Math 130: Intro Stats
Final Exam
December 19, 2012

Directions: Before you leave, you must turn in both this exam sheet and any statistical tables. If not, you will receive a significant grade reduction. You are allowed to use a calculator and a two-sided sheet of notes for this exam. All cell phones, PDAs, iPods, laptops, etc, should be turned off and put out of sight. You may not discuss the exam with anyone but me. In total, this exam is worth 200 points. You have the entire period to complete this exam.

Part I – Multiple Choice: There is only ONE correct response per question. Each question is worth 5 points. There are a total of 10 questions for a combined total of 50 points. Clearly circle or write your answer in front of each question.

<table>
<thead>
<tr>
<th>Part I</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td></td>
</tr>
<tr>
<td>Possible Points</td>
<td>50</td>
</tr>
</tbody>
</table>

Part II – Free Response: You must show all work in order to receive full credit. Each question is worth a different amount of points and this value is noted in the table below. There are a total of 8 questions with multiple parts for a combined total of 150 points.

<table>
<thead>
<tr>
<th>Part II</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Possible Points</td>
<td>12</td>
<td>13</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>13</td>
<td>12</td>
<td>25</td>
<td>150</td>
</tr>
</tbody>
</table>

Here is my suggestion:
Read all questions before beginning and try to complete the ones you know best first.

GOOD LUCK!!!
PART I: MULTIPLE CHOICE. Choose the one alternative that best completes the statement or answers the question.

1. Olivia wants to learn a foreign language. To get an idea of how satisfied other students were after taking a foreign language course, she decides to take a random sample of 20 students. If she randomly selects 5 students from French, 5 from German, 5 from Spanish, and 5 from Chinese, then what sampling method did she use?
   A) simple random sampling  
   B) stratified random sampling  
   C) cluster sampling  
   D) sampling with replacement  
   E) systematic sampling

2. True or False: In hypothesis testing, a result is statistically significant if it is unlikely that chance alone can explain the observed results.
   A) True  
   B) False

3. Absorption rates into the body are important considerations when manufacturing a generic version of a brand-name drug. A pharmacist reads that the absorption rate into the body of a new generic drug (G) is the same as its brand-name counterpart (B). She has a research friend of hers run a small experiment to test $H_0: \mu_G - \mu_B = 0$ against $H_A: \mu_G - \mu_B \neq 0$. Which of the following would be a Type I error?
   A) Deciding that the absorption rates are the same, when in fact they are.  
   B) Deciding that the absorption rates are different, when in fact they are.  
   C) Deciding that the absorption rates are the same, when in fact they are not.  
   D) The researcher cannot make a Type I error, since he has run an experiment.  
   E) Deciding that the absorption rates are different, when in fact they are not.

4. A sports fan selected a random sample of 100 games from each of the NBA, the NFL, the NHL, and Major League Baseball to see if overtimes (or extra innings) are equally likely to occur in all four sports? Which inference procedure would you use to analyze this situation?
   A) Goodness of Fit  
   B) Homogeneity test  
   C) Chi-square of Independence  
   D) Analysis of Variance  
   E) Regression

5. Does regular exercise decrease the risk of cancer? A researcher finds 200 women over 50 who exercise regularly, pairs each with a woman who has a similar medical history but does not exercise, then follows the subjects for 10 years to see which group develops more cancer.
   A) This is a prospective study.  
   B) This is a randomized experiment.  
   C) This is a retrospective study.  
   D) This is a survey.  
   E) None of the above.
6. Which statement correctly compares t-models to the normal models?
   I. t distributions are also mound shaped and symmetric.
   II. t distributions are more spread out than the normal distribution.
   III. As degrees of freedom increase, the variance of t distributions becomes larger.
   A) I only
   B) II only
   C) I and II only
   D) II and III only
   E) I, II, and III

7. What statement is true about both $p$ and $\bar{y}$?
   A) They are both parameters
   B) They are both statistics
   C) They are both symbols pertaining to means
   D) $\bar{y}$ is a statistic and $p$ is a parameter
   E) $\bar{y}$ is a parameter and $p$ is a statistic

8. In an Intro Stats class, 57% of students eat breakfast in the morning and 80% of students floss their teeth. Forty-six percent of students eat breakfast and also floss their teeth. What is the probability that a student from this class eats breakfast but does not floss their teeth?
   A) 9%
   B) 11%
   C) 34%
   D) 57%
   E) 91%

9. Which statement is not true about confidence intervals?
   A) A confidence interval is an interval of values computed from sample data that is likely to include the true population parameter value.
   B) An approximate formula for a 95% confidence interval is sample estimate $\pm$ margin of error.
   C) A confidence interval between 20% and 40% means that the population proportion definitely lies between 20% and 40%.
   D) A 99% confidence interval procedure has a higher probability of producing intervals that will include the population parameter than a 95% confidence interval procedure.
   E) Confidence intervals are (by definition) statistical inference procedures.

10. Which of the following correlation values indicates the strongest linear relationship between two quantitative variables?
    A) $r = -0.65$
    B) $r = -0.30$
    C) $r = 0.00$
    D) $r = 0.11$
    E) $r = 0.60$
PART II: FREE RESPONSE. Write the word or phrase that best completes each statement or answers the question.

President Martin in her second year at the College would like to know more about Amherst students and has formed a committee of twenty nine to analyze a campus-wise online survey recently done via SurveyMonkey.com. The sample consists of responses from n = 184 randomly selected students and is believed to be representative of the student body at Amherst.

The Independence Assumption is satisfied, so you do NOT need to worry about checking the randomization condition and the 10% condition in questions below. (You may, however, need to check other conditions when necessary.)

1. Three committee members, Danny, Hayley, and Sara, would like to know if the percentage of varsity athletes at Amherst is as the advertised figure, 32%. They check with the survey results and find that among 184 students in the survey, 50 are varsity athletes.

   A. Based on the survey results, construct and interpret a 95% confidence interval for the proportion of varsity athletes at Amherst College.

   B. Explain the meaning of ”95% confidence” in Part A.
2. Three committee members, Nick, Greg, and Brad, are very interested in understanding the personal lifestyle choices made by varsity athletes at the College. In particular, they noticed that a large percentage of varsity athletes in the sample \((28/50 = 56\%)\) use Crest toothpaste, while \(12/50 = 24\%\) choose Colgate and \(10/50 = 20\%\) choose others. Based on the toothpaste industry market report, the market shares of Colgate versus Crest are 34.2% and 31.7%, respectively. They would like to know if Amherst varsity athletes’ choices on toothpaste match with the national trend.

A. What is the appropriate analysis to perform (be specific) and state appropriate hypotheses.

   Analysis:

   Null:

   Alternative:

B. Complete the analysis and state your conclusion carefully.
3. Another three committee members, Eric, John, and Jonathan, would also like to investigate the social lives of varsity students at the College. In particular, they are wondering if there is an association between students’ athletic status and whether or not they live in the Socials (Stone/Crossett/Pond/Coolidge).

A. What is the appropriate analysis to perform (be specific) and state appropriate hypotheses.

Analysis:

Null:

Alternative:

B. The two-way table below summarizes the result from the survey. Observed (Expected) is the table setup. Use the observed counts to answer questions in this part.

<table>
<thead>
<tr>
<th>Athletic Status \ Socials</th>
<th>YES</th>
<th>NO</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Varsity Athlete</td>
<td>18 (9.24)</td>
<td>32 (40.76)</td>
<td>50</td>
</tr>
<tr>
<td>Non-Varsity Athlete</td>
<td>6 (     )</td>
<td>34 (     )</td>
<td>40</td>
</tr>
<tr>
<td>Non-Athlete</td>
<td>10 (     )</td>
<td>84 (76.63)</td>
<td>94</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>150</td>
<td>184</td>
</tr>
</tbody>
</table>

Do you think athletic status is independent of whether or not students live in the socials? Please explain. (Do NOT do any inference here.)

C. Some expected counts for inference are missing in the table. Compute and fill in those missing expected counts in the parentheses.

D. How many degrees of freedom?
E. When these three members present their proposal to the committee, another member in the committee point out a potential problem in conducting the above analysis. Based on the two-way table given below, what would be the potential problem of including all 184 participants in the analysis?

<table>
<thead>
<tr>
<th>Live in the socials</th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>7</td>
<td>20</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>No</td>
<td>33</td>
<td>27</td>
<td>40</td>
<td>50</td>
</tr>
</tbody>
</table>

F. After fixing the potential problem, the updated two-way table below summaries the corresponding results. *Observed (Expected)[Chi-square component]* is the table setup.

<table>
<thead>
<tr>
<th>Athletic Status \ Socials</th>
<th>YES</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Varsity Athlete</td>
<td>18 ( 8.95 )</td>
<td>17 ( 26.05 )</td>
</tr>
<tr>
<td>Non-Varsity Athlete</td>
<td>6 ( 6.9 )</td>
<td>21 ( 20.1 )</td>
</tr>
<tr>
<td>Non-Athlete</td>
<td>10 ( 18.15 )</td>
<td>61 ( 52.85 )</td>
</tr>
</tbody>
</table>

Assume the assumptions for the test are all fine. The test statistic works out to be 17.38 with a p-value equal to .000168. State your complete conclusion in context.
4. Similar to the aforementioned members, Kristin, Wyatt, Lauren, and Andrew are eager to learn more about the Amherst varsity student-athletes. One of their interests is to study if students’ athletic status affects the average number of hours they sleep per night. Perform an ANOVA to look for differences in the mean number of sleeping hours per night for students in three different groups: Varsity Athlete, Non-Varsity (Club) Athlete, and Non-Athlete.

A. State appropriate hypotheses.

B. What assumptions need to hold in order for the ANOVA to be valid?

Assuming the assumptions all hold, use the partial output from R below to complete the test and multiple comparisons (if appropriate).

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Varsity</td>
<td>2</td>
<td>17.805</td>
<td>8.902</td>
<td>?????</td>
</tr>
<tr>
<td>Residuals</td>
<td>181</td>
<td>216.103</td>
<td>1.194</td>
<td></td>
</tr>
</tbody>
</table>

Simultaneous Confidence Intervals
Multiple Comparisons of Means: Tukey Contrasts
95% family-wise confidence level

Linear Hypotheses:

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>lwr</th>
<th>upr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non Athlete - Club Athlete == 0</td>
<td>-0.53351</td>
<td>-1.02010</td>
<td>-0.04692</td>
</tr>
<tr>
<td>Varsity Athlete - Club Athlete == 0</td>
<td>0.14500</td>
<td>-0.40178</td>
<td>0.69178</td>
</tr>
<tr>
<td>Varsity Athlete - Non Athlete == 0</td>
<td>0.67851</td>
<td>0.22735</td>
<td>1.12968</td>
</tr>
</tbody>
</table>
C. What is the missing value of the test statistic?

D. What is the P-value of the hypothesis test?

E. What is your conclusion?

F. What are the results of multiple comparisons?

G. The normality assumption is somehow questionable given the sample sizes of these three groups (see below). Thus, it’s probably a good idea to run a non-parametric test to verify the above conclusion. Name a nonparametric alternative to this test procedure.

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Club Athlete</td>
<td>7.07500</td>
<td>0.9510453</td>
<td>40</td>
</tr>
<tr>
<td>Non Athlete</td>
<td>6.54149</td>
<td>1.1834349</td>
<td>94</td>
</tr>
<tr>
<td>Varsity Athlete</td>
<td>7.22000</td>
<td>1.0159945</td>
<td>50</td>
</tr>
</tbody>
</table>
5. On the other hand, three committee members, Tyler, Katie, and Arne have some very different curiosities. For instance, they use Rcmdr to obtain the descriptive statistics below for the variable, Water, the number of cups of water drinking in a day.

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>0%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>100%</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>5.722826</td>
<td>3.898202</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>7.25</td>
<td>30</td>
<td>184</td>
</tr>
</tbody>
</table>

A. What is the shape of distribution/histogram? (symmetric/ right-skewed/ left-skewed.)

B. Are there any outliers? Explain.

C. When describing center and spread, which set of summary statistics should we use? Why?

They would further like to check if students drink more cups of water than coffee in a day. Two tests are performed and the corresponding Rcmdr outputs are given below:

**Output (I): Two Sample t-test**

data: Cups by Type  
t = 15.8692, df = 231.766, p-value < 2.2e-16  
alternative hypothesis: true difference in means is greater than 0  
sample estimates: mean in group Water  
mean in group Coffee  
5.7228261  
0.8627717

**Output (II): Paired t-test**

data: Survey$Water and Survey$Coffee  
t = 16.9518, df = 183, p-value < 2.2e-16  
alternative hypothesis: true difference in means is greater than 0  
sample estimates: mean of the differences  
4.860054
D. Indicate what inference procedure/test you would use to answer this question. Explain.

E. Write appropriate hypotheses.

F. Assuming all assumptions are satisfied, use the correct Rcmdr output above to complete the test procedure. Specify the value of the test statistic and the P-value and make sure to state your conclusion in context. Specify the significance level you use in this part.

G. What confidence level is associated with the test above? Calculate the corresponding confidence interval with this confidence level. Besides the information given in the outputs above, you may also need some of the following:

\[ \text{St. Dev.('Water')} = 3.90, \text{ St. Dev.('Coffee')} = 1.436, \text{ St. Dev.('Water - Coffee')} = 3.89; \]
\[ \text{N('Water')} = \text{N('Coffee')} = 184. \]
6. Another three members, David, Christina, and James are more into the comparison between pre-med and non-premed students. From the preliminary analysis (shown below), it seems that the proportion of pre-med students who are night-only people is greater than the proportion of non-premed students who are night-only. They would like to know if this observed difference is significant or not.

<table>
<thead>
<tr>
<th>Night people?</th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-med</td>
<td>83</td>
<td>70</td>
</tr>
<tr>
<td>Non-premed</td>
<td>7</td>
<td>24</td>
</tr>
</tbody>
</table>

A. Test an appropriate hypothesis and state your conclusion.

B. Explain what your P-value means in the context of the problem.
7. It’s well-known that the left hemisphere of the brain deals predominantly with analytical thinking while the right hemisphere controls creative thinking. Thus four female members, Noelle, Emma, Maggie, and Erika, would like to examine the association between students’ handedness and their hobbies. Suppose that 10% of Amherst students are left-handed.

A. What is the probability that one out of five randomly selected Amherst students is left-handed?

B. In a random sample of 184 students, what is the probability that no more than 16 students are left-handed?

8. Unsurprisingly, several committee members, Claudia, Danielle, Ashley, Daniel (Chun), Marco and Daniel (Park) are very interested in students’ academic life and performances. Some of them decide to take a closer look at the relationship between the number of nights they go out per week and the number of hours they spend doing schoolwork per week. The resulting Rcmdr output is attached below.

```
Call: lm(formula = Schoolwork ~ GoOut, data = Survey)

Coefficients:  
             Estimate Std. Error t value Pr(>|t|)  
(Intercept)  26.2737    1.5805 16.623  < 2e-16 ***  
GoOut         -2.4047    0.9003  -2.671 0.00825  **  

Residual standard error: 11.18 on 182 degrees of freedom  
Multiple R-squared: 0.03772,  Adjusted R-squared: 0.03243  
F-statistic: 7.135 on 1 and 182 DF,  p-value: 0.008247
```
The plots attached below are the scatterplot, the residuals plot (residuals vs. fitted values), and a histogram of the residuals (in order).

A. What is the value of the correlation coefficient?

B. Based on the Rcmdr output, what is the response variable and what is the explanatory variable? Also, what is the equation of the regression line?

C. For a student who usually goes out once per week and on average spend 20 hours on schoolwork each week, what is his/her residual?
D. Is there an association between these two quantitative variables? Write appropriate hypotheses, check and explain if the assumptions for regression satisfied, provide the corresponding test statistic and P-value, and then state your conclusion about the association.

E. Interpret the value of the sample slope in the context of the problem. Then, create a 95% confidence interval for the true slope and explain in context what your interval means.

F. Is the variable ‘GoOut’ a good predictor for the variable ‘Schoolwork’? Explain and use some related statistics to support your answer.