STATISTICS 7 MIDTERM EXAM 1

NAME: KEY ______________ Seat Number: ____________

Student ID#: ____________________ Discussion Section: _______

You may use one page of notes (both sides) and a calculator.

Free response questions: *Show all work or you may not receive any partial credit.* Total of 60 points for free response; 5 points for each part of each question unless shown otherwise.

1. In one of the discussion sections participants were asked to toss a penny into a bowl 20 times in an unimpaired state, and 20 times with an impairment (covering one eye and/or using the non-dominant hand). If the penny landed in the bowl it was a "hit." Results for one student were as follows:

Condition:	**Missed**	**Hit**	**Total**
Impaired	8	12	20
Unimpaired (Baseline)	6	14	20

(1 point each) Fill in the blanks for this student's data:

Risk of missing when Impaired: _8/20 = .4___ Baseline Risk: __6/20 = .3_ Relative Risk: __.4/.3 = 1.33__

Odds of missing to hitting when Impaired: _8/12 = 2/3___ Odds ratio: $\frac{8/12}{6/14} = 14/9 = 1.56$

2. A survey of a random sample of 1186 adults in the United States included the following two questions:
Q1: Do you favor or oppose the death penalty for those convicted of murder?
Q2: Do you think the use of marijuana should be made legal or not?
Results:

	Marijuana legal	**Marijuana not legal**	**Total**
Favor death penalty	313	461	774
Oppose death penalty	165	247	412
Total	478	708	1186

a. What is the overall (conservative) margin of error for this survey? (Round to 2 decimal places.)

$$\frac{1}{\sqrt{1186}} = .029 \text{, } \textit{round to .03 or 3\%}$$

b. Find a 95% confidence interval for the *percent* of the population that thinks marijuana should be legal.

$$\frac{478}{1186} = .403 \text{, } \textit{so the interval is 40.3\% ± 3\% or 37.3\% to 43.3\%}$$

c. State the null hypothesis of interest in this situation.

There is no relationship between opinion on the death penalty and opinion on legalization of marijuana for the population of US adults.

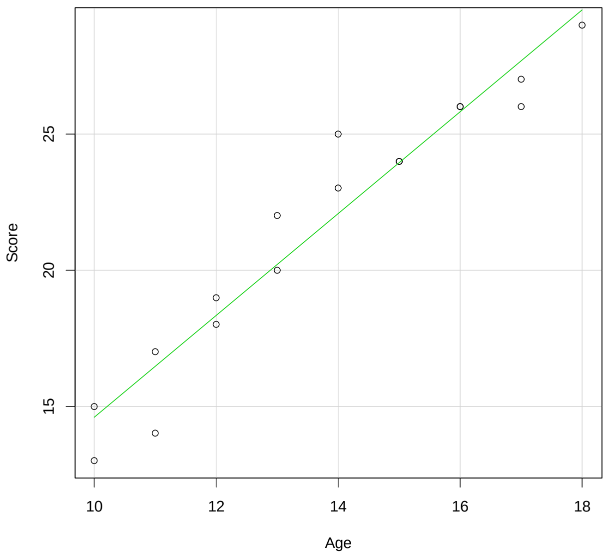
d. The R Commander results for a chi-square test for this situation are

```
X-squared = 0.0171, df = 1, p-value = 0.8961
```

Is the relationship statistically significant? Justify your answer using the output provided.

No. Justification can either be that the p-value of .8961 > .05, or that the chi-square value of .0171 < 3.84.

3. The scatter plot below shows age versus reading comprehension score for students taking a summer reading program at a local library. Regression results from R Commander are shown on the right.




```
Coefficients:
               Estimate
(Intercept)   -4.0753
Age             1.8687

Multiple R-squared: 0.9327, Adjusted R-
squared: 0.9282
```

a. Fill in the blanks for the slope and intercept in the regression equation. Round them to 2 decimal places:

$$\hat{y} = -4.08 + 1.87x$$

a. [*Answer part a in the box above, to the right of the graph.*]
b. What is the correlation between age and reading comprehension score? (Round to 3 decimal places.)

$\sqrt{.9327} = .966$. *(The correlation is positive because the slope is positive.)*

c. What is the predicted reading comprehension score for a 10-year old?

$$\hat{y} = -4.08 + 1.87x = -4.08 + 1.87(10) = -4.08 + 18.7 = 14.62$$

d. One of the 10-year old children had a reading comprehension score of 13. What is the residual for that child?

$$y - \hat{y} = 13 - 14.62 = -1.62$$

e. Does the intercept have a useful interpretation in this situation? If so, give the interpretation. If not, explain why not.

No. It would be the predicted reading score for someone with age = 0, but that does not make sense.

f. Does the slope have a useful interpretation in this situation? If so, give the interpretation. If not, explain why not.

Yes. The slope of 1.87 means that on average, reading comprehension score is predicted to go up by 1.87 points for each year of increase in age (for children aged 10 to 18).

4. (1 pt each) The weights of the adult population of a particular breed of dog are bell-shaped with a mean of 50 pounds and standard deviation of 4 pounds. What is the z-score corresponding to each of the following?

   a. A dog that weighs 60 pounds.

   $$z = \frac{60 - 50}{4} = 2.5$$

   b. A dog for which 50% of the population weighs more and 50% weighs less.

   z = 0

   c. A dog for which 16% of the population weighs more and 84% weighs less.

   z = +1 (based on the Empirical Rule, 68% of z-scores are between –1 and +1, so 16% are above +1 and 16% are below –1

   d. A dog whose weight is at the 2.5^{th} percentile.

   z = –2 (similar argument to part c; 95% are between –2 and +2, so 2.5% are below –2)

   e. A dog that weighs 3 standard deviations less than the mean.

   z = –3

Multiple choice questions: There are 10 questions worth 4 points each (40 points total). *Circle your answer.*

The following scenario is for Questions 1 to 4: A study was done to see if GPA is related to where students prefer to sit in a classroom. Students were asked where they prefer to sit in a large class (front, middle, back) and to report their GPA. It was found that the mean GPA was highest for students who prefer to sit in the front of the room and lowest for students who prefer to sit in the back of the room. However, it was also found that students who prefer the back of the room drink more alcohol than the other students, on average. All of these results were found to hold equally for males and females.

1. What is the response variable for this study?
   A. Where a student prefers to sit
   B. The student's GPA
   C. How much alcohol the student drinks
   D. Gender

2. What is the explanatory variable for this study?
   A. Where a student prefers to sit
   B. The student's GPA
   C. How much alcohol the student drinks
   D. Gender

3. Which of the following is clearly a confounding variable in this study?
   A. Where a student prefers to sit
   B. The student's GPA
   C. How much alcohol the student drinks
   D. Gender

4. Can it be concluded that sitting in the front of the room causes students to have higher GPAs?
   A. Yes, because this was clearly a randomized experiment
   B. No, because this was clearly an observational study
   C. Yes if the students were a random sample of all students, and no otherwise.
   D. It depends on whether this was a randomized experiment or an observational study, and it isn't clear which it was.

5. The correlation between age and a measure of strength for people between the ages of 20 and 50 is –0.4. Which of the following statements is correct?
   A. Age explains 40% of the variability in strength.
   B. Age explains –40% of the variability in strength.
   C. Age explains 16% of the variability in strength.
   D. Age explains –16% of the variability in strength.

6. In a large data set an analyst noticed that the values were approximately bell-shaped and ranged from 30 to 90 except that there was a large outlier of 900. The analyst realized that it should have been 90 instead, and changed it. Which of the following would *not* be affected by this change?
   A. Mean
   B. Standard deviation
   C. Range
   D. Interquartile range

The following scenario is for Questions 7 and 8: The SAT and ACT are two standardized tests used to predict College GPA. To determine which is best, a random sample of college students who had taken both the SAT and ACT was selected. A regression analysis relating y = College GPA to x = SAT score found SSTO = 7.9 and SSR = 4.0.

7. Which of the following would definitely be true for a regression analysis relating y = College GPA to x = ACT score using this same data set?
   A. SSR = 4.0
   B. SSTO = 7.9
   C. Both SSR = 4.0 and SSTO = 7.9
   D. Neither SSR = 4.0 or SSTO = 7.9.

8. For a regression analysis of y = College GPA and x = ACT score using this same data set, which of the following results would indicate that ACT scores are a better predictor of College GPA than are SAT scores? (Remember that for y = College GPA and x = SAT score, SSTO was 7.9 and SSR was 4.0.)
   A. SSTO > 7.9
   B. SSTO < 7.9
   C. SSR > 4.0
   D. SSR < 4.0

9. For chi-square tests, the alternative hypothesis is a statement about:
   A. A relationship in the population
   B. A relationship in the sample
   C. Statistical significance in the population
   D. Statistical significance in the sample

10. For each of the 40 years from 1961 to 2000, x = number of blood donors and y = number of appendicitis operations in the United States were recorded. A scatter plot of these 40 pairs of data revealed a moderately strong linear association. The most likely explanation for this association is:
   A. Donating blood is causing appendicitis in some people.
   B. Having had an operation leads people to want to donate blood.
   C. Larger cities have more hospitals, so there is more opportunity to donate blood and also more opportunity to have an operation in large cities than in small ones.
   D. The US population size has increased over those 40 years, so the association is explained by a third variable, the population in that year.