## Here are the questions that were asked on the survey:

1. GENDER: What is your gender? (Female, Male)
2. AGE: What is your age (in years)?
3. WEIGHT: What is your current weight (in pounds)?
4. HEIGHT: What is your height in feet and inches? (These data have been changed to inches)
5. NUCLEAR SAFETY: How safe would you feel if a nuclear energy plant were built near where you live? (Extremely safe, Very safe, Moderately safe, Slightly safe, Not at all safe)
6. TALK POLITICS: How many days in a typical week do you talk about politics with family or friends?
7. WASH HANDS: In a typical day, about how many times do you wash your hands?
8. LE CAMERAS: Should law enforcement officers be required to wear a camera on their uniform while on duty? (Yes, No)
9. CLOTHING ARTICLES: How many articles of clothing are you wearing right now?
10. CLOTHING PURCHASE: How much money did you spend on your last clothing purchase? (in US dollars)
11. LOWEST GAS PRICE: What is the lowest gas price you recall seeing at the gas station? (in US dollars)
12. FITNESS: About how much time per week (on average) do you devote to physical fitness? (Between zero and 2 hours, Between 2 and 5 hours, Between 5 and 9 hours, Between 9 and 15 hours, Over 15 hours per week)
13. SEXUAL PREDATOR: Do you have good reason to think you have ever been in contact with a sexual predator over the internet? (Yes, No)

A total of $\mathbf{8 0 0}$ students responded to the MATH1530 class survey during the spring semester of 2015. The data for students were recorded in the Minitab worksheet Sp15Math1530CSTPData.MTW.

The Minitab worksheet is set up as follows:

| C1: | ID (an identification number given to each student in the data file) |
| :--- | :--- |
| C2: | GENDER |
| C3: | AGE (yrs) |
| C4: | WEIGHT (lbs) |
| C5: | HEIGHT (in) |
| C6: | NUCLEAR SAFETY |
| C7: | TALK POLITICS |
| C8: | WASH HANDS |
| C9: | L E CAMERAS |
| C10: CLOTHING ARTICLES |  |
| C11: CLOTHING PURCHASE (\$) |  |
| C12: LOWEST GAS PRICE (\$) |  |
| C13: FITNESS |  |
| C14: SEXUAL PREDATOR |  |

Problem 1: Identify Variable Type. Which of these questions from the class survey measured 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. AGE
b. NUCLEAR SAFETY
c. WASH HANDS
d. CLOTHING PURCHASE
e. FITNESS

Categorical
Categorical
Categorical
Categorical
Categorical

| Quantitative | Neither |
| :--- | :--- |
| Quantitative | Neither |
| Quantitative | Neither |
| Quantitative | Neither |
| Quantitative | Neither |

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 (a). Then calculate the mean age of these first 10 observations and report the value below.

| n | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| AGE (yrs) | 20 | 18 | 19 | 19 | 26 | 19 | 20 | 20 | 19 | 19 |

The mean age of the first 10 students is $\underline{19.9}$ years. (Type the value into the space provided)
Identify the type of sampling method you have just used: CONVENIENCE SAMPLING
b. Next, select a random sample of size $\mathrm{n}=10$ (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 | 745 | 773 | 493 | 704 | 53 | 451 | 600 | 408 | 64 | 141 |
| AGE (yrs) | 21 | 28 | 58 | 18 | 20 | 20 | 18 | 21 | 19 | 22 |

Calculate and report the mean age for your random sample of 10 students. The sample mean age is $\underline{\mathbf{2 4 . 5}}$ years.
Identify the type of sampling method you have just used: SIMPLE RANDOM SAMPLING
(NOTE: Your random sample will probably differ from my random sample.)
c. Let's treat all the students who responded to the survey as a population for the purposes of this problem. Use Minitab to calculate the mean age for all 800 (?) observations included in the data set and report this value below.

The mean age of the population is $\underline{21.293}$ years.
d. Compare the population mean you found in Part (c) to the sample means you found in Parts (a) and (b). Which sample provided a closer estimate of the population mean age in this case?
IN THIS PARTICULAR CASE THE SAMPLE OF CONVENIENCE GAVE A MEAN AGE THAT IS CLOSER TO THE POPULATION MEAN AGE (BUT IT WON'T NECESSARILY BE TRUE FOR EVERYONE). THIS RESULT MAY SEEM SURPRISING BUT IF THE DATA ARE ENTERED INTO THE SPREADSHEET IN NO PARTICULAR ORDER, THEN THE SAMPLE OF CONVENIENCE MAY BE JUST AS GOOD AS A RANDOM SAMPLE. IN GENERAL, HOWEVER, IT IS MORE DIFFICULT TO RULE OUT HIDDEN SOURCES OF BIAS WHEN USING HAPHAZARD SAMPLING METHODS.

Name:
E-number:
Section Number:
Problem 3(F): If you are female then do this problem. (Omit this page/problem if you are male.)
Hand-Washing. Question 7 of the survey asked students, "In a typical day, about how many times do you wash your hands?"
a. Create an appropriate graph to display the distribution of the variable called WASH HANDS and insert it here.

b. Which of the following best describes the modality of the distribution shown in your graph? Underline your answer. Unimodal
c. Which of the following best describes the shape of the distribution? Underline your answer. Skewed right
d. Using Minitab, calculate the basic statistics for the data collected on WASH HANDS and copy \& paste the Minitab output here.

Descriptive Statistics: WASH HANDS (times per day)

| Variable | N | $\mathrm{N} *$ | Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum | Range |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| WASH HANDS | 794 | 6 | 8.594 | 10.826 | 0.000 | 4.000 | 5.000 | 10.000 | 100.000 | 100.000 |

e. Choose statistics that are appropriate for the shape of the distribution to describe the center and spread of WASH HANDS.
i. Which statistic will you use to describe the center of the distribution? (Type name of statistic here.) THE MEDIAN
(WE NEED RESISTANT MEASURES OF CENTER \& SPREAD BECAUSE OF THE SKEWNESS.)
ii. What is the value of that statistic? (Type value here.) 5 TIMES PER DAY
iii. Which statistic(s) will you use to describe the spread of the distribution? THE INTERQUARTILE RANGE (or IQR)
iv. What is(are) the value(s) of that(those) statistic(s)? 6 TIMES PER DAY
f. Are there any outliers in this distribution? Justify your answer.


Problem 3(M): If you are male then do this problem. (Omit this page/problem if you are female.)
Talking Politics. Question 6 of the survey asked students, "How many days in a typical week do you talk politics with family or friends?"
a. Create an appropriate graph to display the distribution of the variable called TALK POLITICS and insert it here.

b. Which of the following best describes the modality of the distribution shown in your graph? Underline your answer. Unimodal
c. Which of the following best describes the shape of the distribution? Underline your answer. Skewed right
d. Using Minitab, calculate the basic statistics for the data collected on TALK POLITICS and copy \& paste the Minitab output here.

## Descriptive Statistics: TALK POLITICS (days)

| Variable | $N$ | $N^{*}$ | Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum | Range | IQR |
| :--- | ---: | :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| TALK POLITICS | 794 | 6 | 1.5340 | 1.8416 | 0.0000 | 0.0000 | $\mathbf{1 . 0 0 0 0}$ | 2.0000 | 7.0000 | 7.0000 | $\mathbf{2 . 0 0 0 0}$ |

e. Choose statistics that are appropriate for the shape of the distribution to describe the center and spread of TALK POLITICS.
i. Which statistic will you use to describe the center of the distribution? (Type name of the statistic here.) THE MEDIAN (BECAUSE OF THE SKEWNESS, WE NEED RESISTANT MEASURES OF CENTER \& SPREAD.)
ii. What is the value of that statistic? (Type value here.) 1 DAY PER WEEK
iii. Which statistic(s) will you use to describe the spread of the distribution? THE INTERQUARTILE RANGE (or IQR)
iv. What is(are) the value(s) of that(those) statistic(s)? 2 DAYS PER WEEK
f. Are there any outliers in this distribution? Justify your answer.


Problem 4: Height versus Weight. It is not surprising to see a fairly strong association between height and weight in elementary school children. Does the same hold true for college-aged students? Questions 3 and 4 asked students to give their current weight in pounds (WEIGHT) and their height in feet and inches. From the heights supplied by students we have converted the data into total height in inches (HEIGHT). We are specifically interested in seeing whether we can use a student's height to accurately predict that person's weight.
a. Create an appropriate graph to display the relationship between WEIGHT and HEIGHT. 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. NOT SURPRISINGLY, TALLER STUDENTS TEND TO BE HEAVIER.
c. Describe the form of the relationship between WEIGHT and HEIGHT.

THE FORM IS SOMEWHAT LINEAR. IT WOULD SEEM REASONABLE TO FIT A STRAIGHT LINE TO THESE DATA.
d. Report the value of the correlation between this pair of variables? $r=\underline{0.482}$
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 OF THE ASSOCIATION IS WEAK TO MODERATE.
f. Using Minitab, obtain the equation for the least squares regression of WEIGHT on HEIGHT. Copy \& paste the output here.

## Regression Analysis: WEIGHT (lbs) versus HEIGHT (in)

The regression equation is: WEIGHT $=-176+5.09$ (HEIGHT)
g. Interpret the value of the slope in the least squares regression equation you found in part (f).

ON AVERAGE, WE EXPECT WEIGHT TO INCREASE BY ABOUT 5 POUNDS FOR EACH ADDITIONAL INCH OF HEIGHT.
h. Use the regression equation in part (e) to predict the weight for a student who is 67 inches tall. (Show your math.)

Predicted weight $=-176+5.09(67)=\underline{165 \mathrm{lbs}}$.
i. How well does the regression equation fit the data? Explain. Justify your answer with appropriate plot(s) and summary statistics.

| Least-Squares Regression of WEIGHT on HEIGHT PREDICTED WEIGHT $=-175.5+5.091$ (HEIGHT) |  | THE FACT THAT THE ASSOCIATION IS NOT A STRONG ONE IS SHOWN CLEARLY IN |
| :---: | :---: | :---: |
|  |  | THE FITTED LINE PLOT. THERE ARE MANY POINTS THAT ARE SCATTERED FAR AWAY FROM THE REGRESSION LINE. |
|  |  | THE SQUARED CORRELATION INDICATES THAT ONLY 23\% OF THE VARIATION WE |
|  |  | OBSERVED IN STUDENTS' WEIGHTS IS EXPLAINED BY THEIR HEIGHTS IN THE |
|  |  | LINEAR REGRESSION MODEL. THE OTHER SIDE OF THE COIN, THEN, IS THAT |
|  |  | ABOUT 77\% OF THE VARIATION IN WEIGHT IS EXPLAINED BY OTHER FACTORS SUCH AS GENETICS, NUTRITION, AMOUNT OF PHYSICAL ACTIVITY, ETC. |

Name:
E-number:
Section Number:


Problem 5: Physical Fitness versus Weight. You may have noticed from your analysis in Problem 4 that height does not explain 100\% of the variation that we have observed in students' heights. Is it possible that the amount of time students devote to physical fitness each week may help us to better understand their weights?
a. Question 12 of the survey asked students, "About how much time per week (on average) do you devote to physical fitness?" We have named this variable FITNESS. Create a suitable graph to display the distribution of FITNESS and insert it here.

EITHER A PIE CHART OR A BAR CHART WILL WORK WELL TO DISPLAY THE DISTRIBUTION.

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

Between 2 \& 5 hours
c. Create side-by-side boxplots to display students' weights for the different levels of FITNESS. (Go to Graph > Boxplot > One Y with Groups > OK. Select WEIGHT for the "Graph variables" slot and FITNESS for the "Categorical variables for grouping" slot.) Insert your graph here.

d. Use Minitab to calculate the basic statistics of WEIGHT for each level of FITNESS. Copy and paste the output here.

Descriptive Statistics: WEIGHT (lbs)

| Variable | FITNESS (hrs per week) | N $\mathrm{N}^{*}$ | Mean | StDev | Minimum | Q1 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| WEIGHT (lbs) | Between 0 and 2 hours | 2375 | 170.04 | 48.87 | 88.00 | 134.50 |
|  | Between 2 and 5 hours | 2463 | 164.20 | 44.26 | 95.00 | 132.00 |
|  | Between 5 and 9 hours | 1521 | 163.52 | 39.79 | 89.00 | 135.25 |
|  | Between 9 and 15 hours | 1000 | 169.97 | 45.25 | 102.00 | 137.00 |
|  | Over 15 hours per week | 560 | 190.21 | 45.24 | 104.00 | 157.50 |
| Variable <br> WEIGHT (lbs) | FITNESS (hrs per week) | Median | Q3 | Maximum | Range | IQR |
|  | Between 0 and 2 hours | 160.00 | 200.00 | 335.00 | 247.00 | 65.50 |
|  | Between 2 and 5 hours | 151.50 | 181.25 | 345.00 | 250.00 | 49.25 |
|  | Between 5 and 9 hours | 158.50 | 180.00 | 350.00 | 261.00 | 44.75 |
|  | Between 9 and 15 hours | 165.00 | 190.00 | 380.00 | 278.00 | 53.00 |
|  | Over 15 hours per week | 185.00 | 214.25 | 300.00 | 196.00 | 56.75 |

e. With regard to FITNESS levels, which group of students has the lowest mean weight?
f. Discuss the results: Describe the distributions of WEIGHT for the different levels of FITNESS as well as draw comparisons (i.e., What do they have in common?) and contrasts (i.e., How are they different?) between these distributions. Are there any surprises in the results? Explain why you think so, or why not.

(FYI: THIS TYPE OF DISPLAY IS NOT REQUIRED, BUT IS INFORMATIVE NONETHELESS.) DOTPLOTS, ALTHOUGH NOT DISCUSSED BY DR. MOORE (ET. AL.) IN CHAPTER 1 OF ES2, DISPLAY A SLIGHTLY DIFFERENT PERSPECTIVE OF THESE DISTRIBUTIONS. FOR ONE THING, THEY GIVE US A GLIMPSE OF MODALITY, A FEATURE THAT IS LOST IN THE BOXPLOTS. SECONDLY, LIKE THE BOXPLOTS, THEY ALSO MAKE IT EASY TO COMPARE AND CONTRAST SEVERAL DISTRIBUTIONS, WHICH IS SOMETHING DIFFICULT TO ACHIEVE WITH HISTOGRAMS. MY DESCRIPTIONS GIVEN BELOW ARE BASED ON WHAT I SEE IN THE SIDE-BY-SIDE BOXPLOTS (FOR SYMMETRY AND OUTLIERS), THE DOTPLOTS (FOR MODALITY), AND THE BASIC STATISTICS GIVEN BY MINITAB (FOR CENTER AND SPREAD).

BETWEEN 0 AND 2 HOURS PER WEEK: THE SHAPE IS UNIMODAL AND MODERATELY SKEWED TO THE RIGHT. THE MEDIAN WEIGHT FOR STUDENTS WHO REPORTED THIS LEVEL OF FITNESS IS 160 POUNDS. THIS DISTRIBUTION SHOWS THE LARGEST AMOUNT OF SPREAD AS INDICATED BY THE IQR ( 65.5 LBS.) AND STANDARD DEVIATION (48.87 LBS.); HOWEVER, THE RANGE IN WEIGHTS IS THE SECOND SMALLEST (247 LBS.). THERE ARE FOUR HIGH OUTLIERS RANGEING FROM 321 TO 335 LBS.
between 2 AND 5 HOURS PER WEEK: THE SHAPE IS UNIMODAL AND MODERATELY SKEWED TO THE RIGHT. THE MEDIAN WEIGHT IS 151.5 LBS. AND THE IQR IS 49.25 LBS. THESE STUDENTS HAVE THE LOWEST MEDIAN WEIGHT (THOUGH NOT THE LOWEST MEAN WEIGHT). THERE ARE 9 HIGH OUTLIERS RANGEING FROM 265 TO 345 LBS. THE LARGEST NUMBER OF STUDENTS (246) REPORTED THIS FITNESS LEVEL. (NOTE: ALL OF THESE DISTRIBUTIONS HAVE SOME RIGHT SKEWNESS AND HIGH OUTLIERS.)

BETWEEN 5 AND 9 HOURS PER WEEK: THIS DISTRIBUTION APPEARS TO BE BIMODAL AND SKEWED TO THE RIGHT. THE MEDIAN WEIGHT IS 158.5 LBS. AND THE IQR IS 44.75 LBS. THESE STUDENTS HAVE THE SMALLEST MEAN WEIGHT ( 163.52 LBS.), AS WELL AS THE SMALLEST IQR AND STANDARD DEVIATION (39.79 LBS.). THERE ARE 6 HIGH OUTLIERS RANGEING FROM 250 TO 350 LBS.

BETWEEN 9 AND 15 HOURS PER WEEK: THIS APPEARS TO BE A MULTIMODAL DISTRIBUTION THAT IS ALSO SKEWED TO THE RIGHT. THE MEDIAN WEIGHT IS 165 LBS. AND THE IQR IS 53 LBS. THE HIGHEST MAXIMUM WEIGHT IS FOUND IN THIS GROUP ( 380 LBS.), AS WELL AS THE LARGEST OVERALL RANGE IN WEIGHTS ( 278 LBS.). THERE ARE 5 HIGH OUTLIERS RANGEING FROM 270 TO 380 LBS.

OVER 15 HOURS PER WEEK: IT APPEARS THAT THIS DISTRIBUTION IS BIMODAL AND ONLY SLIGHTLY SKEWED TO THE RIGHT. THE MEDIAN WEIGHT IS 185 LBS., WHICH IS THE HIGHEST MEDIAN WEIGHT AMONG THE DIFFERENT LEVELS OF FITNESS. THE MEAN WEIGHT IS ALSO HIGHEST FOR THESE STUDENTS ( 190 LBS.). THE IQR IS 56.75 LBS. AND THE OVERALL RANGE IN WEIGHTS, AT 196 LBS., IS THE SMALLEST FOR THIS GROUP OF STUDENTS. THIS IS ALSO THE SMALLEST GROUP OF STUDENTS WITH ONLY 56 OF THEM REPORTING THIS LEVEL OF FITNESS. THERE IS ONLY ONE HIGH OUTLIER AT 300 LBS.

SURPRISES? WELL, IT MAY SEEM A BIT PARADOXICAL THAT STUDENTS CLAIMING TO DEVOTE THE LARGEST AMOUNT OF TIME TO PHYSICAL FITNESS, OVER 15 HOURS PER WEEK, ALSO HAVE THE HIGHEST MEAN AND MEDIAN WEIGHTS. HOWEVER, PERHAPS THEY ARE MORE INTERESTED IN BODY BUILDING THAN CONTROLLING WEIGHT. MUSCLE WEIGHS MORE THAN FAT.

Problem 6 (Even): 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.)
Gender and Nuclear Safety. Question 5 in the survey asked students "How safe would you feel if a nuclear energy plant were built near where you live?" (Students could choose one of these options: Extremely safe, Very Safe, Moderately safe, Slightly safe, or Not at all safe.) Is there a relationship between gender and students' opinions about nuclear safety?
a. Create an appropriate graph to display the relationship between GENDER and NUCLEAR SAFETY. You don't want to display information for students that didn't answer both of these questions on the survey, so click on Data Options > Group Options and remove the checks in the boxes beside "Include missing as a group" and "Include empty cells." Insert your graph here.

## EITHER A STACKED BAR CHART OR A CLUSTERED BAR CHART WILL WORK WELL TO DISPLAY THE ASSOCIATION.


b. Create an appropriate two-way table to summarize the data. Click on Options > Display missing values for... and put a dot in the circle beside "No variables." Insert your table here.

c. SUPPOSE WE SELECT ONE STUDENT AT RANDOM: (Calculate the following probabilities and show your work.)
i. What is the probability that this student is a female and feels "very safe"? $\quad \mathrm{P}=24 / 796=0.030150753$ (OR A 3.02\% CHANCE)
ii. What is the probability that this student is either a male or that he/she feels "very safe"?

$$
P=(343+69-45) / 796=0.461055276(\text { OR } 46.11 \%)
$$

iii. What is the probability that this student feels "not at all safe" given that the student selected is a female?

$$
P=180 / 453=0.397350993 \text { (OR 39.74\%) }
$$

iv. What is the probability that this student is a male given that the student selected feels "not at all safe"?

$$
P=91 / 271=0.335793357 \text { (OR 33.58\%) }
$$

Name:
E-number:
Section Number:
d. Do you think there may be an association between GENDER and NUCLEAR SAFETY? Why or why not? Explain your reasoning based on what you see in your graph.

YES, THERE DOES APPEAR TO BE AN ASSOCIATION. IT IS EASY TO SEE THAT THE PATTERN OF RESPONSES, IN THE CLUSTERED BAR CHART, IS DIFFERENT FOR FEMALES THAN IT IS FOR MALES. IT LOOKS LIKE FEMALES TEND TO BE MORE CONCERNED THAN MALES ABOUT THE SAFETY OF HAVING A NUCLEAR ENERGY PLANT NEAR THEIR HOME.

Problem 6 (Odd): 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.)
Gender and Physical Fitness. You are already familiar with the variable called FITNESS. Now we want to investigate further to see if there is a relationship between a student's gender and the amount of time devoted to physical fitness per week.
a. Create an appropriate graph to display the relationship between GENDER and FITNESS. Insert your graph here.

EITHER A STACKED BAR CHART OR A CLUSTERED BAR CHART WILL WORK WELL TO DISPLAY THE ASSOCIATION.

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

Tabulated statistics: GENDER, FITNESS (hrs per week)
Rows: GENDER Columns: FITNESS (hrs per week)

|  | Between <br> 0 and 2 <br> hours | Between <br> 2 and 5 <br> hours | Between <br> 5 and 9 hours | Between 9 and 15 hours | Over 15 hours per week | All |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Female | 158 | 162 | 82 | 44 | 9 | 455 |  |  |
|  | 34.73 | 35.60 | $18 . \overline{02}$ | 9.67 | 1.98 | 100.00 |  |  |
|  | 65.29 | 65.06 | 53.59 | 44.00 | 16.07 | 56.88 |  |  |
|  | 19.75 | 20.25 | 10.25 | 5.50 | 1.13 | 56.88 |  |  |
| Male | 84 | 87 | 71 | 56 | 47 | 345 |  |  |
|  | 24.35 | 25.22 | 20.58 | 16.23 | 13.62 | 100.00 |  |  |
|  | 34.71 | 34.94 | 46.41 | 56.00 | 83.93 | 43.13 |  |  |
|  | 10.50 | 10.88 | 8.88 | 7.00 | 5.88 | 43.13 |  |  |
| All | $\underline{242}$ | 249 | 153 | 100 | 56 | 800 | Cell Contents: | Count |
|  | 30.25 | 31.13 | 19.13 | 12.50 | 7.00 | 100.00 |  | \% of Row |
|  | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |  | \% of Column |
|  | 30.25 | 31.13 | 19.13 | 12.50 | 7.00 | 100.00 |  | \% of Total |

c. SUPPOSE WE SELECT ONE STUDENT AT RANDOM: (Calculate the following probabilities and show your work.)
i. What is the probability that this student is a male and devotes over 15 hours per week to physical fitness?

$$
P=47 / 800=0.05875 \text { (OR A 5.88\% CHANCE) }
$$

ii. What is the probability that this student is either a female or that he/she devotes between 5 and 9 hours to physical fitness?

$$
P=(455+153-82) / 800=0.6575(\text { OR } 65.75 \%)
$$

iii. What is the probability that this student devotes between zero and 2 hours per week to physical fitness given that the student selected is a female? $\quad P=158 / 455=0.347252747$ ( $O R 34.73 \%$ )
iv. What is the probability that this student is a female given that the student selected devotes between zero and 2 hour per week to physical fitness? $\quad \mathrm{P}=\mathbf{1 5 8 / 2 4 2}=0.652892652$ ( $O R 65.29 \%$ )
d. Do you think there may be an association between GENDER and FITNESS? Why or why not? Explain your reasoning based on what you see in your graph.

YES, THERE DOES APPEAR TO BE AN ASSOCIATION. THE PATTERNS SHOWN IN THE CLUSTERED BAR CHART ARE SOMEWHAT SIMILAR BETWEEN THE TWO GENDERS BUT THE PERCENTAGES ARE DIFFERENT. FEMALE STUDENTS TEND TO DEVOTE FEWER HOURS PER WEEK TO PHYSICAL FITNESS COMPARED TO THE MALES.

Problem 7: Lowest Gas Prices. Survey question \#11 asked, "What is the lowest gas price you recall seeing at the gas station?" However, people who work with college students on a regular basis might wonder if they really pay attention to such details as the price of gasoline. We may be able to use our sample data to perform a test to see if this is true. AAA reports in their Daily Fuel Gauge Report* that the average price of regular grade gasoline, in the state of Tennessee, was $\$ 1.922$ per gallon during the first week of February (when many of our Math-1530 students took the survey). The price of regular gas is lower than the other grades, so if students are reporting the lowest price, I will assume it is probably for regular.
*http://fuelgaugereport.aaa.com/states/tennessee/
a. Create a suitable graph to display the distribution of gas prices reported by our sample of college students and insert it here.

b. Describe the distribution shown in your graph.

THE DISTRIBUTION IS UNIMODAL AND ONLY SLIGHTLY SKEWED TO THE LEFT BUT THE BOXPLOT SHOWS NUMEROUS HIGH AND LOW OUTLIERS. HERE IS THE 5-NUMBER SUMMARY:

$$
\begin{aligned}
& \text { MINIMUM }=\underline{0.00} \$ / \text { GALLON (WHICH LEAVES ONE WONDERING WHERE THEY MIGHT HAVE GONE TO GET FREE GASOLINE) } \\
& Q_{1}=\underline{1.72} \$ / \text { GALLON MEDIAN }=\underline{1.85} \$ / \text { GALLON } \quad Q_{3}=\underline{1.90 ~} \$ / \text { GALLON MAXIMUM }=\underline{4.00} \$ / \text { GALLON }
\end{aligned}
$$

c. Perform a test of significance to see if all college students would truly report low gas prices on average. If this claim is true, then the average price reported by students should be less than the average price reported by AAA. For this test, the null hypothesis is that the average price reported by students is the same as the average price reported by AAA. Thus,

$$
\mathrm{H}_{0}: \mu=\$ 1.922 \text { per gallon }
$$

Write the correct alternative hypothesis for the test.
$H_{a}: \mu<\$ 1.922$ per gallon
d. Use Minitab to perform the appropriate test. Copy and paste the output for the test here.

e. What is the name of your test statistic and what is its value? WE ARE PERFORMING A ONE-SAMPLE t -TEST AND $\mathrm{t}=-10.91$.
f. What is the $P$-value for the test? $P=0.000$
g. State your decision regarding the hypothesis being tested. SINCE THE P-VALUE IS VERY SMALL WE SHOULD REJECT Ho.
h. State your conclusion. COLLEGE STUDENTS APPEAR TO BE PAYING ATTENTION. ON AVERAGE, THEY DO REPORT LOW GAS PRICES.
i. Is the P-value valid in this case? What assumptions are you making in order to carry out this test? SEE "USING THE $t$ PROCEDURES" ON P. 364 OF ES2.
FIRSTLY, WE NEED A SRS OF COLLEGE STUDENTS (BUT WE HAVE A SAMPLE OF CONVENIENCE, HEAVILY LOADED WITH FRESHMEN). SECONDLY, WE HAVE A LARGE SAMPLE ( $\mathrm{n}=789$ ) SO THE t-PROCEDURES SHOULD BE ROBUST AGAINST ALL THOSE OUTLIERS. IN OUR FAVOR, THE DEGREE OF SKEWNESS IS RELATIVELY SMALL AND THERE ARE BOTH HIGH AND LOW OUTLIERS, WHICH TEND TO balance out in this case.

Bonus Problem: Sexual Predators on the Internet. According to the online child safety website, PureSight*, "one in five U.S. teenagers who regularly log on to the Internet says they have received an unwanted sexual solicitation via the Web." (NOTE: One in 5 is the same as $20 \%$.) Is the same true for the population of students enrolled in U.S. colleges and universities? Survey question \#13 asked our Math1530 students, "Do you have good reason to think you have ever been in contact with a sexual predator over the internet?" In the data worksheet, we call this variable SEXUAL PREDATOR.
*(http://www.puresight.com/Pedophiles/Online-Predators/online-predators-statistics.html)
a. Create an appropriate graph to display the distribution of SEXUAL PREDATOR and insert it here.

EITHER A PIE CHART OR A BAR CHART WILL WORK WELL TO DISPLAY THE DISTRIBUTION.


IT MAY ALSO BE HELPFUL TO GET MINITAB TO TALLY UP THE RESPONSES FOR THIS VARIABLE (GO TO STAT > TABLES > TALLY INDIVIDUAL VARIABLES.)

Tally for Discrete Variables: SEXUAL PREDATOR

| SEXUAL |  |  |
| :---: | :--- | :---: |
| PREDATOR | Count | Percent |
| No | 635 | 80.08 |
| Yes | 158 | 19.92 |
| N $=$ | 793 |  |
| $*=$ | 7 |  |

b. How many of the students surveyed said "yes" (in response to survey question \#13)? 158 STUDENTS
c. What proportion of our sample said "yes?" ABOUT 0.199 OR 19.9\% OF THE STUDENTS SURVEYED SAID YES.
d. Assume (for the purpose of this problem) that we may treat our 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 would say "yes" to the survey question (based on our sample data). Copy and paste the Minitab output here.

Cl for One Proportion: SEXUAL PREDATOR
Event = Yes

|  | Xariable | N | Sample p | 95\% |
| :--- | ---: | ---: | ---: | ---: | ---: |
| SEXUAL PREDATOR | 158 | 793 | 0.199243 | $\mathbf{( 0 . 1 7 1 4 4 3 , ~ 0 . 2 2 7 0 4 4 )} \quad$ Using the normal approximation. |

HERE ARE THE RESULTS IF WE CHOOSE NOT TO USE A NORMAL APPROXIMATION (ALTHOUGH IT SHOULD BE OKAY TO DO SO SINCE WE HAVE A LARGE SAMPLE). NOTICE THAT THE 95\% CONFIDENCE INTERVALS ARE VERY SIMILAR EITHER WAY.

## Cl for One Proportion: SEXUAL PREDATOR

Event = Yes

| Variable | X | N | Sample p | 95\% CI |
| :--- | ---: | ---: | ---: | ---: |
| SEXUAL PREDATOR | 158 | 793 | 0.199243 | $(0.171972,0.228771)$ |

Name:
E-number:
Section Number:
e. Interpret the confidence interval you reported in part (d).

WE ARE 95\% CONFIDENT THAT THE TRUE PERCENTAGE OF ALL U.S. COLLEGE/UNIVERSITY STUDENTS THAT WOULD SAY YES TO THE SURVEY QUESTION IS SOMEWHERE BETWEEN 17 AND $23 \%$.
f. What do you think? Do our results contradict the claim made at the PureSight website or do they appear to agree with it? EXPLAIN. TWENTY PERCENT IS WITHIN OUR 95\% CONFIDENCE INTERVAL, SO OUR RESULTS APPEAR TO BE IN AGREEMENT WITH THE CLAIM MADE AT PureSight .
(IN FACT, WE CAN USE THE 95\% CONFIDENCE INTERVAL TO PERFORM A TWO-SIDED TEST OF HYPOTHESES FOR A ONE-SAMPLE TEST OF A POPULATION PROPORTION. IN THIS CASE, OUR TEST WOULD NOT BE CONSIDERED SIGNIFICANT AT THE 5\% ALPHA LEVEL. FOR A TEST OF $H_{0}: p=20 \%$ VERSUS $H_{a}: p \neq 20 \%$, WE SHOULD NOT REJECT THE NULL HYPOTHESIS.)

