Statistics Qualifying Examination

Answer all questions and show all work.
This exam is closed-note/book. You need to use a calculator.

1. Suppose that Y_1 and Y_2 have the joint distribution $f_{Y_1,Y_2}(y_1, y_2) = y_1 y_2/2$ if $0 \leq y_2 \leq y_1 \leq 2$ and 0 otherwise.

   (a) Find the marginal probability density functions (pdfs) of Y_1 and Y_2.

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

   (c) Find the values of $E(Y_2|Y_1 = 1)$ and $Var(Y_2|Y_1 = y_1)$.

   (d) Find the pdf of $U = Y_1 - Y_2$.

2. The following questions are related to change of variables techniques where the transformation function is monotone (part a) or piece-wise monotone (part b), respectively.

   (a) Let X be a continuous random variable with the pdf $f_X(x)$ and support S_X. Let $Y = g(X)$ where $g(x)$ is a strictly monotonic function on S_X.
   Then, prove that the pdf of Y is given by

   $$f_Y(y) = f_X(g^{-1}(y)) \ \frac{d[g^{-1}(y)]}{dy}$$

   where the support of Y is $S_Y = \{y : y = g(x), x \in S_X\}$.

   (b) Let Z follow the standard normal distribution. Prove that a new random variable V such that $V = Z^2$ follows the chi-square distribution with the degrees of freedom one, i.e., $\chi^2(1)$.

3. Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with the pdf as

$$f(x) = \begin{cases} \dfrac{1}{\theta}, & 0 < x < \theta; \\ 0, & \text{elsewhere.} \end{cases}$$

Let $Y_1 < Y_2 < \ldots < Y_n$ be the order statistics.

(a) Derive the pdf of the sample maximum Y_n.

(b) Show that Y_n converges in probability to θ, i.e., Y_n is a consistent estimator for θ.

(c) Find the joint pdf of the sample minimum and the sample maximum, (Y_1, Y_n).

(d) The range is defined as $R = Y_n - Y_1$. Find the joint pdf of (R, Y_n).

(e) Based on (d), are R and Y_n independent? Clearly justify your answer.

4. Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with the pdf

$$f(x; \beta) = \frac{1}{2}\beta^{-3}x^2 \exp\Big(-\frac{x}{\beta}\Big),$$

where $x > 0$ and $\beta > 0$.

(a) Find the maximum likelihood estimator $\hat{\beta}$ of β.

(b) Is $\hat{\beta}$ an unbiased estimator of β? Clearly justify your answer.

(c) Is $\hat{\beta}$ an efficient estimator of β? Clearly justify your answer.

(d) Let $Y = \sum_{i=1}^{n} X_i$. Show that Y is a complete and sufficient statistic for β.

(e) Find the minimum-variance unbiased estimator for β.

5. A substance used in biological and medical research is shipped by airfreight to users in cartons of 1,000 ampules. Data were collected on the number of times the carton was transferred from one aircraft to another over the shipment route (X) and the number of ampules found to be broken upon arrive (Y). A simple linear normal error regression model has been fitted to the data. The SAS output from PROC REG is shown below with many values missing.

```
                                   The REG Procedure
                                     Model: MODEL1
                                   Dependent Variable: Y

                   Number of Observations Read           8
                   Number of Observations Used           8

                                    Analysis of Variance

                                  Sum of            Mean
Source                DF          Squares         Square        F Value     Pr > F
Model                  ***        *********      *********     *********    0.0004
Error                  ***        *********        2.66667
Corrected Total        ***        144.00000


Root MSE                  *********          R-Square    *********
Dependent Mean            *********          Adj R-Sq     0.8704
Coeff Var                 *********

                                     Parameter Estimates

Parameter                            Standard
Variable    DF      Estimate         Error          t Value        Pr > |t|
Intercept    1      10.00000       *********       *********       <.0001
X            1       4.00000       *********       *********      *********
```

Answer the following questions based on the available information given in the above SAS output.

(a) Calculate the percentage of variability in the number of ampules found to be broken upon arrive explained by the simple linear regression model.

(b) Conduct an F test to decide whether or not there is a linear association between the number of times a carton is transferred (X) and the number of broken ampules (Y). Use $\alpha = 0.05$. You must clearly state the null and the alternative hypotheses, the sampling distribution under the null hypothesis, decision rule using p-value, and your conclusion.

(c) Interpret the result that the estimation of β_1 is $b_1 = 4$.

(d) Test the hypothesis $H_0 : \beta_1 = 3$ versus $H_1 : \beta_1 < 3$. Use $\alpha = 0.05$.

Below is a list of selected percentage points for the standard normal distribution with Z_α implying that $P(Z > Z_\alpha) = \alpha$ and the t-distribution with v degrees of freedom, where $t_{v,\alpha}$ implies $P(T_v > t_{v,\alpha}) = \alpha$.

$$\begin{aligned} Z_{0.10} &= 1.28; & Z_{0.05} &= 1.64; & Z_{0.025} &= 1.96; \\ t_{6,\,0.025} &= 2.45; & t_{7,\,0.025} &= 2.36; & t_{8,\,0.025} &= 2.31; \\ t_{6,\,0.05} &= 1.94; & t_{7,\,0.05} &= 1.89; & t_{8,\,0.05} &= 1.86. \end{aligned}$$

6. Suppose that you were hired as a statistical consultant by a client to investigate the association between advertising and sales of a particular product. Data are collected on the sales of that product (Y) in 50 different markets, along with advertising budgets for the product in each of those markets for three different media: TV(X_1), radio(X_2), and newspaper(X_3). You have decided to use the normal linear regression model. Results on a collection of models are included below, where MSE is the error mean square and R^2 represents the coefficient of determination.

Model index	MSE	R^2	variables in the model
1	13.259	0.622	X_1
2	17.817	0.492	X_2
3	34.211	0.025	X_3
4	3.167	0.912	X_1 X_2
5	12.768	0.644	X_1 X_3
6	17.982	0.498	X_2 X_3
7	3.191	0.913	X_1 X_2 X_3
8	34.367	0.000	None of X's

(a) Comment on the following statement: "The best model selected based on R^2 is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$ since it is the model with the largest R^2."

(b) Carry out a forward selection prodecure to decide the best model. Use a significant entry level $\alpha = 0.05$.

Below is a list of selected percentage points for the F-distribution with v_1 and v_2 degrees of freedom, where $F_{v_1,v_2,\alpha}$ implies $P(F_{v_1,v_2} > F_{v_1,v_2,\alpha}) = \alpha$.

$F_{1,50\,0.05} = 4.034$; $F_{1,49\,0.05} = 4.038$; $F_{1,48\,0.05} = 4.043$; $F_{1,47\,0.05} = 4.047$; $F_{1,46\,0.05} = 4.052$.

7. In an experiment, the amount of radon released in shower was investigated. Radon-enriched water was used, and five different orifice diameters were tested in shower heads. The data from the experiment are shown in the following table (the response is the percentage of radon released).

Orifice Diameter					Mean	Std.D
0.40	87	88	89	93	89.25	2.62995564
0.60	74	73	76	77	75.00	1.82574186
0.80	69	71	70	72	70.50	1.29099445
1.00	76	72	74	74	74.00	1.63299316
1.20	89	92	84	89	88.50	3.31662479

The ANOVA table obtained from SAS is given below.

```
                              Analysis of Variance

                                   Sum of             Mean
Source                 DF         Squares           Square       F Value

Model                   4        1230.200       307.550000         60.11
Error                  15          76.750         5.116667
Corrected Total        19        1306.950
```

(a) Perform pairwise comparison for the five different orifice diameters in terms of their corresponding radon released, using Bonferroni and Tukey's methods, *respectively* (use $\alpha = 0.1$). Summarize and compare the results, and comment on which comparison method is preferred.

df	13	14	15
$t_{0.0025;df}$	3.3725	3.3257	3.2860
$t_{0.005;df}$	3.0123	2.9768	2.9467
$t_{0.01;df}$	2.6503	2.6245	2.6025
$t_{0.05;df}$	1.7709	1.7613	1.7531

(df_1, df_2)	(4,13)	(4,14)	(4,15)	(5,13)	(5,14)	(5,15)
$q_{0.025;df_1,df_2}$	4.6941	4.6375	4.5893	5.0041	4.9399	4.8851
$q_{0.05;df_1,df_2}$	4.1509	4.1105	4.0760	4.4529	4.4066	4.3670

(b) Notice that the amount of released radon changes when the size of orifice varies from 0.40 to 1.20 in diameter. An analyst wants to study the functional relationship between the response and the diameter. She obtains the complete set of orthogonal contrasts from Table IX in Montgomery:

$$\begin{array}{lrrrrr}
C1: & -2 & -1 & 0 & 1 & 2 \\
C2: & 2 & -1 & -2 & -1 & 2 \\
C3: & -1 & 2 & 0 & -2 & 1 \\
C4: & 1 & -4 & 6 & -4 & 1
\end{array}$$

The contrast sum of squares for $C1$, $C3$, and $C4$ and their testing results are given below. Obtain the estimate of the contrast $C1$.

```
Contrast        DF         Contrast SS     Mean Square     F Value       Pr>F
C1              1            2.500000        2.500000        0.49      0.4952
C2              *             ********       ********        ****      ******
C3              1            0.625000        0.625000        0.12      0.7316
C4              1            1.289286        1.289286        0.25      0.6230
```

(c) Note that the contrast SS, Mean square, F-value and Pr $> F$ for $C2$ are missing. Recover these values and test if $C2$ is significant (Use $\alpha = 0.05$).

(df_1, df_2)	$(1, 13)$	$(1, 14)$	$(1, 15)$	$(2, 13)$	$(2, 14)$	(2, 15)
$F_{0.025;df_1,df_2}$	6.4143	6.2979	6.1995	4.9653	4.8567	4.7650
$F_{0.05;df_1,df_2}$	4.6672	4.6001	4.5431	3.8056	3.7389	3.6823

8. An experiment was conducted to determine the effects of four different pesticides on the yield of fruit from three different varieties of a citrus tree. Eight trees from each variety were available and the four pesticides were then randomly assigned to *two* trees of each variety. Yields of fruit (in bushels per tree) were obtained after the test period. The mean yields for each combination of pesticide and variety are given below.

		Variety		
Pesticide	1	2	3	Pesticide Means
1	44	48	67	53.00
2	52.5	62.5	88.5	67.83
3	40.5	47.5	65.5	51.17
4	50.5	79	92	73.83
Variety Means	46.875	59.25	78.25	

Suppose that the following statistical model is used to fit the data:

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

where τ_i $(i = 1, 2, 3, 4)$ and β_j $(j = 1, 2, 3)$ are the main effects of pesticide and variety, and $(\tau\beta)_{ij}$ are their interactions, respectively. For parameter estimation, we impose the following constraints as we had in the lecture notes: $\sum_i \tau_i = \sum_j \beta_j = \sum_i (\tau\beta)_{ij} = \sum_j (\tau\beta)_{ij} = 0$.

Some SAS output is given below:

```
The GLM Procedure
Dependent Variable: yield

Source                  Sum of Squares
Model                      6680.458333
Error                       507.500000
Corrected Total            7187.958333

Source                      Type I SS
pesticide                 2227.458333
variety                   3996.083333
pesticide*variety          456.916667
```

(a) Provide the degrees of freedom corresponding to each of the sums of squares in the output, which are marked by "???" below.

Source	Degrees of freedom
Model	???
Error	???
Corrected Total	???
pesticide	???
variety	???
pesticide * variety	???

(b) Are the effects of the pesticides on yield *dependent* on the variety of citrus tree? Conduct an appropriate test to answer this question. To get full credits, give hypotheses, a test statistic, determine its degrees of freedom, use an appropriate value from the table below, state the p-value (or its range), and give your conclusion using $\alpha = 0.05$.

(df_1, df_2)	$(5, 10)$	$(5, 11)$	$(5, 12)$	$(5, 13)$	(5,14)
$F_{0.05;df_1,df_2}$	3.3258	3.2039	3.1059	3.0254	2.9582
(df_1, df_2)	$(6, 10)$	$(6, 11)$	$(6, 12)$	$(6, 13)$	(6,14))
$F_{0.05;df_1,df_2}$	3.2172	3.0946	2.9961	2.9153	2.8477

(c) Use Tukey's method to perform a pairwise comparison for the three different *varieties*. Report the critical difference **and** report your results of comparison (using $\alpha = 0.05$). You can report the result as we have done in class by labeling significantly different combinations with different Latin letters.

(df_1, df_2)	$(2, 10)$	$(2, 11)$	$(2, 12)$	$(2, 13)$	$(2, 14)$
$q_{0.025;df_1,df_2}$	3.7247	3.6672	3.6204	3.5817	3.5491
$q_{0.05;df_1,df_2}$	3.1511	3.1127	3.0813	3.0552	3.0332
(df_1, df_2)	$(3, 10)$	$(3, 11)$	$(3, 12)$	$(3, 13)$	$(3, 14)$
$q_{0.025;df_1,df_2}$	4.4740	4.3913	4.3243	4.2687	4.2220
$q_{0.05;df_1,df_2}$	3.8768	3.8196	3.7729	3.7341	3.7014

(d) Suppose that Pesticides 1 and 2 are sold by Company A while Pesticides 3 and 4 are sold by Company B. Conduct a test or construct a confidence interval to compare the effectiveness of Company A's pesticides to the effectiveness of Company B's pesticides. Show your work and provide your conclusion (using $\alpha = 0.05$).

(df_1, df_2)	$(1, 10)$	$(1, 11)$	$(1, 12)$	$(1, 13)$	$(1, 14)$
$F_{0.05;df_1,df_2}$	4.9646	4.8443	4.7472	4.6672	4.6001
df	10	11	12	13	14
$t_{0.05;df}$	1.8125	1.7959	1.7823	1.7709	1.7613
$t_{0.025;df}$	2.2281	2.2010	2.1788	2.1604	2.1448