**MATH-1530 CAPSTONE TECHNOLOGY PROJECT (100 POINTS) SPRING SEMESTER 2015**

**Directions:**

1. Type your Name, E number, and Section number in the header to the right of the colons. Double click on the header

and then click on the toolbar “Close header and footer” in the menu list.

2. **DO YOUR OWN WORK! It is academic misconduct to copy or seek assistance from other people, or to share your**

**work with other students.** **Any academic misconduct on this project will result in a grade of 0 and a written report**

**sent to the dean’s office**.

3. The Capstone Project counts for 10% of your final grade in this course.

4. The project is due via digital dropbox on D2L on **April 30, 2015 at the beginning of class**. Seriously, no late projects will be graded. Don’t wait till the last minute to start working—you know how computer technology can fail at ETSU without even a moment’s notice.

5. The first 2 problems will probably fit on a single page. After that, please start each problem at the top of a new page.

NOTE: As you type in your answers and insert graphical displays, it will advance the next problems along so you may

need to operate the *backspace* or *delete* keys in order to make the problems start back at the top of a new page.

6. Another problem that can arise from typing in answers, discussions, etc. is that auto-formatting kicks in and wants to

insert new numbers or letters into the outline format of this document. I find that one of the best ways to get rid of

these unwanted additions is to click on the back-arrow (undo) icon immediately to indicate to M.S. Word that the auto-

formatting is not desired. It doesn’t always work, mind you, M.S. Word can be very stubborn about auto-formatting.

7. Insert all graphs and R output within a problem as requested. (DO NOT ATTACH AT THE END.)

8. Please take pity on your poor teacher and make it easier for me to find your answers/discussions. Use might use a

different font for your answers or make them bold print. If you are using a color printer (with fresh ink cartridges) you

could highlight in yellow (other colors will obscure your typing in the printed version) or use a different color of ink for

your responses.

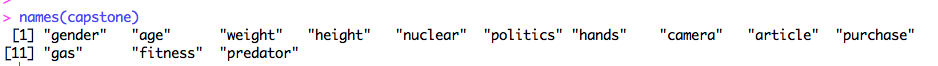
9. NEATNESS COUNTS! Give me a clean, professional presentation—Bonus points may be involved.

10. **Do not hand in these 1st two pages—just the problems, please.**

**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:** 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. **POLITICS:** How many days in a typical week do you talk about politics with family or friends?
7. **HANDS:** In a typical day, about how many times do you wash your hands?
8. **CAMERAS:** Should law enforcement officers be required to wear a camera on their uniform while on duty? (Yes, No)
9. **ARTICLES:** How many articles of clothing are you wearing right now?
10. **PURCHAS**E**:** How much money did you spend on your last clothing purchase? (in US dollars)
11. **GAS:** 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. **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 811 students responded to the MATH1530 class survey during the spring semester of 2015. The name of the data file is Capstone R Data.txt. Do not download the data. Click on the file directly in D2L. Click anywhere in the file. Hit cltr A (command A for a Mac) and then hit cltr C (command C for a Mac). Now open up R and click in R and hit cltr V (command V for a Mac). Then hit enter. Below is a print screen of what you should see if you correctly loaded the data set in R**

****The R data file is set up as follows (Note: When using the variables, use all lowercase):

|  |
| --- |
| **gender** |
| **age** |
| **weight** |
| **height** |
| **nuclear** |
| **politics** |
| **hands** |
| **camera** |
| **article** |
| **purchase** |
| **gas** |
| **fitness** |
| **predator** |

**MATH-1530 CAPSTONE TECHNOLOGY PROJECT SPRING SEMESTER 2015**

**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).

* 1. **AGE** Categorical Quantitative Neither
  2. **NUCLEAR SAFETY** Categorical Quantitative Neither
  3. **WASH HANDS** Categorical Quantitative Neither
  4. **CLOTHING PURCHASE** Categorical Quantitative Neither
  5. **FITNESS** Categorical 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. In R, type: **age[1:10]**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| n | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **AGE (yrs)** |  |  |  |  |  |  |  |  |  |  |

The mean age of the first 10 students is **\_\_\_\_\_** years. (Type the value into the space provided)  
Identify the type of sampling method you have just used: **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

b. Next, select a random sample of size n = 10. In R, type: **sample(age,10)**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| n | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **AGE (yrs)** |  |  |  |  |  |  |  |  |  |  |

Calculate and report the mean age for your random sample of 10 students. The sample mean age is **\_\_\_\_\_\_** years.  
Identify the type of sampling method you have just used: **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

c. Let’s treat all the students who responded to the survey as a *population* for the purposes of this problem. Use R 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 **\_\_\_\_\_\_** 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?

**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 ***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 Bimodal Multimodal

c. Which of the following best describes the shape of the distribution? Underline your answer.

Skewed left Symmetric Skewed right

d. Using R, calculate the basic statistics for the data collected on ***hands*** and copy & paste the R output here.

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

i. Which statistic will you use to describe the center of the distribution? (Type name of statistic here.)

ii. What is the value of that statistic? (Type value here.)

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

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

f. Are there any outliers in this distribution? If so, what are their values? 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 ***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 Bimodal Multimodal

c. Which of the following best describes the shape of the distribution? Underline your answer.

Skewed left Symmetric Skewed right

d. Using R, calculate the basic statistics for the data collected on ***politics*** and copy & paste the R output here.

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

i. Which statistic will you use to describe the center of the distribution? (Type name of the statistic here.)

ii. What is the value of that statistic? (Type value here.)

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

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

f. Are there any outliers in this distribution? If so, what are their values? 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.

c. Describe the *form* of the relationship between *weight* and *height*.

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

e. Based on the information displayed in the graph and the correlation you just reported, how would you describe the *strength* of the association?

f. Using R, obtain the equation for the least squares regression of ***weight*** on ***height***. Copy & paste the output here.

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

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 = \_\_\_\_\_\_\_\_\_\_**

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

**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.

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

Between 0 & 2 hours Between 2 & 5 hours Between 5 & 9 hours Between 9 & 15 hours Over 15 hours

c. Create side-by-side boxplots to display students’ weights for the different levels of ***fitness***. Insert your graph here.

d. Use R to calculate the basic statistics of ***weight*** for each level of ***fitness***. Copy and paste the output here.   
In R, type: **tapply(weight, fitness,summary)**  
In R, also type: **tapply(weight, fitness,sd)**

e. With regard to ***fitness*** levels, which group of students has the lowest mean weight? (Please underline or highlight one option.)

Between 0 & 2 hours Between 2 & 5 hours Between 5 & 9 hours Between 9 & 15 hours Over 15 hours

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.

**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***. Insert your graph here.

b. Create an appropriate two-way table to summarize the data. Insert your table here. In R, type: **addmargins(table(gender,nuclear))**

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”?

ii. What is the probability that this student is either a male *or* that he/she feels “very safe”?

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

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

d. Do you think there may be an association between ***gender*** and ***nuclear***? Why or why not? Explain your reasoning based on what you see in your graph.

**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.

b. Create an appropriate two-way table to summarize the data. Insert your table here. In R, type: **addmargins(table(gender,fitness))**

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?

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?

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?

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?

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.

**Problem 8:** **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.

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,

**Ho**: μ = $1.922 per gallon

Write the correct alternative hypothesis for the test.

**Ha**:

d. Use R 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?

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

g. State your decision regarding the hypothesis being tested.

h. State your conclusion.

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

**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 at U.S. colleges and universities? Survey question #13 asked our Math-1530 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 ***predator***.

\*(http://www.puresight.com/Pedophiles/Online-Predators/online-predators-statistics.html)

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

b. How many of the students surveyed said “yes?” In R, type: **addmargins(table(predator))**

c. What proportion of our sample 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 R 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 R output here.   
In R, type: **prop.test(164,811,p=.2)**

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

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.