Math 17: Intro Stats
Final Exam
May 9, 2011

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>15</td>
<td>25</td>
<td>20</td>
<td>25</td>
<td>25</td>
<td>12</td>
<td>12</td>
<td>16</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. Based on her past experience, a professor knows that the probability distribution for \( X = \) number of students who come to her office hours on Friday is given below:

<table>
<thead>
<tr>
<th>( k )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(X = k) )</td>
<td>0.10</td>
<td>0.20</td>
<td>0.50</td>
<td>0.15</td>
<td>0.05</td>
</tr>
</tbody>
</table>

What is the probability that at least two students come to office hours on Friday?
A) 0.30  B) 0.50  C) 0.70  D) 0.80  E) 0.90

2. Sandra wants to learn an Asian language. To get an idea of how satisfied other students were after taking an Asian language course, she decides to take a random sample of 40 students. Suppose that she randomly select 10 students from Chinese, 10 from Japanese, 10 from Korean, and 10 from Indian. 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

3. Which of the following statements is true for tossing a fair coin (i.e. \( \Pr(\text{Head}) = 0.5 \)) if the first 100 tosses of the coin result in 100 heads (assuming independence)?
A) The chance that the next toss will be heads is almost certain  
B) The chance that the next toss will be tails is almost certain  
C) The chance that the next toss will be heads is near 0  
D) The chance that the next toss will be tails is near 1  
E) The chance that the next toss will be heads is 1/2

4. An ice cream stand reports that 12% of the cones they sell are “jumbo” size. You want to see what a “jumbo” cone looks like, so you stand and watch the sales for a while. What is the probability there is exactly one jumbo among the first six cones sold?
A) 6%  B) 12%  C) 38%  D) 53%  E) 84%

5. In a survey, students are asked how many hours they study in a typical week. A five-number summary of the responses is: 2, 9, 14, 20, 60. Which interval describes the number of hours spent studying in a typical week for about 25% of the students sampled?
A) 2 to 14  B) 2 to 60  C) 9 to 20  D) 14 to 60  E) 20 to 60
6. A supermarket claims that their checkout scanners correctly price 99.8% of the items sold. How many items would you expect to buy, on average, to find one that scans incorrectly?  
A) 2  
B) 99.8  
C) 200  
D) 500  
E) 998

7. Which of the following is true about confidence intervals?  
I. A confidence interval is an interval of values computed from sample data that is likely to include the true population parameter value.  
II. The probability that a 95% CI for a population mean contains the population mean is 0.95.  
III. The probability that a 95% CI for a population mean contains the sample mean is 1.  
A) I only  
B) II only  
C) III only  
D) I and II  
E) I and III

8. Which of the following is true about hypothesis testing?  
I. A significance test gives a p-value of 0.03. Using this information, we can say that the probability that the null hypothesis is true is 0.03.  
II. A test which was significant at the alpha-level = 0.05, will also be significant at the alpha-level = 0.01.  
III. A decrease in power means that your sample size may have decreased.  
A) I only  
B) II only  
C) III only  
D) I and II  
E) I and III

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

10. What are pie charts useful for summarizing?  
A) A large number of quantitative variables  
B) The relationship between an explanatory variable and a response variable  
C) The five-number summary  
D) A single categorical variable, if there are not too many categories  
E) The margin of error for certain proportions
PART II: FREE RESPONSE. Write the word or phrase that best completes each statement or answers the question.

President Marx, in his last year of presidency, would like to know more about students at Amherst and has formed a committee of thirty one to analyze a campus-wise online survey recently done via SurveyMonkey.com. The sample consists of responses from n = 147 randomly selected students and is believed to be representative of the student body at Amherst.

The Independence Assumption is satisfied, so do NOT worry about checking the randomization condition and the 10% condition in Q1 – Q8. (You DO need to check other conditions when doing inference below.)

1. Many committee members are very interested in the lives of athletes on Amherst campus. Nick (Sorrentino), Harriet, and Nolen have found that among 147 students in the survey, forty-four participate in varsity sports, while the other 103 students don’t. Sheng, Nick (Egen) and Koby, another three committee members, would then like to further investigate if varsity athletes, on average, have lower stress level than non-athletes.

   A. Indicate what inference procedure you would use to investigate this question. Why?

   B. Write appropriate hypotheses.

   C. Assume the assumptions for the test are met. Partial Rcmdr output is given below. Use the output to complete the test procedure at a .10 significance level.

```
data: StressLevel by VarsityAthlete
t = -2.7537, df = 69.44, p-value = ---------
alternative hypothesis: true difference in means is less than 0
sample estimates:     mean in group Yes mean in group No
                  5.579545            6.684466
```

P-value:

Conclusion:
D. Three committee members, Carolyn, Marithe and Laura are more interested in knowing the average stress level of all Amherst College students. They knew that this average should be somewhere between 5.579545 and 6.684466, on a scale of 1 to 10. Use the Rcmdr output and information given on page 3 to find out this overall average.

2. Speaking of the stress level, three committee members, James, Esther, and Josef, has a theory. They believe that people who sleep more tend to have lower stress level and would like to know if the number of sleeping hours per night is a good predictor of the stress level (units). The plots attached below are the scatterplot, the residuals plot, and a histogram of the residuals (in order), along with the regression analysis for the data.

```
Call: lm(formula = StressLevel ~ SleepHr, data = Survey2011)
Coefficients:
            Estimate Std. Error  t value Pr(>|t|)
(Intercept)   12.1169     1.0157  11.9295   <2e-16 ***
SleepHr       -0.8480     0.1476  -5.7436  5.23e-08 ***
```

Residual standard error: 1.921 on 145 degrees of freedom
Multiple R-squared: 0.1854, Adjusted R-squared: 0.1798
F-statistic: 33 on 1 and 145 DF, p-value: 5.234e-08

A. What is the value of the correlation coefficient?

B. Based on the output, what is the equation of the regression line?

C. For a student who sleeps 7 hours per day and reports a stress level of 5 (units), what is his/her residual?
D. Is there an association between hours sleeping per night and the stress level? 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. (If any assumption looks questionable, name it and proceed with caution.)

E. Create a 95% confidence interval for the true slope and explain in context what your interval means.

F. Is the number of sleeping hours per night a good predictor of the stress level? Explain and use statistics to support your answer.
3. Unsurprisingly, many committee members would like to know how well Amherst students perform in academia. Three of them, Peter, Kelley, and Doug, use Rcmdr to obtain the descriptive statistics for the variable, GPA. The output is shown below.

<table>
<thead>
<tr>
<th>mean</th>
<th>sd</th>
<th>0%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>100%</th>
<th>n</th>
<th>NA</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPA</td>
<td>3.555175</td>
<td>0.2985053</td>
<td>2.7</td>
<td>3.4</td>
<td>3.6</td>
<td>3.77</td>
<td>4</td>
<td>143</td>
</tr>
</tbody>
</table>

A. What is the shape of the corresponding 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? Mean and standard deviation, or median and IQR? Why?

To get a whole picture of all students’ academic performance, they decide to call the Registrar and are informed that a Normal model with a mean of 3.56 and a standard deviation of 0.15 is appropriate for Amherst students’ GPAs.

D. What percent of Amherst students achieve a GPA greater than 3.80?
E. What is the probability that the average GPA of 31 randomly selected students will be less than 3.50?

4. Also interested in the comparison of athletes versus non-athletes, three committee members, Alex, Eric and Angelina investigate the average number of studying hours per week. They are 95% confident that, compared to athlete students, non-athlete students on average study between 2.97 and 9.77 hours more per week. As a similar investigation, they would like to perform a one-way ANOVA to look for differences in the mean number of hours studied in the library per week among different class years. Use an α-level = .10.

A. State appropriate hypotheses.

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

Assuming the assumptions hold, use the partial output from R below to complete the test.

<table>
<thead>
<tr>
<th></th>
<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>ClassYear</td>
<td>3</td>
<td>603.3</td>
<td>201.1</td>
<td>???</td>
<td>0.01988 *</td>
</tr>
<tr>
<td>Residuals</td>
<td>143</td>
<td>8490.1</td>
<td>59.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

C. What is the missing value of the test statistic?
D. What is your conclusion from this test?

E. Interpret the results of multiple comparisons.

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>2012 - 2011 == 0</td>
<td>3.1761</td>
<td>-1.7875</td>
<td>8.1398</td>
</tr>
<tr>
<td>2013 - 2011 == 0</td>
<td>-1.6199</td>
<td>-6.4519</td>
<td>3.2122</td>
</tr>
<tr>
<td>2014 - 2011 == 0</td>
<td>-1.7719</td>
<td>-6.6793</td>
<td>3.1355</td>
</tr>
<tr>
<td>2013 - 2012 == 0</td>
<td>-4.7960</td>
<td>-9.3085</td>
<td>0.2835</td>
</tr>
<tr>
<td>2014 - 2012 == 0</td>
<td>-4.9480</td>
<td>-9.5411</td>
<td>-0.3550</td>
</tr>
<tr>
<td>2014 - 2013 == 0</td>
<td>-0.1520</td>
<td>-4.6025</td>
<td>4.2985</td>
</tr>
</tbody>
</table>

F. The homoskedasticity assumption is in fact questionable, so it’s probably better to run a non-parametric test to verify the above conclusion. First use the Rcmdr output below to explain why this assumption might be questionable and then 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>2011</td>
<td>6.310345</td>
<td>6.296844</td>
<td>29</td>
</tr>
<tr>
<td>2012</td>
<td>9.486486</td>
<td>12.243817</td>
<td>37</td>
</tr>
<tr>
<td>2013</td>
<td>4.690476</td>
<td>5.731760</td>
<td>42</td>
</tr>
<tr>
<td>2014</td>
<td>4.538462</td>
<td>6.244350</td>
<td>39</td>
</tr>
</tbody>
</table>
5. With a slightly different interest, three committee members, Elliot, Jimmy and Dylan investigate the differences among students from various regions and find that students’ plan after graduation is somehow associated with where they are from at an alpha-level of 0.10. They would like to further investigate if there is also an association between students’ desires to study abroad and where they are from. Use an alpha-level of 0.10.

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

Analysis:

Null:

Alternative:

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

<table>
<thead>
<tr>
<th>Region\ StudyAbroad</th>
<th>YES</th>
<th>NO</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northeast (NE)</td>
<td>35 (35.24)</td>
<td>35 (34.76)</td>
<td>70</td>
</tr>
<tr>
<td>Midwest (MW)</td>
<td>9 (8.56)</td>
<td>8 (8.44)</td>
<td>17</td>
</tr>
<tr>
<td>West</td>
<td>9 (11.58)</td>
<td>14 ( )</td>
<td>23</td>
</tr>
<tr>
<td>South</td>
<td>16 (10.07)</td>
<td>4 (9.93)</td>
<td>20</td>
</tr>
<tr>
<td>Other countries (O.C.)</td>
<td>5 ( )</td>
<td>12 ( )</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td>74 ( )</td>
<td>73 ( )</td>
<td>147</td>
</tr>
</tbody>
</table>

What is the conditional distribution of Region for students planning to study abroad?

NE_______ MW_______ West_______ South_______ O.C._______.

For students who don’t plan to study abroad?

NE_______ MW_______ West_______ South_______ O.C._______.

C. Write one sentence or two to describe differences in the conditional distributions above. Based on those differences, does it appear that the study abroad plan is related to region? (Do not do any inference yet.)
D. Some expected counts for inference are missing in the table. Compute and fill in those missing expected counts in the parentheses.

E. How many degrees of freedom?

F. Assume the assumptions for the test are met. The test statistic works out to be 11.2218 with a P-value equal to .02418. State your complete conclusion in context.

G. Compute the chi-square component of the cell {South\ YES}. Does this cell arouse your suspicion? What additional information is provided by this cell?

6. While the new science center is on its way, some committee members, like Akosua, Joshua, Junyeop, and Mingzu, are curious about the true proportion of the science majors at Amherst. They notice that among 147 students in the survey, 71 are science majors.

A. Based on the survey results, construct and interpret a 90% confidence interval for the proportion of Amherst students who are science majors.
B. Explain the meaning of “90% confidence” in part A.

7. On a different issue, Rachel, Dana, and Jack would like to know if students on financial aid are more likely to have an on-campus job. The results are shown in the table below.

<table>
<thead>
<tr>
<th>Financial Aid \ Job on Campus</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>52</td>
<td>31</td>
</tr>
<tr>
<td>No</td>
<td>26</td>
<td>38</td>
</tr>
</tbody>
</table>

Test an appropriate hypothesis and state your conclusion.
8. As the 2012 election is approaching, three committee members, Eric, Michael, and Jasper, suggest investigating the validity of Amherst’s liberal stereotype. In particular, they would like to answer the question: “Do Amherst students tend to think that other students would rate the success of Obama’s presidency, on a scale of 1 to 10, higher than they do?”

A. Explain why this is an example of paired data.

B. Write appropriate hypotheses (in words and in symbols).

C. Assuming all assumptions are satisfied to perform the test, use the partial Rcmdr output below to complete the test procedure. Make sure to state your conclusion in context.

```
data: ObamaSuccess_Peers and ObamaSuccess_Self
t = 3.7508, df = 113, p-value = 0.0001399
alternative hypothesis: true difference in means ('Peers - Self') is greater than 0
sample estimates: mean of the differences
                 0.7429825
```

D. Explain what the P-value means in the context of the problem.

E. Given your conclusion above, which type of error (Type I or Type II) could be made? Explain this type of error in context.