Stat 111 – Final (3 hours)

Name: ______________________________

You may use a calculator and three 8.5” by 11” sheets of notes (front and back), which you will keep. Please show all your work, including all calculations, and explain your answers. Whenever needed, please round numbers (including intermediate calculations) to the nearest 0.001.

Cell phones and any other electronic devices (aside from your calculator) are not permitted. No interaction of any sort is allowed with your classmates.

Here is the breakdown of problem values:

Conceptual Questions:

1. (8 points): ______________________________________________________

2. (6 points): ______________________________________________________

3. (4 points): _____________________________________________________

4. (4 points): _____________________________________________________

Word Problems:

1. (34 points): _____________________________________________________

2. (11 points): _____________________________________________________

3. (14 points): _____________________________________________________

Total (83 points): ___________________________________________________
I Conceptual Questions

Please answer the following in no more than 3-4 sentences each.

1. (8 points) For each of the following, circle the type of hypothesis test that would be most appropriate to analyze the question at hand. There may be more than one reasonable answer; if so, pick one.

(a) The American National Election Studies (ANES) collects data on voter attitudes and intentions as well as demographic information. In their 2012 data set, among other things they gathered the following information for each survey participant:
   - region: where the survey participant is from (levels: Northeast, North Central, South, and West), and
   - direction: whether the respondent feels things in this country are generally going in the right direction or things have pretty seriously gotten off on the wrong track. (levels: Positive, Negative)

We want to evaluate whether people from different regions have different opinions about the direction of the country. For each combination of region and direction, there were at least 30 survey respondents who were from that region and had that opinion about the direction the country was going.

i. two sample t-test for difference in population means
ii. paired two sample t-test for difference in population means
iii. ANOVA
iv. two sample z-test for difference in population proportions
v. $\chi^2$ test of goodness of fit
vi. $\chi^2$ test of homogeneity or independence

(b) Air quality measurements were collected in a random sample of 25 country capitals in 2013, and then again in the same cities in 2014. We would like to use these data to compare average air quality between the two years.

i. two sample t-test for difference in population means
ii. paired two sample t-test for difference in population means
iii. ANOVA
iv. two sample z-test for difference in population proportions
v. $\chi^2$ test of goodness of fit
vi. $\chi^2$ test of homogeneity or independence
(c) We wish to assess the gender-related salary gap. We take a random sample of men and women in the United States and compare their salaries. Based on an initial plot, the standard deviation of salaries seems to be different for men and women.

i. two sample t-test for difference in population means
ii. paired two sample t-test for difference in population means
iii. ANOVA
iv. two sample z-test for difference in population proportions
v. $\chi^2$ test of goodness of fit
vi. $\chi^2$ test of homogeneity or independence

(d) We want to find out if there is a connection between the type of music a person listens to while studying and how well they do on a memory test. We randomly assign subjects to listen to either hip hop, classical music, or no music. Based on initial plots of the data, the standard deviations of test scores for the three groups are similar.

i. two sample t-test for difference in population means
ii. paired two sample t-test for difference in population means
iii. ANOVA
iv. two sample z-test for difference in population proportions
v. $\chi^2$ test of goodness of fit
vi. $\chi^2$ test of homogeneity or independence
2. (6 points) Pick one of the following three terms. Define the term, give an example, and describe why this concept is important enough to put on a final. Clearly indicate which term you have chosen.

   Confounding          Lurking Variable          Simpson’s Paradox

3. (4 points) Someone comes to you with a data set and a question they want to answer based on it. What do you do before you start fitting models and conducting hypothesis tests? No points for "load the data into R", although I recognize that that can be a major task =)

4. (4 points) I was curious about how much weight the average European swallow could carry, so I captured 42 swallows at random and measured how much weight they could carry. Based on that sample, I calculated a confidence interval. In my analysis, I wrote “I am 95% confident that the average European swallow can carry a weight between 0.5 ounces and 1.2 ounces.” What does that actually mean?
II  Word Problems

1. (34 points) What can we predict about earnings for movies? Let’s predict gross earnings in millions of dollars as a function of the movie’s budget in millions of dollars, and the movie’s genre (in this problem we’ll look at Action, Adventure, or Thriller). The accompanying packet contains some plots and linear model output from R. Use that information to answer the following questions:

   (a) (4 points) Check the assumptions for the linear model. Extra Credit: If any assumptions are violated, suggest something you could do to address those limitations.

   Regardless of your answer to part (a), let’s proceed using the results from this model.

   (b) (2 points) Write down the model equation using symbols like $\beta_0, \beta_1$ for the coefficients.

   (c) (1 point) Write down the model equation again, filling in the estimated coefficients from the R output.
(d) (2 points) In the R output, there is a variable called *genreAdventure* (this should appear in your answers to parts b and c). What is that variable? Be as clear and concrete as you can (I want you to list possible values and explain when that variable takes those values).

(e) (2 points) Based on this model, what is the predicted earnings for an Action movie with a budget of 100 million dollars?

(f) (2 points) Based on this model, what is the predicted earnings for an Adventure movie with a budget of 80 million dollars?

(g) (6 points) Does this model explain a statistically significant amount of variation in movie earnings? State a null and alternative hypothesis in terms of the coefficients you set up in part (b), write down the p-value from the R output, and draw a conclusion in the context of the problem.
(h) (6 points) After accounting for the effects of the movie’s budget, does the movie’s genre explain a statistically significant amount of variation in movie earnings? State a null and alternative hypothesis in terms of the coefficients you set up in part (b), write down the p-value from the R output, and draw a conclusion in the context of the problem.

(i) (6 points) After accounting for the effects of the other variables in the model, is there a statistically significant difference in earnings for Action and Adventure movies? State a null and alternative hypothesis in terms of the coefficients you set up in part (b), write down the p-value for this test from the R output, and draw a conclusion in the context of the problem.
(j) (3 points) Write down a confidence interval for the coefficient in the model describing the relationship between a movie’s budget and its earnings (this is in the R output). Interpret the interval in context.
2. (11 points) The China Health and Nutrition Survey aims to examine the effects of the health, nutrition, and family planning policies and programs implemented by national and local governments. Among other variables, it collects information on number of hours Chinese parents spend taking care of their children under age 6. The side-by-side box plots and overlaid density plots in the supplementary package show the distribution of this variable by educational attainment of the parent. Also provided below is the ANOVA output for comparing average hours across educational attainment categories.

<table>
<thead>
<tr>
<th></th>
<th>Sum Sq</th>
<th>Df</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td>4142.09</td>
<td>4</td>
<td>1035.52</td>
<td>1.26</td>
<td>0.28</td>
</tr>
<tr>
<td>Residuals</td>
<td>653047.83</td>
<td>794</td>
<td>822.48</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) (2 points) Write the null and alternative hypotheses for testing for a difference between the average number of hours spent on child care across educational attainment levels.

(b) (4 points) Check all assumptions for testing this hypothesis using an ANOVA test. Extra credit: If any of the assumptions do not hold with this data set, suggest a strategy for solving that problem.
(c) (2 points) Intuitively, what does the F value in the ANOVA table above measure?

(d) (3 points) Regardless of your answer to part (b), go ahead with conducting the hypothesis test. In context, what is the conclusion of the hypothesis test?
3. (14 points) Microhabitat factors associated with forage and bed sites of barking deer in Hainan Island, China were examined from 2001 to 2002. In this region woods make up 4.8% of the land, cultivated grass plot makes up 14.7%, and deciduous forests makes up 39.6%; the remaining 40.9% of the land is some other habitat type. Of the 426 sites where the deer forage, 4 were categorized as woods, 16 as cultivated grassplot, and 61 as deciduous forests. The table below summarizes these data.

<table>
<thead>
<tr>
<th></th>
<th>Woods</th>
<th>Cultivated grassplot</th>
<th>Deciduous forests</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
<td>16</td>
<td>61</td>
<td>345</td>
<td>426</td>
</tr>
</tbody>
</table>

(a) (2 points) Is the proportion of foraging sites in each habitat type equal to the proportion of land with that habitat type in this region? Write the null and alternative hypotheses for testing if barking deer prefer to forage in certain habitats over others.

(b) (3 points) What hypothesis test will you use to check this assumption? Check if the assumptions and conditions required for this test are satisfied.
(c) (9 points) Regardless of your answer to part (b), perform the mechanics of the hypothesis test. State your conclusion in the context of the problem. You may use the following facts, where $t_3$ is a t distribution with 3 degrees of freedom and $\chi^2_3$ is a $\chi^2$ distribution with 3 degrees of freedom:

If $T \sim t_3$ then $P(T < 1.637) = 0.9$, $P(T < 2.353) = 0.95$, and $P(T < 3.182) = 0.975$
If $T \sim t_4$ then $P(T < 1.533) = 0.9$, $P(T < 2.132) = 0.95$, and $P(T < 2.776) = 0.975$
If $X \sim \chi^2_3$ then $P(X < 6.251) = 0.9$, $P(X < 7.815) = 0.95$, and $P(X < 9.348) = 0.975$
If $X \sim \chi^2_4$ then $P(X < 7.779) = 0.9$, $P(X < 9.488) = 0.95$, and $P(X < 11.143) = 0.975$