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 a random variable Z follows the standard normal distribution whose probability density function is written by

$$f_Z(z) = (2\pi)^{-1/2} e^{-z^2/2}.$$

The moment generating function of a random variable X is defined as $M(t) = \mathrm{E}(e^{tX})$ for $-h < t < h$ with some $h > 0$.

a. Prove that moment generating function of Z is $M_Z(t) = e^{t^2/2}$.

b. Suppose that Y is a random variable such that $Y = \mu + \sigma Z$ where μ and σ are unknown, but fixed constants.

Using the moment generating function of Z in part (a), show that the moment generating function of Y is $M_Y(t) = e^{\mu t + \sigma^2 t^2/2}$.

c. Find the mean and variance of Y using the moment generating function of Y in part (b).

2. Suppose that we conduct a clinical study for a vaccine against a certain type of flu. At random, 200 people are assigned to two groups: treatment group, in which people receive the vaccine, and control group, in which people receive a placebo.

The following table shows the status of the 200 people six months after getting the vaccine or the placebo.

	Not infected	Infected with flu
Treated group	85	15
Control group	70	30

Let p_t and p_c be the infection rate of the treatment group and the control group, respectively.

a. Find an *approximate* 95% confidence interval for $p_t - p_c$.

b. *Interpret* the 95% confidence interval in plain language.

3. Let $X_1, \ldots, X_n$ be a random sample of size $n > 1$ from a distribution that is Bernoulli(θ), $0 < \theta < 1$.

a. Find the minimum variance unbiased estimator (MVUE) $\tilde{\delta}$ of $\delta = \theta(1-\theta)$. Carefully justify that the obtained estimator is MVUE.

b. Find the the asymptotic distribution of $\sqrt{n}(\tilde{\delta} - \delta)$.

4. Let $X_1, \ldots, X_n$ be a random sample from a population with the probability density function (pdf) as

$$f_X(x; \alpha, \beta) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right), & x > 0, \alpha > 0, \beta > 0; \\ 0, & \text{otherwise.} \end{cases}$$

which is a Gamma (α, β) distribution. Assume that the parameter β is unknown and α is known.

a. Find the exact Likelihood Ratio test for the hypotheses $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$. Specifically

(a-i) Show that the likelihood ratio depends only on $S = 2\sum_{i=1}^n X_i/\beta_0$. That is, S is the test statistic used here.

(a-ii) What is the sampling distribution of S under H_0?

(a-iii) Explicitly state the decision rule of a size α_0 test in the form "Reject H_0 if $S \leq c_1$ or $S \geq c_2$" or "Reject H_0 if $S \geq c_3$" with the constants c_1 and c_2 or c_3 specified.

b. Obtain the power function of the test obtained in Part (a). Express the results in terms of a standard distribution.

5. A multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

with $\epsilon \sim N(0, \sigma^2 \mathbf{I})$ is fitted to a dataset ($\mathbf{I}$ is an identity matrix).

Refer to the software output below and answer the following questions.

Model: MODEL1
Dependent Variable: Y

Number of Observations Read	81
Number of Observations Used	81

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	******	138.32691	34.58173	26.76	<.0001
Error	******	98.23059	1.29251		
Corrected Total	******	236.55750			

Root MSE	1.13689	R-Square	0.5847
Dependent Mean	15.13889	Adj R-Sq	0.5629
Coeff Var	7.50970		

```
## Coefficients:
##               Estimate    Std. Error  t value  Pr(>|t|)
## (Intercept) 12.20059     0.57796     21.11    <.0001
## X1          -0.14203     NA          NA       <.0001
## X2           0.28202     0.06317      4.46    <.0001
## X3           0.61934     1.08681      0.57    0.5704
## X4           0.00000792  0.00000138   5.72    <.0001
```

We also have got two types of sums of squares.

```
Variable      Type I SS
X1            14.70852
X2            (1)
X3            8.38142
X4            (2)

Variable      Type III SS
X1            57.15802
X2            (3)
X3            0.41975
X4            42.32496
```

a. Fill in the numbered blanks **(1)**, **(2)**, and **(3)**. *[No need to fill in the blanks denoted by ******.]*

b. Fill in the blanks denoted by NA.

c. For the full model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$, test the following hypotheses (using $\alpha = 0.05$):

$$H_0 : \beta_2 = \beta_3 = 0 \text{ vs. } H_a : \text{ not all equals } 0$$

Clearly specify the test statistic, its sampling distribution under the null hypothesis and your conclusion. To answer this question, you may need to use (some of) the following values.

(df_1, df_2)	$(2, 74)$	$(2, 75)$	$(2, 76)$	$(2, 77)$	$(2, 78)$
$F(0.975; df_1, df_2)$	3.879036	3.876416	3.873867	3.871387	3.868972
$F(0.95; df_1, df_2)$	3.120349	3.118642	3.116982	3.115366	3.113792
(df_1, df_2)	$(3, 74)$	$(3, 75)$	$(3, 76)$	$(3, 77)$	$(3, 78)$
$F(0.975; df_1, df_2)$	3.298180	3.295668	3.293225	3.290847	3.288532
$F(0.95; df_1, df_2)$	2.728280	2.726589	2.724944	2.723343	2.721783

6. 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

Source	DF	Sum of Squares	Mean 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 using Bonferroni and Tukey's methods, *respectively* (use $\alpha = 0.1$). Summarize and compare the results, and comment on which 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:

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

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

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

7. Seven different hardwood concentrations are being studied to determine their effect on the strength of the paper produced. As days may differ, the analyst carried out the experiment described below:

Hardwood	Days							
Concentration (%)	1	2	3	4	5	6	7	$Y_{i.}$
2	114				120		117	351
4	126	120				119		365
6		137	117				134	388
8	141		129	149				419
10		145		150	143			438
12			120		118	123		361
14				136		130	127	393

a. What kind of design is used by the experimenter? Determine the design parameters, and give a model to analyze this data set together with the assumptions.

b. Some SAS output is given below. Test whether the seven hardwood concentrations are different in terms of their effect on the strength of the paper produced (use $\alpha = 0.05$). To get full credits, please give the hypotheses, test statistic, p-value and your conclusion.

The GLM Procedure

Dependent Variable: strength C

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	2431.714286	202.642857	9.62	0.0017
Error	8	168.571429	21.071429		
Corrected Total	20	2600.285714			

R-Square	Coeff Var	Root MSE	strength Mean
0.935172	3.550557	4.590363	129.2857

Source	DF	Type I SS	Mean Square	F Value	Pr > F
concentration	6	2037.619048	339.603175	16.12	0.0005
day	6	394.095238	65.682540	3.12	0.0701

Source	DF	Type III SS	Mean Square	F Value	Pr > F
concentration	6	1317.428571	219.571429	10.42	0.0021
day	6	394.095238	65.682540	3.12	0.0701

Here is some output from `means` (left) and `lsmeans` (right), respectively.

Level of concentration	N	strength	
		Mean	Std Dev
2	3	117.000000	3.0000000
4	3	121.666667	3.7859389
6	3	129.333333	10.7857931
8	3	139.666667	10.0664459
10	3	146.000000	3.6055513
12	3	120.333333	2.5166115
14	3	131.000000	4.5825757

concentration	strength LSMEAN	LSMEAN Number
2	116.857143	1
4	120.714286	2
6	131.857143	3
8	140.000000	4
10	**********	5
12	124.142857	6
14	128.428571	7

c. Obtain the least squares mean (adjusted mean) for hardwood concentration 10%.

8. An engineer is designing a battery for use in a device that will be subject to some extreme variations in temperatures. She decides to test three plate materials at three temperature levels, resulting two factors at three levels each. Four batteries are tested at each combination of plate material and temperature, and all 36 tests are run in random order. The experiment and the resulting observed battery life data are given in the table below. A longer life is preferred. The overall mean battery life of the sample is 105.53.

Material Type	Temperature		
	15	70	125
1	130, 155, 74, 180	34, 40, 80, 75	20, 70, 82, 58
2	150, 188, 159, 126	136, 122, 106, 115	25, 70, 58, 45
3	138, 110, 168, 160	174, 120, 150, 139	96, 104, 82, 60

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

$$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk}; k = 1, 2, 3, 4$$

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

Here are the outputs from `MEANS` statement of `PROC GLM` in SAS:

Level of		life	
temp	N	Mean	Std Dev
15	12	144.833333	31.6940870
70	12	107.583333	42.8834750
125	12	64.166667	25.6721757

Level of		life	
type	N	Mean	Std Dev
1	12	83.166667	48.5888751
2	12	108.333333	49.4723676
3	12	125.083333	35.7655455

Level of	Level of		life	
temp	type	N	Mean	Std Dev
15	1	4	134.750000	45.3532432
15	2	4	155.750000	25.6173769
15	3	4	144.000000	25.9743463
70	1	4	57.250000	23.5990819
70	2	4	119.750000	12.6589889
70	3	4	145.750000	22.5444006
125	1	4	57.500000	26.8514432
125	2	4	49.500000	19.2613603
125	3	4	85.500000	19.2786583

a. Estimate grand mean μ, main effect τ_1 (temperature=15), and interaction effect $(\tau\beta)_{23}$ (temperature=70, material type=3).

The ANOVA table from SAS is given on the next page, with some values covered by $\star\star\star$. *[NO need to calculate all these values.]*

b. Do the effects of temperatures on the battery life *depend* on the material types? Conduct an appropriate test to answer this question. To get full credits, give hypotheses, a test statistic, determine its degrees of freedom, state the p-value, and give your conclusion using $\alpha = 0.05$.

c. Use Tukey's method to perform a pairwise comparison for *all treatment combinations*. 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)	$(3, 26)$	$(3, 27)$	$(4, 26)$	$(4, 27)$	$(8, 26)$	$(8, 27)$	$(9, 26)$	$(9, 27)$
$q_{0.025;df_1,df_2}$	3.9580	3.9472	4.3177	4.3046	5.0838	5.0651	5.2049	5.1852
$q_{0.05;df_1,df_2}$	3.5142	3.5064	3.8796	3.8701	4.6519	4.6378	4.7733	4.7584

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	***	59416.22222	7427.02778	***	<.0001
Error	***	***	***		
Corrected Total	***	77646.97222			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
temp	***	39118.72222	***	***	<.0001
type	***	10683.72222	***	***	0.0020
temp*type	***	****	***	***	0.0186

d. Now assume that it is given that *Temperature=70*. We need to carry out multiple comparison on the material type effects using Bonferroni method with $\alpha = 0.06$. To get full credits, you just need to report the critical difference *without the comparison results*.

df	3	4	26	27
$t_{0.005;df}$	5.8409	4.6041	2.7787	2.7707
$t_{0.01;df}$	4.5407	3.7469	2.4786	2.4727
$t_{0.02;df}$	3.1824	2.7764	2.1620	2.1578
$t_{0.06;df}$	2.3534	2.1318	1.6076	1.6056