Applied Statistics Qualifier Examination  
(Part II of the STAT AREA EXAM)  
June 1, 2016; 11:00AM-1:00PM

Instructions:

(1)  The examination contains 4 Questions. You are to answer 3 out of 4 of them. *** Please only turn in solutions to 3 questions ***

(2)  You may use up to 4 books and 4 class notes, plus your calculator and the statistical tables.

(3)  NO computer, internet, or cell phone is allowed in the exam.

(4)  This is a 2-hour exam due by 1:00 PM.

Please be sure to fill in the appropriate information below:

I am submitting solutions to QUESTIONS _____, _____, and _____ of the applied statistics qualifier examination. Please put your name on every page of your exam solutions, and add page number for solutions to each question individually.

There are __________ pages of written solutions.

Please read the following statement and sign below:
This is to certify that I have taken the applied statistics qualifier and have used no other person as a resource nor have I seen any other student violating this rule.

____________________________________
(Signature)

____________________________________
(Name)
1. A research institute claimed that biostatisticians (BIOS) are paid more than statisticians (STAT) in the drug development industry. Twenty students from each major who graduated in 2014 were surveyed and their salaries, \( x \) (in thousands) were recorded.

<table>
<thead>
<tr>
<th>STATS</th>
<th>39.6, 44.8, 19.7, 14.0, 40.4, 3.58, 41.4, 16.7, 16.6, 8.78, 19.6, 136, 31.9, 37.1, 10.4, 49.9, 9.08, 15.8, 67.6, 29.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIOS</td>
<td>53.1, 104, 12.8, 5.15, 4.93, 148, 15, 45.3, 44.9, 20.7, 54.8, 219, 189, 125, 31.4, 39.8, 17.6, 4.62, 143, 25</td>
</tr>
</tbody>
</table>

(a) Test the research institute's claim at \( \alpha = 0.05 \). State your assumptions and conclusion.

(b) Define High salaries as \( x \geq T \) and Low salaries \( x < T \). The analyst was unsure the best cutoff \( T \), thus he considered three possible values (i) \( T=40 \), (ii) \( T=50 \), and (iii) \( T= \) median of the 40 observations. Using these information, test the research institute's claim at \( \alpha = 0.05 \). State your assumptions and conclusion.

(c) Compare the results from (a) and (b). Identify the advantages and disadvantages of the two approaches. Identify any potential pitfalls of the current study design and propose an alternative sampling scheme and/or analysis strategy to address these pitfalls.

(d) (This part is unrelated to the above) In the following, the *italicized* statement is either True or False. Indicate whether you believe that the statement is True or False, and justify your answer.
Suppose you have observations \((Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)\) with \( n \geq 3 \) and you want to fit the simple linear regression model \( Y_i=\beta_0+\beta_1 x_i+\epsilon_i \). *In these circumstances \( \beta_0 \) will always be estimable.*
Name: ______________________________

2. For a $1 \times J$ two-way contingency table with row and column random variables denoted as $X$ and $Y$, one kind of association measure between $X$ and $Y$, called $\tau$, is calculated based on the so-called variation measure

$$V(Y) = \sum_{j=1}^{J} \pi_{+j} (1 - \pi_{+j}) = 1 - \sum_{j=1}^{J} \pi_{+j}^2$$

a. Show $V(Y)$ is the probability that two independent observations on $Y$ fall in different categories.

b. Show that $V(Y) = 0$ when $\pi_{+j} = 1$ for some $j$ and $V(Y)$ takes maximum value of $(J - 1)/J$ when $\pi_{+j} = 1/J$ for all $j$.

c. For the proportional reduction in variation, show that $E[V(Y|X)] = 1 - \sum_{l} \sum_{j} \frac{\pi_{lj}}{\pi_{l}}$. 
3. Consider the multiple regression model

\[ y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \]

where \( \varepsilon_i \) are i.i.d. \( N(0, \sigma^2) \) random variables, or

\[ Y = X\beta + \varepsilon, \]

where \( Y = (y_1, \ldots, y_n)^T \), \( X = (x_{ij})_{1 \leq i \leq n, 1 \leq j \leq p} \) and \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \). Let \( x_i = (x_{i1}, \ldots, x_{ip})^T \), \( H = X(X^TX)^{-1}X^T \) and \( h_{ii} \) is the \( i^{th} \) diagonal element of \( H \). Let \( \hat{\beta} \) be the least square estimate of \( \beta \) given observations \( (x_j, y_j), 1 \leq j \leq n \) and \( e_i = y_i - x_i^T \hat{\beta} \) be the \( i^{th} \) residual. Suppose the \( i^{th} \) observation \( (x_i, y_i) \) is deleted, we then first get the least square estimate \( \hat{\beta}_{(-i)} \) using \( \{(x_j, y_j); j = 1, \ldots, i-1, i+1, \ldots, n\} \) and then compute the prediction error \( e_{(-i)} = y_i - x_i^T \hat{\beta}_{(-i)} \). Show that

\[ e_{(-i)} = \frac{e_i}{1 - h_{ii}} \]
Name: _______________________

4. A statistician has been hired by a research team to design a study to evaluate curricula to teach students in college science course. There is a measure of the quality of a student’s performance called \( Y \), the result of a standardized examination. There are \( I = 2 \) curricula. The effect of the \( i \)-th curriculum is fixed and represented by \( \alpha_i \) with the constraint that \( \sum_{i=1}^{I} \alpha_i = 0 \). For each curriculum, the research team proposes to take a random sample of \( J = 4 \) colleges and teach the curriculum in that college. They will select a random sample of \( K = 3 \) classes in each college to receive standardized instruction in the curriculum. They will select \( L = 5 \) students at random from each class (each class has more than 5 students) and measure each sampled student’s performance. Note that the students are nested within classes and that classes are nested within colleges, and that colleges are nested within curriculum and that the design is balanced.

The researchers will use the model

\[
Y_{ijkl} = \mu + \alpha_i + B_{ij} + C_{ijk} + \sigma_Z Z_{ijkl}.
\]

The random variables \( B_{ij} \) (representing the college contributions) are normal and independently distributed with expected value 0 and variance \( \sigma_B^2 \). The random variables \( C_{ijk} \) (representing the class contributions) are normal and independently distributed with expected value 0 and variance \( \sigma_C^2 \). The random variables \( Z_{ijkl} \) are independent and identically distributed normal random variables with mean 0 and variance 1. The set of \( B_{ij} \) are independent of the set \( C_{ijk} \), and both are independent of the set of \( Z_{ijkl} \).

(1). State the formula for each relevant sum of squares in the Analysis of Variance Table and state its expected value.

(2). Specify the test of the null hypothesis that \( \alpha_i = 0 \) for \( i = 1, \ldots, I \) and specify its distribution under the usual alternative hypothesis.

(3). This null hypothesis will be tested at level of significance 0.01. The research team has asked the consultant to calculate the power of the test when \( \alpha_1 = 10 \) and \( \alpha_2 = -10 \), \( \sigma_B = 20 \), \( \sigma_C = 5 \), and \( \sigma_z = 30 \). What is the power of the test?