Statistics Qualifying Exam

9:00 am - 1:00 pm, Monday, May 2, 2016

1. If Y_1 and Y_2 have a joint distribution given by

$$f(y_1, y_2) = \begin{cases} \frac{1}{2} y_1 y_2, & 0 \leq y_2 \leq y_1 \leq 2; \\ 0, & \text{elsewhere.} \end{cases}$$

   (a) Find the marginal distributions of Y_1 and Y_2.

   (b) Find the conditional distribution of Y_2 given $Y_1 = y_1$.

   (c) Find the values of $E(Y_2|Y_1 = 1)$ and $Var(Y_2|Y_1 = 1)$.

   (d) What is the density of $U = Y_1 - Y_2$?

2. Suppose that $X_1, , X_n$ is a random sample from Uniform$(\theta, 2\theta)$ distribution, where $\theta > 0$ is an unknown parameter.

   (a) Find the maximum likelihood estimator (MLE) for θ.

   (b) Prove that the MLE is consistent.

3. Let $X_1, \ldots, X_n$ is a random sample from a distribution with pdf as $f(x|\theta) = \theta^{-1} \exp(-x/\theta), x \geq 0, \theta > 0$.

   (a) Find the MLE of $P(X \leq 2)$.

   (b) Find the minimum variance unbiased estimator (MVUE) of $P(X \leq 2)$.

   (c) Derive the exact likelihood ratio test for the following hypothesis

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0.$$

   Please carefully specify the likelihood ratio Λ, the test statistic, the sampling distribution of the test statistic under the null hypothesis, and the decision rules of a size α test.

   (d) For the decision rule derived in Part (b), obtain the distribution of the test statistic under a general alternative (that is $H_1 : \theta \neq \theta_0$) and use it to obtain the power function of the test.

4. Let $X_1, \ldots, X_n$ be a random sample from the pdf $f(x|\theta) = \exp[-(x - \theta)]$, where $-\infty < \theta < \infty$ and $x \geq \theta$.

   (a) Find the minimal sufficient statistic for θ.

   (b) Let $Y_1 < Y_2 < \cdots < Y_n$ be the ordered sample, and define $R_i = Y_n - Y_i$, $i = 1, \ldots, n-1$. Show that the set $(R_1, R_2, \ldots, R_n)$ is ancillary for θ.

   (c) Assume it is given that $Y_1 = \min_i X_i$ is a complete sufficient statistic. Show that Y_1 and $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ are independent, where $\bar{X} = \sum_{i=1}^n X_i/n$.

   (d) Now prove the condition given in Part (c). That is, show that $Y_1 = \min_i X_i$ is a complete sufficient statistic.

5. The following is part of ANOVA table for a simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where ε_i's are i.i.d. from $N(0, \sigma^2)$, $i = 1, \ldots, n$, and n is the number of observations. (No need to complete the table.)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	***	******	252378	105.88	<.0001
Error	23	******	******		
Corrected Total	***	******			

(a) Compute the coefficient of determination R^2.

(b) Assume that we now know the least square estimate of β_1 is $b_1 = 3.57$. Construct a two-sided t-test of whether or not $\beta_1 = 3$. State the null and alternative hypotheses, the value of the test statistic, the sampling distribution under the null hypothesis and the decision rule.

6. A gardener is interested in studying the relationship between fertilizer and tomato yield. The gardener has two gardens (1 and 2). He divides each into 9 plots. Three fertilizer application rates (3, 5, and 7 units/acre) are assigned to the plots in garden 1 in a completely randomized fashion. The same three fertilizer application rates (3, 5, and 7 units/acre) are assigned to the plots in garden 2 in a completely randomized fashion. Thus there are three plots for each combination of garden and fertilizer application rate. After some initial analyses, the gardener decides to base his analysis on the following SAS code and output.

(a) Note that rate was not included in the class statement. What would the Model and Error DF change to if rate were included in the class statement? That is, complete the following tables by filling in the missing values for DF (*you only need to provide DF for ???, but do not need to calculate any of the Sum Squares*).

Source	DF	Sum of Sum Squares
Model	???	
Error	???	
Corrected Total	???	

Source	DF	Type I SS
garden	???	
rate	???	
rate*garden	???	

(b) Estimate the equation of the regression line relating yield to fertilizer application rate in garden 1.

(c) Estimate the equation of the regression line relating yield to fertilizer application rate in garden 2.

(d) Is there a significant difference between the slopes of the two regression lines? To get full credits, give an appropriate test statistic, p-value, and conclusion, using $\alpha = 0.05$.

(e) Suppose the gardener were to apply 7 units of fertilizer per acre to all plots in both gardens. Which garden would have the higher expected yield?

```
proc glm;
  class garden;
  model yield=garden rate garden*rate / solution;
run;
```

The GLM Procedure
Dependent Variable: yield

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	58.88888889	19.62962963	27.33	<.0001
Error	14	10.05555556	0.71825397		
Corrected Total	17	68.94444444			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
garden	1	2.72222222	2.72222222	3.79	0.0719
rate	1	52.08333333	52.08333333	72.51	<.0001
rate*garden	1	4.08333333	4.08333333	5.69	0.0318

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		-1.11 B	0.90993803	-1.22	0.2422
garden	1	3.69 B	1.28684670	2.87	0.0123
garden	2	0.00 B	.	.	.
rate		1.33 B	0.17299494	7.71	<.0001
rate*garden	1	-0.58 B	0.24465179	-2.38	0.0318
rate*garden	2	0.00 B	.	.	.

7. An experiment is conducted to study the effects of loading frequencies (Frequency) and environmental conditions (Environment) on fatigue crack growth at a constant 22 MPa stress for a particular material. The data from this experiment are shown below (the response is crack growth rate):

	Environment		
Frequency	1: Air	2: H_2O	3: SaltH_2O
1	2.29, 2.47, 2.12	2.86, 3.03, 2.73	4.93, 4.75, 5.06
2	3.15, 2.88, 2.56	4.00, 4.44, 4.70	3.10, 3.24, 3.98
3	2.24, 2.71, 2.81	4.00, 4.30, 3.20	4.86, 4.26, 5.20

Some summary statistics are give below.

```
grand mean: 3.551
frequency    MEAN        |environment    MEAN
1           3.360        | 1             2.581
2           3.561        | 2             3.696
3           3.731        | 3             4.376

frequency environment   MEAN
1             1         2.293
1             2         2.873
1             3         4.913
2             1         2.863
2             2         4.380
2             3         3.440
3             1         2.587
3             2         3.833
3             3         4.773
```

Suppose the following statistical model is used to fit the data.

$$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk}; k = 1, 2, 3$$

where $\tau_i (i = 1, 2, 3)$ and $\beta_j (j = 1, 2, 3)$ are the main effects of frequency, the main effects of environment, respectively, and $(\tau\beta)_{ij}$ are their interactions. For parameter estimation, we impose the following constraints: $\sum_i \tau_i = \sum_j \beta_j = \sum_i (\tau\beta)_{ij} = \sum_j (\tau\beta)_{ij} = 0$.

(a) Calculate the estimates of τ_1, τ_2, τ_3 and $(\tau\beta)_{23}$.

(b) Calculate the sum of squares due to frequency.

(c) The ANOVA of the data was done in SAS and the output is shown. Notice that some quantities are removed. Test if the main effects of frequency are significant and if the interactions of frequency and environment are significant (using $\alpha = 5\%$).

```
Dependent Variable: crack
                        Sum of
Source        DF        Squares          Mean Square      F Value Pr > F
Model          8        22.719               2.840        22.11   <.0001
Error         18         2.312               0.128
Coed Total    26        25.031

Source                   DF          Type I SS     Mean Square  F Value   Pr > F
frequency                *           *****         *****        ****      *****
environment              2           14.773        7.387        57.51     <.0001
frequency*environment    *           *****         *****        ****      *****
```

(d) Calculate the critical difference (CD) for the treatment pairwise comparison using Tukey's method. Report the results for the following (*and only the following*) pairs: (1, 1) versus (1,3); (2, 2) versus (3, 2).

8. An experiment was run to determine whether four specific temperatures affect the production of a certain type of chemical compound. The experiment led to the following data. The response is the amount of the chemical compound produced over a period of time.

Temperature	Response					Mean $\bar{Y}_{i.}$	St.D. s_i
100	8.71	10.47	9.62	11.55	10.25	10.12	1.05
125	24.12	25.23	26.08	20.30	21.15	23.38	2.54
150	30.37	31.14	35.62	30.19	27.03	30.87	3.09
175	28.24	24.15	22.02	26.46	27.44	25.66	2.55

(a) Generate (approximately) the plot of log St.D. ($log\ s_i$) versus log Mean ($log\ \bar{Y}_{i.}$) by hand.

(b) Based on your plot, is the constant variance assumption for ANOVA valid? What remedy you can recommend? Derive the remedy explicitly.

The modified data are given below.

Temperature	Transformed Response					Mean $\bar{Y}_{i.}$	St.D. s_i
100	2.16	2.35	2.26	2.45	2.33	2.310	0.1079
125	3.18	3.23	3.26	3.01	3.05	3.146	0.1106
150	3.41	3.44	3.57	3.41	3.30	3.426	0.0966
175	3.34	3.18	3.09	3.28	3.31	3.240	0.1032

(c) It is known that $SST = 3.839$. Choose an appropriate model, construct the ANOVA table, and test if the temperatures have different effects on the transformed response.

(d) What is the estimate for τ_1 (the treatment effect at temperature 100) under the constraint $\sum_i \tau_i = 0$?

(e) Let μ_1, μ_2, μ_3 and μ_4 be the treatment means at temperatures 100, 125, 150 and 175, respectively, and let $L = \mu_1 - 2\mu_2 + \mu_3$. What is the estimate for L?

(f) Test $H_0 : L = 0$ vs $H_1 : L \neq 0$.

(g) One decides to use the contrasts based on orthogonal polynomials to further model the relationship between temperature and the transformed response. Part of the SAS code and output is given below. Fill in the DF, Contrast SS, Mean Square and F Value for the linear contrast.

```
contrast 'linear'       temperature -3 -1  1 3;
contrast 'quadratic'    temperature  1 -1 -1 1;
contrast 'cubic'        temperature -1  3 -3 1;
-------------------------------------------------------------------
Contrast     DF  Contrast SS        Mean Square     F Value     Pr > F
linear       *   *********          ********        ******      <.0001
quadratic    1   1.30560500         1.30560500      119.07      <.0001
cubic        1   0.00202500         0.00202500        0.18      0.6731
```