision and conclusion for the test.

Based on the chi-square test results, the p-value is 0.693. Therefore, we fail to reject the null hypothesis and conclude there is not a significant relationship between gender and belief of same-sex marriage.

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.**

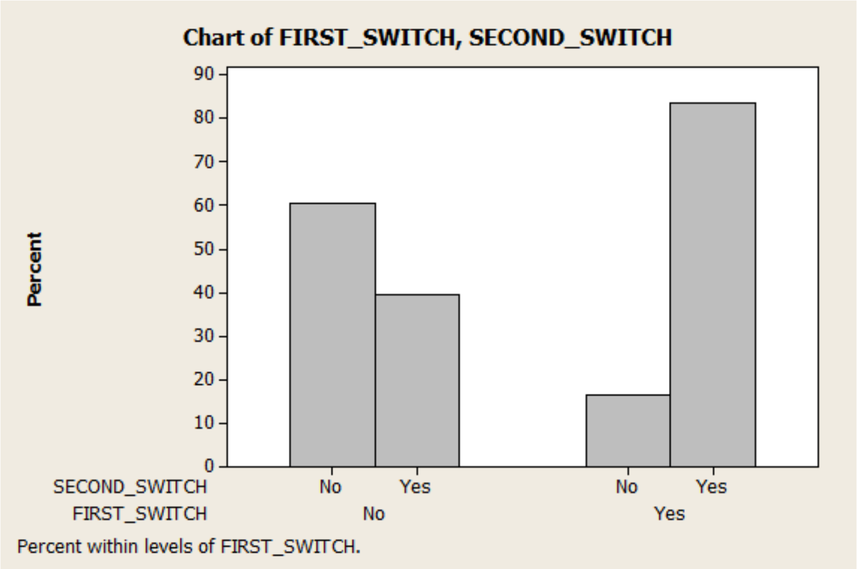
Name:
E-number:
Section Number:

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.)** **Overpopulation.** Overpopulation is one of the world's impending problems. In the new Dan Brown (author of the Da Vinci Code) book *Inferno*, the following question was asked: "What if you were told that if you flipped a switch, you would kill half the world's population?"

   Question 20 from the survey asked students "Would you flip the switch?" **Question 21** from the survey asked students ""What if you were told that if you didn't flip the switch right then, the human race would be extinct in the next hundred years. Would you flip the switch then?" We want to check if there is a relationship between the choice of the flipping the switch based on question 20 and the choice of flipping the switch based on the secondary question 21. 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



   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: FIRST_SWITCH, SECOND_SWITCH

```
Rows: FIRST_SWITCH   Columns: SECOND_SWITCH

         No  Yes  All

No      317  208  525
Yes      14   71   85
All     331  279  610

Cell Contents:      Count
```

Name:
E-number:
Section Number:

SUPPOSE WE SELECT ONE STUDENT AT RANDOM:

c. Find the probability that the student chose to flip the switch based on question 20 and chose to flip the switch based on question 21.

   71/610 = 0.1164 = 11.64%

d. Find the probability that a student does not choose to flip the switch based on question 20 or does not flip the switch based on question 21.

   525/610 + 331/610 – 317/610 = 539/610 = 0.8836 = 88.36%

e. Find the probability that a student chooses to flip the switch based on question 20 given the chose to flip the switch based on question 21.

   71/279 = 0.2545 = 25.45%

f. Find the probability that a student chooses to flip the switch based on question 21 given the chose to flip the switch based on question 20.

   71/85 = 0.8353 = 83.53%

g. Carry out a test for the hypothesis that there is no relationship between the choice of the flipping the switch based on question 20 and the choice of flipping the switch based on the secondary question 21.ETSU students. Use α=0.05.

   i. State the null and alternative hypothesis.

   H_0: There is no relationship between choice of flipping the switch based on question 20 and the choice of flipping the switch based on the secondary question 21.

   H_a: There is a relationship between choice of flipping the switch based on question 20 and the choice of flipping the switch based on the secondary question 21.

Name:
E-number:
Section Number:

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

Tabulated statistics: FIRST_SWITCH, SECOND_SWITCH

```
Rows: FIRST_SWITCH   Columns: SECOND_SWITCH

            No    Yes    All

No         317    208    525
         284.9  240.1  525.0

Yes         14     71     85
          46.1   38.9   85.0

All        331    279    610
         331.0  279.0  610.0

Cell Contents:      Count
                    Expected count


Pearson Chi-Square = 56.834, DF = 1, 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 56.834

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 choice of flipping the switch based on question 20 and the choice of flipping the switch based on the secondary question 21.

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.

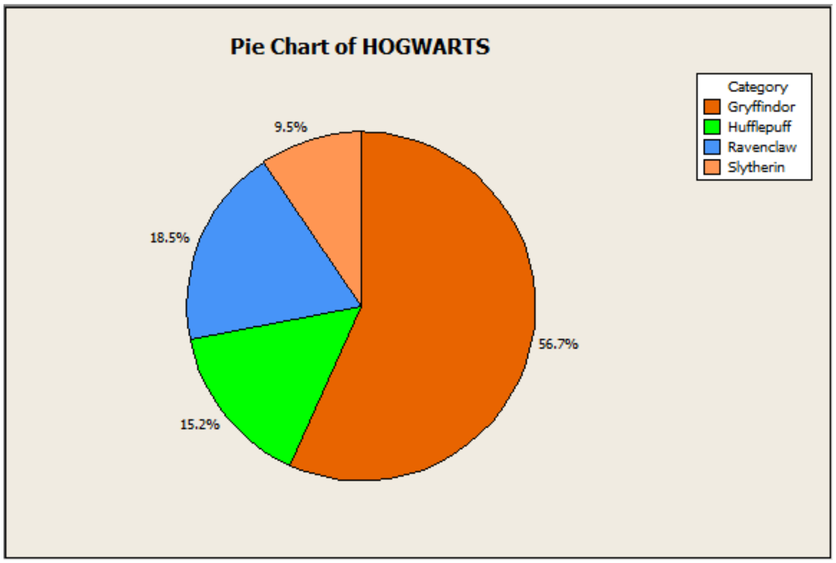
* **The data is a random sample from the population.** Here the problems states to assume the students who took the class survey are from an SRS of ETSU students.

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. **Hogwarts School of Witchcraft and Wizardry.** Question 17 from the survey asked “If you could ask the Sorting Hat to place you in any House at Hogwarts, which House would you choose?” We suspect that the proportion of students that choose to be placed the Ravenclaw House at Hogwarts is different than 25%. Assume the students who took the class survey are from an SRS of ETSU students. Use the survey data to conduct a confidence interval and perform a hypothesis test. Answer the following questions.

a. Create pie graph of the variable Hogwarts and insert it here.



b. Based on our sample data, use Minitab to calculate a 90% confidence interval for the proportion of all ETSU students who would choose to be placed in Ravenclaw House at Hogwarts.
Use Stat ➔ Basics Statistics ➔ 1 Proportion

Test and CI for One Proportion

```
Sample     X     N   Sample p          90% CI
1        113   610   0.185246   (0.159373, 0.211119)

Using the normal approximation.
```

c. State the proper interpretation of the confidence interval you obtained in part (a).

We are 90% confident that the true proportion of all ETSU students who would choose be placed in Ravenclaw House at Hogwarts is between 15.94 and 21.11 percent.

d. Perform an appropriate hypothesis test to determine if the proportion of ETSU students that would choose Ravenclaw is not 25%. State the null and alternative hypotheses.

H_0: p = 0.25 vs. H_a: p $\neq$ 0.25

Name:
E-number:
Section Number:

e. What is the sample proportion of the respondents who chose Ravenclaw?

Use Stat ➔ Tables ➔ Tally Individual Variables

Tally for Discrete Variables: HOGWARTS

```
  HOGWARTS  Count  Percent
Gryffindor    346    56.72
Hufflepuff     93    15.25
 Ravenclaw    113    18.52
 Slytherin     58     9.51
        N=    610
```

The proportion is 0.1852 or 18.52%.

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

Use Stat ➔ Basics Statistics ➔ 1 Proportion

Test and CI for One Proportion

```
Test of p = 0.25 vs p not = 0.25


Sample    X    N  Sample p         90% CI            Z-Value  P-Value
1       113  610  0.185246  (0.159373, 0.211119)    -3.69    0.000

Using the normal approximation.
```

g. Which test statistic are you using? Report its value.

The z-test statistic is used and the value is -3.69.

h. What is the P-value for the test?

The p-value is 0.000

i. State your decision and conclusion using a significance level of α=0.10.

Since the p-value is less than α, we conclude that there is enough evidence that the proportion of ETSU students that would choose Ravenclaw is not 25%.

Name:
E-number:
Section Number:

j. Examine the data. What are the conditions for inference for this test? Are they met in this case? Explain.

Conditions for inference about a population using a large sample test:

* **The data is a random sample from the population.** 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.**

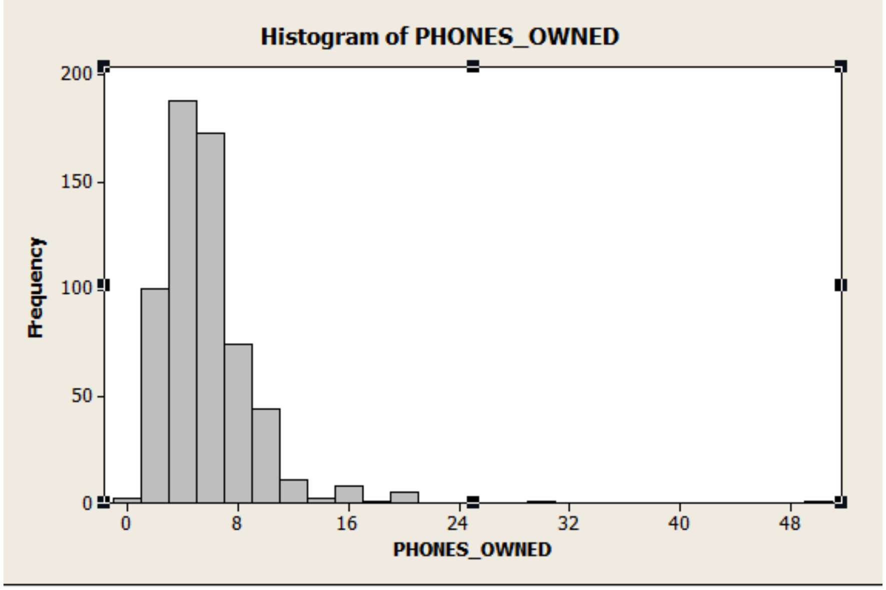
* **The population is at least 10 times as large as the sample. This is true. The sample size is 610.**

* **The sample size n is so larger that both np_o and $n(1 – p_0)$ are 10 or more. Here $n = 610$, $p_0 = 0.25$. Hence we have $np_0 = 152.5$ and $n(1 – p_0) = 475.5$.**

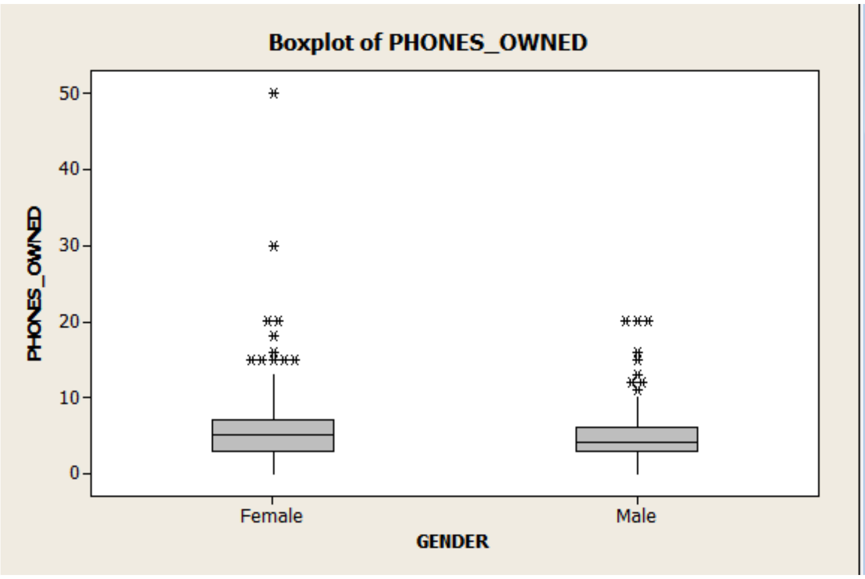
Name:
E-number:
Section Number:

9. **Number of Cell Phones Owned.** A marketing analyst wishes to know if males or females buy more cell phones 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 females buy more cell phones and thus, have owned more cell phones in their lifetime compared to males. Questions 7 from the survey asked students "How many cell phones have you owned?" Assume that the students who responded the survey are a SRS of all ETSU students. Is there good evidence to support the idea that female students at ETSU have owned more cell phones, on average, than male students?

a. Create an appropriate graph to display the distribution of number of cell phones owned 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 cell phones owned between female and male students. Interpret the confidence interval.

Use Stat ➔ Basics Statistics ➔ 2 sample t

Two-Sample T-Test and CI: PHONES_OWNED, GENDER

```
Two-sample T for PHONES_OWNED

GENDER    N  Mean  StDev  SE Mean
Female  372  5.49   4.11     0.21
Male    238  4.75   3.23     0.21


Difference = mu (Female) - mu (Male)
Estimate for difference:  0.737
95% CI for difference:  (0.150, 1.324)
T-Test of difference = 0 (vs not =): T-Value = 2.47  P-Value = 0.014  DF = 582
```

We are 95% confident that the true mean difference in the number of cell phones owned between female and male students is between 0.150 and 1.324 phones.

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

Use Stat ➔ Basics Statistics ➔ 2 sample t

H$_0$: $\mu_1 - \mu_2 = 0$ **versus** **H$_a$:** $\mu_1 - \mu_2 > 0$

Two-Sample T-Test and CI: PHONES_OWNED, GENDER

```
Two-sample T for PHONES_OWNED

GENDER    N  Mean  StDev  SE Mean
Female  372  5.49   4.11     0.21
Male    238  4.75   3.23     0.21


Difference = mu (Female) - mu (Male)
Estimate for difference:  0.737
95% lower bound for difference:  0.245
T-Test of difference = 0 (vs >): T-Value = 2.47  P-Value = 0.007  DF = 582
```

d. What is the value of the test statistic?

2.47

e. What is the P-value for this test?

0.007

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

Since the p-value is less than α, we conclude that there is enough evidence that ETSU female students have owned more cell phones on average than female students.

Name:
E-number:
Section Number:

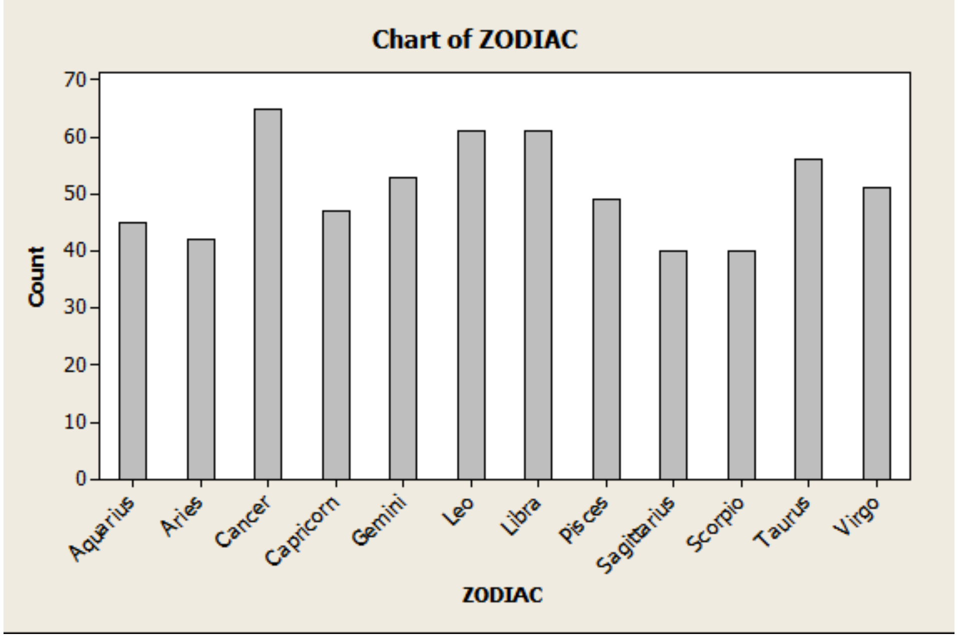
g. What assumptions are we making about the samples for our interpretation to be valid?

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

Name:
E-number:
Section Number:

10. **(BONUS PROBLEM) What's your sign?** (For reasons unknown to anybody, this was thought to be a good pick-up line back in the 70's.) For reasons known only to social scientists, the General Social Survey (GSS) regularly asks its subjects about their astrological sign. Question 4 from the survey asked, "What is your zodiac sign?" If births are spread uniformly across the year, we expect all 12 signs to be equally likely. Assume that the students who responded the survey are from an SRS of all ETSU students.

a. Create an appropriate graph to display the distribution of astrological signs in our sample and comment on the results. Insert the graph here.



It appears that the most common zodiac sign is Cancer and the least common is Sagittarius and Scorpio.

b. Is there good evidence that the zodiac signs of ETSU students are not equally likely? Perform a test of hypotheses using a significance level of α = 0.05. Include the appropriate test output and accompanying graphs from Minitab here.

What is your conclusion? If you concluded that not all signs are equally likely, which signs appear to be furthest from expectations based on the null hypothesis?

Use Stats ➔ Tables ➔ Chi-square Goodness-of-fit (one variable)…

Name:
E-number:
Section Number:

Chi-Square Goodness-of-Fit Test for Categorical Variable: ZODIAC

```
                             Test             Contribution
Category     Observed  Proportion  Expected     to Chi-Sq
Aquarius           45   0.0833333   50.8333       0.66940
Aries              42   0.0833333   50.8333       1.53497
Cancer             65   0.0833333   50.8333       3.94809
Capricorn          47   0.0833333   50.8333       0.28907
Gemini             53   0.0833333   50.8333       0.09235
Leo                61   0.0833333   50.8333       2.03333
Libra              61   0.0833333   50.8333       2.03333
Pisces             49   0.0833333   50.8333       0.06612
Sagittarius        40   0.0833333   50.8333       2.30874
Scorpio            40   0.0833333   50.8333       2.30874
Taurus             56   0.0833333   50.8333       0.52514
Virgo              51   0.0833333   50.8333       0.00055


  N  N*  DF   Chi-Sq  P-Value
610   0  11  15.8098    0.148
```

H_0: $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = p_8 = p_9 = p_{10} = p_{11} = p_{12} = 1/12$
H_a: Not all $p_i = 1/12$

The chi-square test statistics is 15.8098 and the p-value is 0.148.

Since the p-value is greater than α, we fail to reject the null hypothesis and conclude that there is not enough evidence to suggest not all signs are equally likely.