STATISTICS 8, MIDTERM EXAM 1 KEY

NAME: **KEY** (VERSION A3) Seat Number: _____

Last six digits of Student ID#: _____ Circle your Discussion Section: 1 2 3 4

Make sure you have 6 pages. You may use one page of notes (both sides) and a calculator.

Multiple choice questions: There are 16 questions worth 4 points each (16 x 4 pts each = 64 pts). Instructions will be given when those begin on page 4.
Free response questions: Show all work. If you need extra space use the back of the page, but make sure to tell us it's there. Total of 36 points; points for each part of each question are shown with the question.

1. A CBS News Poll conducted between December 17 and 22, 2009 asked a random sample of $n = 563$ married adult Americans, "Would you say that your marriage with your spouse is better, worse or about the same as your parents' marriage?" and 41% said "about the same."

   a. (**4 pts)** What is the conservative margin of error for this poll? Give your answer as a percentage (not a proportion).

   $$\frac{1}{\sqrt{n}} \times 100\% = \frac{1}{\sqrt{563}} \times 100\% = 4.21\%$$ (It's okay if you round it to 4%.)

   b. (**4 pts**) Calculate a conservative 95% confidence interval for the population percentage that would have responded "About the same."

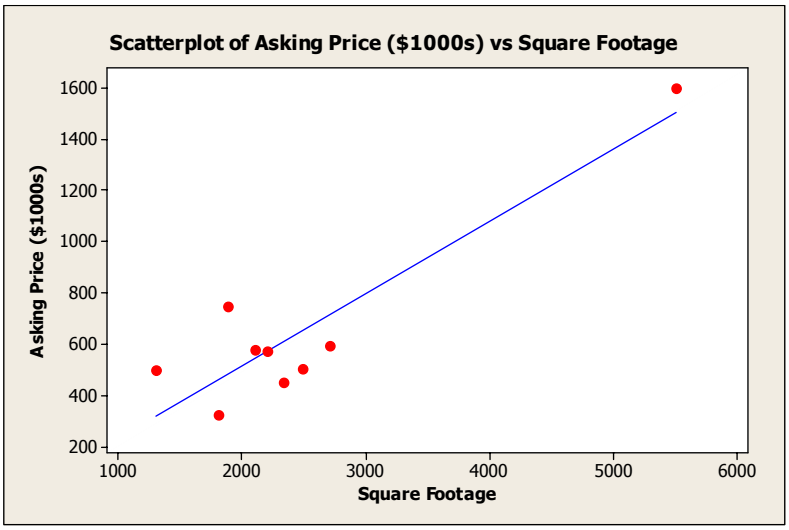
   41% ± 4% or 37% to 45%

   c. (**2 pts**) Based on the interval you found in part (b), could you conclude that in December 2009 less than half of all married adult Americans thought that their own marriage was about the same as their parents' marriage? Explain.

   Yes. The entire interval falls below 50% (half). So we can conclude that the population value is less than 50%.

2. The R Commander output and scatter plot below are based on y = asking price (in thousands of dollars) and x = square feet, for 9 homes in Orange County that were for sale in Spring 2010.
Note: Additional questions about this scenario will be asked in the multiple choice part of the exam.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -48.51517  133.46102  -0.364 0.726950
SqFeet        0.28192    0.04895   5.759 0.000692 ***
```



a. (**3** pts) Write the regression equation to predict asking price based on square feet.

$$\hat{y} = -48.51517 + 0.28192(\textit{square feet})$$

b. (**4 pts**) Predict the asking price for a 2500 square foot house. You can round off the coefficients (intercept and slope) to 2 decimal places.

$$\hat{y} = -48.52 + 0.28(2500) = 651.48$$

Predicted asking price is $651,480

c. (**2 pts**) Write a sentence interpreting the slope in this situation.

The slope of 0.28 means that on average, the asking price (in thousands) goes up by 0.28 for every one inch increase in square feet. This means it goes up $280.

3. In a discussion section (not this quarter) 56 students were asked if they were male or female, and if they had ever been pulled over by an officer. Of the 22 males, 15 said yes. Of the 34 females, 18 said yes.

   a. (**4 pts)** Fill in the contingency table below. Label the rows and columns, and fill in the numbers.

	Been pulled over by an officer?		
	Yes	*No*	Total
Male	15	7	22
Female	18	16	34
Total	33	23	56

   b. (**5 pts**) Find the relative risk of being pulled over for males compared to females *and* write a sentence interpreting the relative risk that someone with no training in statistics would understand.

   Risk for males = 15/22 = .6818

   Risk for females = 18/34 = .5294

   Relative risk = .6818/.5294 = 1.29

   Based on this sample, males are about 1.29 times as likely as females to have been pulled over by an officer.

   c. (**4 pts)** Write the null and alternative hypotheses for a chi-square test for this situation.

   Null hypothesis: There is no relationship between sex and whether or not someone has been pulled over by an officer, in the population represented by these students.
   Alternative hypothesis: There is a relationship between sex and whether or not someone has been pulled over by an officer, in the population represented by these students..

   Note: *You could also write these as:*
   Null: In the population represented by these students, males and females are equally likely to have been pulled over by an officer.
   Alternative: In the population represented by these students, males and females are not equally likely to have been pulled over by an officer.

   d. (**4 pts**) R Commander produced the following results for this situation:

```
X-squared = 1.2819, df = 1, p-value = 0.2575
```

   Use this information to make a conclusion about the hypotheses you wrote in part (c). State your conclusion in statistical terms *and* in the context of this situation.

   The p-value is .2575, which is greater than .05, so do not reject the null hypothesis.
   There is not sufficient evidence to conclude that there is a relationship between sex and whether or not someone has been pulled over by an officer, for the population represented by these students.

MULTIPLE CHOICE

- **You have Exam Version A3. Write this on your Scantron in the space labeled "Test No."**
- **Make sure you write your name on your Scantron sheet.**
- ***Circle* the best answer on this exam paper *and* bubble in the Scantron sheet.**

1. Refer to the scatter plot for house sales, shown in Problem 2 on page 2 of this exam. There is an obvious outlier, which is a house with 5500 square feet and an asking price of $1.6 million. Which of the following is the *most likely* explanation for this point and the most reasonable course of action if the goal is to predict asking price for houses in the 1000 to 3000 square foot range?
   A. It is not clear why the outlier is there, so the house should definitely not be removed from the data.
   B. A mistake was made so the house should be removed from the data.
   C. The house represents natural variability in house sizes and prices and should not be removed because removing it would result in an erroneous estimate of the relationship between size and price, even for houses in the 1000 to 3000 square foot range.
   D. The house in question is a mansion, and thus is not like the bulk of the houses, so the house should not be included in the analysis. A different equation should be found for mansions.

2. Refer (again) to the scatter plot for house sales, shown in Problem 2 on page 2 of this exam. For the data shown, the correlation between square feet and asking price is $r = .91$. If the outlier were to be removed, the correlation would be:
   A. Much lower
   B. Much higher
   C. About the same
   D. There is no way to tell from the information provided.

3. In election years many polls are conducted to estimate what percent of the population is likely to vote for each candidate. Because the true percent can change over time, the polls are not necessarily estimating the same thing every time. Suppose 100 such polls are done and each one is used to find a 95% confidence interval for the percent of the population who plan to vote for a certain candidate at the time of the poll. Which of the following is true about these confidence intervals?
   A. About 95 of them will cover the true percent, and we can determine which ones they are.
   B. About 95 of them will cover the true percent, but there is no way to know which ones they are.
   C. Exactly 95 of them will cover the true percent, and we can determine which ones they are.
   D. Exactly 95 of them will cover the true percent, but there is no way to know which ones they are.

4. Remember that r^2 can be expressed as a proportion or as a percent. Which of the following numbers could *not* be a value for r^2?
   A. 0
   B. −30%
   C. 25%
   D. 0.5

5. Which of the following values of r indicates the strongest linear relationship between x and y?
   A. 0
   B. –0.5
   C. –0.75
   D. +0.6

6. Samples can be chosen using a probability sampling plan and they can be chosen by selecting a simple random sample. Which of the following is true about the relationship between these methods?
   A. Using a probability sampling plan and using a simple random sample are the same thing.
   B. Using a probability sampling plan and using a simple random sample are two completely unrelated methods; neither is a special case of the other.
   C. Using a simple random sample is a special case of using a probability sampling plan.
   D. Using a probability sampling plan is a special case of using a simple random sample.

7. Which of the following is true about the tea and conception example discussed in class?
   A. It was an observational study with tea drinking (or not) as the explanatory variable.
   B. It was an observational study with tea drinking (or not) as the response variable.
   C. It was a randomized experiment with tea drinking (or not) as the explanatory variable.
   D. It was a randomized experiment with tea drinking (or not) as the response variable.

8. According to the website wiki.answers.com, the following are the quartiles and median for the times of 100 swimmers aged 19-29 in the Newport one-mile swim: Q_1 = 24 minutes, median = 26 minutes and 45 seconds, Q_3 = 28 minutes and 21 seconds. About what percent of the swimmers had times in the interval from 24 minutes to 28 minutes and 21 seconds?
   A. 25%
   B. 50%
   C. 75%
   D. 100%

9. Body mass index (BMI) compares height to weight to determine whether an individual is underweight, normal, overweight or obese. Suppose that for a certain age of girls, body mass index values are bell-shaped with a mean of 20 and a standard deviation of 2. A physician would like to warn parents that their daughter may be at risk of obesity if her BMI is in the top 16% of this distribution. What BMI value has about 16% of the distribution above it?
   A. 16
   B. 18
   C. 22
   D. 24

The following scenario applies to Questions 10 and 11: A study of students at a community college found a negative correlation between distance the student commutes to school and grade point average.

10. Which of the following can be concluded from this study?
    A. If a student reduces his commute time his GPA will increase.
    B. Students who commute longer distances have lower GPAs because they have less time to study.
    C. Students who commute longer distances have lower GPAs on average.
    D. Students who commute longer distances have higher GPAs on average.

11. Which of the following could *not* be how this study was done?
    A. It was a survey using a stratified sample of students at the college.
    B. It was an observational study using a random sample of students at the college.
    C. It was an observational study using a convenience sample of students at the college.
    D. It was a randomized experiment using a convenience sample of students at the college.

12. Which of the following could be used in an observational study?
    A. A random sample.
    B. Random assignment to treatments.
    C. Double blind conditions.
    D. All of the above could be used in an observational study.

13. Which of the following is true for a hypothesis test?
    A. With a large sample size even a very strong relationship in the population will result in a p-value > .05.
    B. With a large sample size only a very strong relationship in the population will result in a p-value < .05.
    C. With a small sample size it is easy to obtain a p-value < .05, even if there is no relationship in the population.
    D. With a small sample size it is difficult to obtain a p-value < .05, even if there is a moderate relationship in the population.

14. For which of the following would it be preferable to have a *negative* standardized (z) score?
    A. Your income compared to others born the same year you were born.
    B. The amount you owe on student loans when you graduate (compared to others in your class).
    C. Your test score on a final exam.
    D. None of the above.

15. Which of the following refers to *selection bias*?
    A. Not reaching the individuals selected.
    B. Using biased wording in a question.
    C. Using the wrong sampling frame.
    D. Desire to please the interviewer.

16. Which of the following is *not* true about a case-control study?
    A. The "cases" and "controls" usually represent the different categories of the response variable.
    B. It is difficult to interpret relative risk because of the way the row totals are determined.
    C. It is a special type of randomized experiment.
    D. The "controls" are chosen in a way that helps reduce the impact of confounding variables.