Ph.D. Preliminary Examination
Statistics
May 29, 2015

NOTES:

1. The exam is worth 100 points.
2. Partial credit may be given for partial answers if possible.
3. There are 4 pages in this exam paper.

I have neither given nor received aid on this examination.

Name (print): ____________________________

Student ID  : ____________________________

Signature, Date: _________________________

==================================================================
1. (15 points) Let \( X_1, X_2, \ldots, X_n \) be uniformly distributed on the interval 0 to \( a \).
   (a) Please find the moment estimator for \( a \)
   (b) Is the answer of (a) an unbiased estimator
   (c) Discuss the reasonableness of this estimator

2. (15 points) A test was conducted to investigate whether a die with six sides is fair or not. The die was tossed 180 times, and the following table lists the results.

<table>
<thead>
<tr>
<th>Side</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>26</td>
<td>50</td>
<td>20</td>
<td>22</td>
<td>40</td>
<td>22</td>
</tr>
</tbody>
</table>

Please judge whether there strong evidence that the die is not fair? Use a significance level of 0.1.

3. (20 points) A manufacturing company wants to study whether the product material has an effect on product quality. Test was performed on 20 product samples made of different material. The sample mean is 1.25 and the sample standard deviation is \( s = 0.25 \).

   (a) Test the hypothesis that \( \sigma = 0.10 \) against an alternative specifying that \( \sigma \neq 0.10 \), using \( \alpha = 0.01 \), and draw a conclusion. State any necessary assumptions about the underlying distribution of the data.
   (b) What is the P-value for this test?
   (c) Find another way to answer the question in part (a) rather than using the statistical hypothesis test.

4. (25 points) Regression Analysis
   (a) Consider the model
   \[
   Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \beta_4 x_{i4} + \text{error}_i, \ i = 1, \ldots, n.
   \]
   One wants to test the null hypothesis \( \beta_2 = \beta_3 = \beta_4 = 0 \) against the alternative hypothesis \( \beta_2 \neq 0 \) and/or \( \beta_3 \neq 0 \) and/or \( \beta_4 \neq 0 \). Which class of distributions does the corresponding test statistic have?

   (b) How many parameters have to be estimated for a multinomial regression model with a response variable with 4 levels, 2 continuous predictors, 1 categorical variable with 5 levels and an intercept? How many observations do you at least recommend for fitting this model?

   (c) Decide (with short explanations) whether the following statements are true or false.
i. Leverage points should always be removed from the regression analysis.

ii. One square root-transformed the response variable and now wants to calculate confidence intervals on the original scale. He/she therefore just needs to square the confidence intervals on the transformed scale.

iii. In a regression model a factor with 4 levels served as predictor. One of these levels is not significant and should consequently be removed from the analysis.

iv. There are two predictors in a multiple linear regression which are not significant. The global F-test is highly significant. For this reason, it is better not to remove both predictors simultaneously.

5. (25 points) The fire department of a state uses conventional gasoline for its feet of fire vehicles. For environmental reasons, the state governor wants to switch to a mixture of gasoline and ethanol. But the fire department head objects to this plan, because he believes that it will decrease fuel efficiency and strain the department's budget. They decide to conduct an experiment to estimate the effects of ethanol on fuel efficiency. A consultant proposes two different designs.

**Design 1:** 100 firemen are randomly divided into five groups of 20. Each fireman is given a car to drive for a 12-hour shift. For Group 1, no ethanol (0%) is added to the fuel. For Groups 2, 3, 4 and 5, the levels of ethanol are 2.5%, 5%, 7.5% and 10%, respectively. The response variable is fuel efficiency, measured in miles per gallon.

**Design 2:** 20 firemen are given fire vehicles to drive, and each fireman drives the same car for five consecutive 12-hour shifts. Each fireman drives for one shift at each level of ethanol (0%, 2.5%, 5%, 7.5%, 10%). For each fireman, the five levels of ethanol are given in a random order. The response variable is fuel efficiency, measured in miles per gallon.

(a) What kind of experimental design is Design 1? What kind of experimental design is Design 2?

(b) The state governor says, “Design 1 will obviously give more accurate results, because the sample size is larger. Design 1 uses 100 firemen, but Design 2 uses only 20.” The fire department head says, “The two designs will be equally accurate, because each design uses 100 shifts.” Is either of these statements correct? Which design do you think will give more accurate estimates of treatment effects? (What will determine whether Design 1 or Design 2 is better?)

(c) For Design 1, the consultant proposes to analyze the data in two different ways. Analysis A: Perform a simple linear regression of the response variable, $Y = \text{fuel efficiency}$, on $X = \text{percentage of ethanol}$,

$$Y = \beta_0 + \beta_1 X + \text{error} ;$$

and test the null hypothesis $H_0 : \beta_1 = 0.$
Analysis B: Perform a one-way ANOVA with five treatment groups. Define a contrast $L$ for the linear effect of ethanol with contrast weights $-2; -1; 0; 1; 2$, and test the null hypothesis $H_0 : L = 0$. Will the results of these two analyses be identical? If yes, explain why. If no, explain why. (Are the null hypotheses equivalent? Are the F-statistics and degrees of freedom identical?)

(d) For Design 2, describe two different ways to analyze the data to test for a linear trend. For each method of analysis, explain with formulas how you would compute the test statistic, state the degrees of freedom, and state the assumptions being made.

(e) Consider Design 2. If the order in which the treatments are given has a large effect on the response, it may be advantageous to balance the order so that each treatment appears in each position (1,2,3,4,5) an equal number of times. What kind of design would that be? Write a linear model and ANOVA table that would be appropriate for this new design. For each line in the ANOVA table, give the correct error term for the F-test.