MAT 111 Test 1

Student Name: _____

Answer the questions in the spaces provided.

1. In the General Social Survey from the year 2002, some of the subjects were asked whether or not they owned a gun, and were also asked whether or not they favored capital punishment. The results are given in the two-way table below:

```
##       cappun
## owngun Favor Oppose
##    No    375    199
##    Yes   243     59
```

Here are the row percents for the table:

```
##     Favor Oppose Total
## No  65.33  34.67   100
## Yes 80.46  19.54   100
```

Here are the column percents:

```
##        Favor Oppose
## No     60.68  77.13
## Yes    39.32  22.87
## Total 100.00 100.00
```

Use these tables to answer the questions below.

(a) True or False, and explain briefly: "People who do not own a gun are more likely to oppose capital punishment than are people who own a gun, because in the sample there are 199 non-gun-owners who oppose capital punishment, whereas there are only 59 gun-owners who oppose capital punishment."

(b) Consider the people in the study who favor capital punishment: what percentage of them do not own a gun?

(c) What percentage of the people in the study who do not own a gun are people who oppose capital punishment?

(d) Describe the relationship, in the sample, between **owngun** and **cappun**. Use two relevant percentages to back up your answer.

2. Assume that the subjects in the GSS 2008 survey are a random sample from the population of all adults in the United States. We are interested in knowing whether the data in the previous problem provide strong evidence for a relationship ,in the U.S. population, between **owngun** and **cappun**. To that end, we ask R to perform a χ^2-test of significance, and we get the following results:

```
## Pearson's Chi-squared test with Yates' continuity correction
##
## Observed Counts:
##       cappun
## owngun Favor Oppose
##    No    375    199
##    Yes   243     59
##
## Counts Expected by Null:
##       cappun
## owngun Favor Oppose
##    No  404.9 169.05
##    Yes 213.1  88.95
##
## Contributions to the chi-square statistic:
##       cappun
## owngun Favor Oppose
##    No   2.21   5.30
##    Yes  4.21  10.08
##
##
## Chi-Square Statistic = 21.09
## Degrees of Freedom of the table = 1
## P-Value = 4.389e-06
```

Use these results to answer the following questions:

(a) The first step in a test of significance is to state the Null and Alternative Hypotheses. Please do so in the space below:

(b) If someone believes that **owngun** and **cappun** are unrelated in the population, about how many people in the survey would he/she expect to be gun-owners who oppose capital punishment?

(c) The number of observed gun-owners who oppose capital punishment differs from the expected number in the question above, leading to a contribution to the χ^2-statistic. What is the numerical value of this contribution?

(d) What is the value of the χ^2-statistic? (This is Step Two in a test of significance.)

(e) If someone believes that there is no relationship between **owngun** and **cappun** in the population, then about what would he/she expect the χ^2-statistic to be?

(f) (Step Three) What is the P-value for this test of significance?

(g) If there is no relationship between **owngun** and **cappun** in the population, then about what is the chance that a study like this one would result in a χ^2-statistic at least as big as the one we actually got in this study?

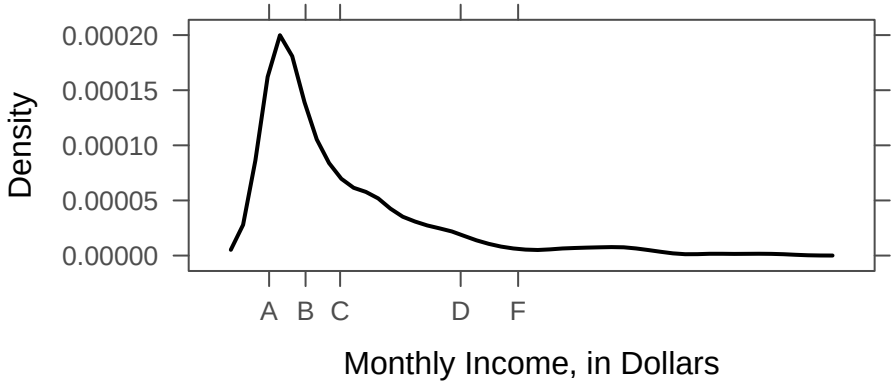
(h) (Step Four) Should we reject the Null Hypothesis, or not reject it? Base your decision on the P-value, using a level of significance ("cut-off value") of $\alpha = 0.05$.

(i) (Step Five) Write a brief, one sentence conclusion in nontechnical language, stating how much evidence the data provide for the Alternative Hypothesis. Do not use the term "Alternative Hypothesis" in your answer.

(j) Like most significance-test procedures, the `xchisq.test` function produces only an approximation to the P-value. In this case, can the approximation be considered trustworthy? What specific numerical features of the output justify your answer?

3. The following graph is a density plot of the monthly incomes of 2000 people.

Figure 1: Monthly Incomes of 2000 People



(a) One of the five labeled points on the x-axis is at the median monthly income. Which one is it?

(b) Another of the the labeled points is at the mean monthly income. Which one is it?

4. A histogram of the times for a large sample of African Dung Beetles to gather and process a ball of dung is roughly bell-shaped, with a mean of 20 minutes and a standard deviation of 3 minutes. Use the Empirical Rule to answer the following questions:

(a) About what percentage of the Dung Beetles require more than 26 minutes to process their balls of dung?

(b) Fill in the blanks: about 68% of the Dung Beetles require between ____ and ____ minutes to process their dung balls.

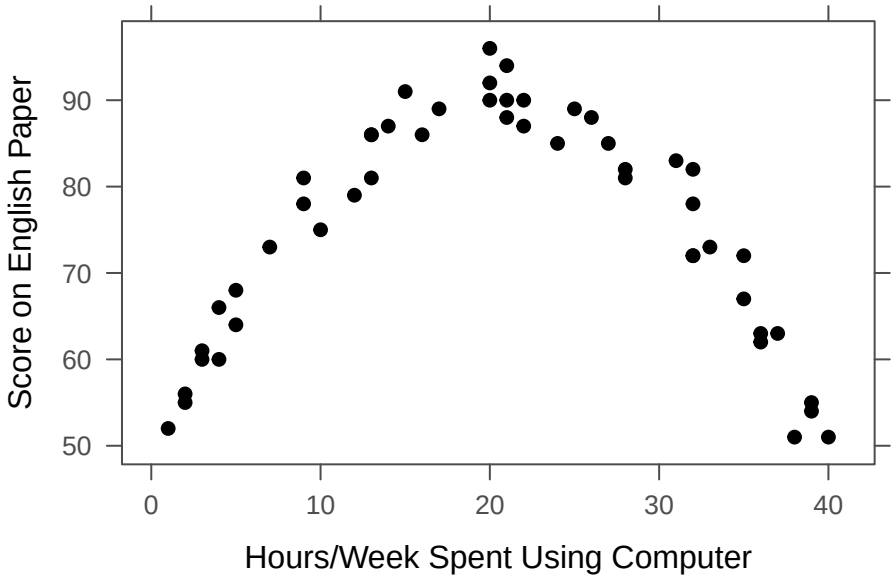
(c) Fill in the blanks: the percentage of Dung Beetles who require less than 15.5 minutes to process their dung balls is somewhere in between ____% and ____%.

5. In a survey of students in an English class, each subject was asked how how many hours per week he or she spent using a computer. The score of each student on the final major essay in the class was also recorded. A scatterplot of these scores appears below, preceded by the first few lines of the data frame **survey** that contains the data itself.

```
head(survey)

##   ComputerHours EnglishScore
## 1             9           81
## 2            26           88
## 3            10           75
## 4            17           89
## 5            37           63
## 6            21           88
```

Figure 2: English Score vs. Computer Hours Per Week



The correlation coefficient for the scatter plot is computed as: follows:

```
with(survey, cor(ComputerHours, EnglishScore))

## [1] -0.01919
```

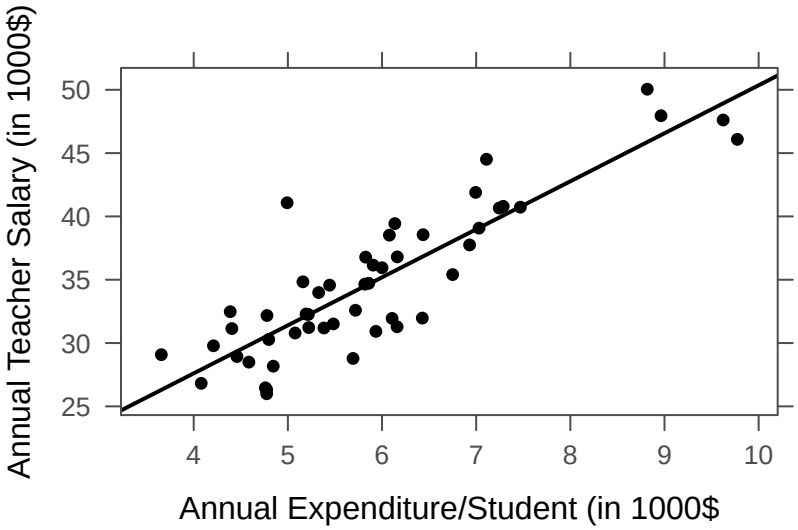
The question is on the next page.

Which of the following of the following is correct? Choose one option and briefly explain your choice.

- ◯ There is a strong negative relationship between **ComputerHours** and **EnglishScore**.
- ◯ There is a weak negative relationship between **ComputerHours** and **EnglishScore**.
- ◯ There is essentially no relationship between **ComputerHours** and **EnglishScore**.
- ◯ There is a weak positive relationship between **ComputerHours** and **EnglishScore**.
- ◯ There is a strong positive relationship between **ComputerHours** and **EnglishScore**.
- ◯ We need to do a regression analysis (`lmGC(EnglishScore ~ ComputerHours, data=survey)`) before we can describe any relationship that exists.
- ◯ None of the above is correct.

6. In the **sat** data from class, we might be interested in studying the relationship (if any) between the amount of money spent to educate students and the salary of their teachers. Below is a scatterplot of **salary**, the mean salary of teachers in a state, versus **expend**, the mean annual expenditure of education funds per student in that state. Both variables are measured in thousands of dollars. The plot includes the regression line.

Figure 3: Teacher Salary vs. Expenditure



Here is the code and the output for a regression analysis:

```
lmGC(salary ~ expend, data = sat)
```

```
##
##  Simple Linear Regression
##
## Correlation coefficient r =  0.8698
##
## Equation of Regression Line:
##
##   salary = 12.44 + 3.792 * expend
##
## Residual Standard Error: s   = 2.962
## R^2 (unadjusted): R^2 = 0.7566
```

Answer the following questions

(a) One of the states (California) appears to be an outlier. For this state, give approximately:

i. the mean annual expenditure per student, in thousands of dollars:

ii. the mean annual teacher salary, in thousands of dollars:

(b) What is the slope of the regression line?

(c) In a complete sentence, give a practical interpretation of the slope. Use the numerical value of the slope in your sentence. (You may round to one decimal place, if you like.)

(d) Suppose that a state were to spend *no money at all on education*. According to the regression line, what would be the predicted annual teacher salary in that state? Is this prediction trustworthy? Why or why not?

(e) Expenditure varies from state to state, and teacher salaries vary from state to state. About what fraction of the variation in teacher salaries is accounted for by variation in expenditure on students?

7. We will ask a series of research questions involving the relationship between two variables. We would like to investigate these questions using numerical techniques. For each question, we indicate which variable should be considered explanatory (EXP) and which variable should be considered the response variable (RESP). After the research questions, we will present some blocks of code, numbered 1 through 9. For each research question, check the number for each code-block that constitutes an appropriate way to explore the question numerically. (**Note:** All of the variables reside in a data frame called CARS.)

   First, the research questions:

   (a) when cars are driven at 40 mph, does distance required to come to a full stop tend to be greater for heavier cars than for lighter ones? EXP = weight of car, in pounds. RESP = stopping distance, in feet. (You may assume that a scatterplot of the data shows a linear relationship.) ◯ 1 ◯ 2 ◯ 3 ◯ 4 ◯ 5 ◯ 6 ◯ 7 ◯ 8 ◯ 9

   (b) Which type of car tends to break down the most: a truck, a passenger sedan, or a sports car? EXP = Type of car. RESP = Number of breakdowns in the past five years. ◯ 1 ◯ 2 ◯ 3 ◯ 4 ◯ 5 ◯ 6 ◯ 7 ◯ 8 ◯ 9

   (c) Is there any relationship the type of a car (truck, passenger sedan, sports car) and the sex of the person who owns it? EXP = Sex of owner (male, female). RESP = Type of car. ◯ 1 ◯ 2 ◯ 3 ◯ 4 ◯ 5 ◯ 6 ◯ 7 ◯ 8 ◯ 9

   Here are the blocks of code that you might use for each research question.

```
# Block 1
favstats(~RESP, data = CARS)
# Block 2
favstats(~EXP, data = CARS)
# Block 3
favstats(RESP ~ EXP, data = CARS)
# Block 4
with(CARS, cor(EXP, RESP))
# Block 5
MyModel <- lm(RESP ~ EXP, data = CARS)
summary(MyModel)
# Block 6
ExpResp <- xtabs(~EXP + RESP, data = CARS)
rowPerc(ExpResp)
# Block 7
chisq.testGC(~EXP + RESP, data = CARS)
# Block 8
xtabs(~EXP, data = CARS)
# Block 9
rowPerc(xtabs(~RESP, data = CARS))
```

8. (You will need to use R for this one.) We are interested in the question of who drives faster at Penn State University: guys or gals. Use **favstats** to find the mean and median fastest speed ever driven for females and for males in the **pennstate1** data frame. Write your answers below:

   i. mean fastest speed for the males

   ii. mean fastest speed for the females

   iii. median fastest speed for the males

   iv. median fastest speed for the males

   v. Based on the above figures, who tends to drive faster in this sample of Penn State students: the guys or the gals?