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, respectively.

   (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. Let $X_1, \ldots, X_n$ be a random sample from a geometric distribution with the probability mass function

$$P(X = x) = (1-p)^{x-1} p,\ x = 1, 2, \ldots;\ 0 < p < 1.$$

   (a) Find the maximum likelihood estimator (MLE) $\hat{p}$ of p.

   (b) Find the Fisher information $I_1(p)$.

   (c) Find the Rao-Cramér lower bound (RCLB) for any unbiased estimator of $1/p$.

   (d) Is $1/\hat{p}$ an efficient estimator of $1/p$? Clearly justify your answer.

3. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with the probability density function (pdf)

$$f(x; \theta) = \theta^2 x e^{-\theta x},\ 0 < x < \infty,\ 0 < \theta < \infty.$$

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

   (b) Find the function of Y that is the unique minimum variance unbiased estimator (MVUE) of θ.

   (c) Using Basu's Theorem, determine $E(X_1/Y)$. Clearly justify your answer.

4. Let X_i be a random variable with pdf $f(x; \theta)$.

· We assume that $f(x;\theta)$ is twice differentiable with respect to θ, and $\int f(x;\theta)d\theta$ is twice differentiable under the integral sign with respect to θ.

(a) Prove that $E\left(\frac{\partial \log f(X_i;\theta)}{\partial \theta}\right) = 0$.

(b) The Fisher information is defined as the variance of the score function, i.e., $I(\theta) = V\left(\frac{\partial \log f(X_i;\theta)}{\partial \theta}\right)$.

Show that the Fisher information can also be expressed as

$$V\left(\frac{\partial \log f(X_i;\theta)}{\partial \theta}\right) = -E\left(\frac{\partial^2 \log f(X_i;\theta)}{\partial \theta^2}\right).$$

(c) Let X_1 and X_2 are iid with common pdf $f(x;\theta)$. Let $Y = u(X_1, X_2)$.

Show that, under some regularity conditions,

$$V(Y) \geq \frac{\left[\frac{\partial E(Y)}{\partial \theta}\right]^2}{2I(\theta)}$$

5. A maker of asphalt shingles is interested in the relationship between sales for 1995 and factors that influence sales. He gathered data from 15 randomly collected different districts. The regressor (covariate) variables are as follows: Promotional accounts (X_1); Active accounts (X_2); Competing Brands (X_3); Potential (X_4). The respond is Y, the Sales (in thousands). A multiple linear regression model has been fitted. Part of the SAS output is listed as in Table 1. Also, a sequence of models has been fitted. The model error sums of squares (SSE) have been recorded in Table 2. All models include the intercept.

   When answering questions related to hypothesis testing, please clearly state your null and alternative hypotheses, test statistic, the distribution under the null hypothesis, the decision rule, and your conclusion, in order to receive full credit. Use $\alpha = 0.05$.

(a) Write down the fitted model when the assumed model is

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

(b) Calculate R^2 for the regression model in Part (a). Interpret it.

(c) Perform the F test to compare the two models:

$$\text{Model 1: } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

$$\text{Model 2: } Y = \beta_0 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

(df_1, df_2)	$(2, 10)$	$(2, 14)$	$(4, 10)$	$(4, 14)$
$F_{0.05;df_1,df_2}$	4.1028	3.7389	3.4781	3.1123
$F_{0.025;df_1,df_2}$	5.4564	4.8567	4.4683	3.8919

(d) Use the forward selection method to select the best model. Use $\alpha = 0.05$.

(df_1, df_2)	$(1, 9)$	$(1, 10)$	$(1, 11)$	$(1, 12)$	$(1, 13)$	$(1, 14)$	$(1, 15)$
$F_{0.05;df_1,df_2}$	5.1174	4.9646	4.8443	4.7472	4.6672	4.6001	4.5431
$F_{0.025;df_1,df_2}$	7.2093	6.9367	6.7241	6.5538	6.4143	6.2979	6.1995

Table 1:

Source	DF	Squares	Square	F Value	Pr > F
Model	4	*******	*******	*******	*******
Error	10	262	*******		
Corrected Total	14	89547			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >
Intercept	1	177.2286	8.7874	20.169	1.98e-09
X1	1	2.1702	0.6737	3.221	0.00915
X2	1	3.5380	0.1092	32.414	1.84e-11
X3	1	-22.1583	0.5454	-40.630	1.95e-12
X4	1	0.2035	0.3189	0.638	0.53760

Table 2: SSE for the model fitted

Number in Model	SSE	Variables in Model
1	81233.40	X4
1	88472.74	X1
1	44682.67	X2
1	32482.88	X3
2	79241.41	X1 X4
2	44395.78	X2 X4
2	28609.71	X3 X4
2	32086.00	X1 X3
2	43967.92	X1 X2
2	534.66	X2 X3
3	43524.28	X1 X2 X4
3	27796.58	X1 X3 X4
3	534.02	X2 X3 X4
3	272.75	X1 X2 X3
4	262.07	X1 X2 X3 X4

6. Researchers were interested in studying the effect of temperature and light level on the growth of bacterial colonies on potato leaflets. Bacteria were inoculated onto a total of 48 leaflets. The leaflets were randomly assigned to treatment with one of four temperatures (10, 15, 20, or 25 °C) and one of three light levels (A=low, B=medium, or C=high). Four weeks after inoculation, the log of the area of the bacterial colony on each leaflet was measured as the response variable. A completely randomized design was used with four leaflets for each combination of temperature and light intensity. Use the SAS code and output on the next page of your exam to answer the following questions.

(a) Were there significant differences among the 12 treatment means? Give an appropriate test statistic, its degrees of freedom, the p-value, and a brief conclusion.

(b) Were there any significant differences among the temperature lsmeans? Give an appropriate test statistic, its degrees of freedom, the p-value, and a brief conclusion.

(c) Two models have been fit to the data. Does the second model fit the data adequately? Give an appropriate test statistic, its degrees of freedom, the p-value, and a brief conclusion.

Now, for each of the three light intensities, suppose there is a linear relationship between the mean of the response variable and temperature.

(d) For low light level A, provide the estimated linear regression equation relating mean response to temperature.

(e) For high light level C, give an 95% confidence interval for the slope of the linear regression equation. Based on this confidence interval, is there evidence that temperature affected bacterial colony growth at high light level C? Explain.

(f) Estimate the difference between the slope for low light level A and the slope for high light level C, and determine if that difference is significantly different from 0. Provide the estimated difference, a test statistic, its degrees of freedom, the p-value, and a brief conclusion.

```
proc glm;
  class light temp;
  model y=light temp light*temp;
run;
```

```
      Class Level Information
Class          Levels    Values
light               3    A B C
temp                4    10 15 20 25
Number of observations    48
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	2420.799306	220.072664	8.50	<.0001
Error	36	932.134425	25.892623		
Corrected Total	47	3352.933731			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
light	2	1588.672888	794.336444	30.68	<.0001
temp	3	441.252123	147.084041	5.68	0.0027
light*temp	6	390.874296	65.145716	2.52	0.0389

Source	DF	Type III SS	Mean Square	F Value	Pr > F
light	2	1588.672887	794.336444	30.68	<.0001
temp	3	441.252123	147.084041	5.68	0.0027
light*temp	6	390.874296	65.145716	2.52	0.0389

```
proc glm;
  class light;
  model y=light temp light*temp / solution;
run;
```

```
   Class Level Information
Class          Levels    Values
light               3    A B C
Number of observations    48
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2224.265059	444.853012	16.55	<.0001
Error	42	1128.668673	26.873064		
Corrected Total	47	3352.933731			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
light	2	1588.672888	794.336444	29.56	<.0001
temp	1	373.276984	373.276984	13.89	0.0006
temp*light	2	262.315188	131.157594	4.88	0.0124

Source	DF	Type III SS	Mean Square	F Value	Pr > F
light	2	12.3292407	6.1646204	0.23	0.7960
temp	1	373.2769837	373.2769837	13.89	0.0006
temp*light	2	262.3151875	131.1575938	4.88	0.0124

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		2.044500000	4.25902782	0.48	0.6337
light	A	-2.307500000	6.02317490	-0.38	0.7036
light	B	1.760000000	6.02317490	0.29	0.7716
light	C	0.000000000	.	.	.
temp		0.276600000	0.23183211	1.19	0.2395
temp*light	A	0.808000000	0.32786011	2.46	0.0179
temp*light	B	-0.141250000	0.32786011	-0.43	0.6688
temp*light	C	0.000000000	.	.	.

7. An experiment is run to investigate the effect of three drugs (D1, D2, and D3) in bringing behavioral changes in two type of mental illnesses schizophrenics(SZ) and depressive (DP). The effeteness of the drugs is tested to 24 patients, 12 for each type of mental illness. For each group of patients, the three drugs are randomly assigned to the 12 patients, 4 for each drug. The data from this experiment are given to the table below.

Mental Illness	Drug (D)		
(MI)	D1	D2	D3
SZ	16, 15, 13, 15	19, 16, 17, 20	23, 18, 16, 18
DP	19, 14, 16, 16	21, 25, 19, 22	24, 27, 30, 21

(a) Write down the model for this experiment along with appropriate assumptions and constraints.

(b) Fill in the missing values labeled (1)-(11) in the SAS GLM output. Show the necessary steps. Note: you do not need to calculate in the order of the labeled missing values.

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	(3)	(4)	(6)	(8)	0.0013
Error	(2)	84.6666667	(7)		
Corrected Total	(1)	(5)			

Source	**DF**	**Type III SS**	**Mean Square**	**F Value**	**Pr > F**
MI	1	(9)	98.0000000	13.89	0.0029
D	2	169.0000000	(10)	(11)	0.0014
MI*D	2	30.3333333	15.1666667	2.15	0.1593

(c) Estimate the main effect of the the first drug (D1).

(d) Carry out the appropriate hypothesis testing procedure on the interaction effect. Clearly state the null and alternative hypotheses, test statistic, degrees of freedom, p-value, and conclusions. Use $\alpha = 0.05$.

(e) Is it meaningful to discuss the main effects of the drug? Explain. If yes, use the Tukey's method to compare the three drugs pairwisely. Use $\alpha = 0.05$.

(df_1, df_2)	$(2, 18)$	$(2, 19)$	$(3, 18)$	$(3, 19)$
$q_{0.025;df_1,df_2}$	3.4578	3.4414	4.0915	4.0682
$q_{0.05;df_1,df_2}$	2.9712	2.9600	3.6093	3.5927

8. The effect of five different ingredients (A, B, C, D, E) on the reaction time of a chemical process is being studies. Each batch of new material is only large enough to permit five runs to be made. Furthermore, each run requires approximately 1.5 hours, so only five runs can be made in one day. The experiment layout and results are given below.

	Day				
Batch	1	2	3	4	5
1	A=6	B=6	D=1	C=8	E=5
2	C=10	E=2	A=8	D=5	B=11
3	B=2	A=8	C=10	E=2	D=7
4	D=1	C=4	E=3	B=4	A=8
5	E=2	D=1	B=3	A=9	C=10

The grand mean $\bar{Y}_{\cdots} = 5.44$. And the level means for batch, day and ingredient are given in Figure 1.

Level of batch	N	Mean	Level of day	N	Mean	Level of Ingredient	N	Mean
1	5	5.20	1	5	4.20	1	5	7.80
2	5	7.20	2	5	4.20	2	5	5.20
3	5	5.80	3	5	5.00	3	5	8.40
4	5	4.00	4	5	5.60	4	5	3.00
5	5	5.00	5	5	8.20	5	5	2.80

Figure 1: The means of the data in Problem 8.

(a) What design is employed for this experiment?

(b) Part of the SAS output for ANOVA is given below in Figure 2. Test if the five ingredients have different effect on the chemical time. To get full credits, give hypotheses, the test statistic, p-value, degrees of freedom, and your conclusion (use $\alpha = 0.05$).

(c) Use Bonferroni method for treatment pairwise comparison. Calculate the critical difference **AND** draw your conclusion (use $\alpha = 0.1$).

(d) Assume that five operators are employed to conduct the experiment, and it is know that the operator can influence the experimental results. Derive an experimental plan that can be used to study the five ingredients using five batches, five operators, and in five days. Some useful squares are given below in Figure 3.

(df_1, df_2)	$(4, 12)$	$(4, 24)$	$(5, 12)$	$(5, 24)$
$F_{0.05;df_1,df_2}$	3.26	2.78	3.11	2.62
$F_{0.025;df_1,df_2}$	4.12	3.38	3.89	3.15

df	4	5	12	24
$t_{0.1;df}$	1.53	1.48	1.36	1.32
$t_{0.05;df}$	2.13	2.02	1.78	1.71
$t_{0.025;df}$	2.78	2.57	2.18	2.06
$t_{0.01;df}$	3.75	3.36	2.68	2.49
$t_{0.005;df}$	4.60	4.03	3.06	2.80

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	218.880	18.2400	5.57	0.0029
Error	12	39.280	3.2733		
Co. Total	24	258.160			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
batch	4	27.760	6.940	2.12	0.1410
day	4	54.560	13.640	4.17	0.0241
ingredient	*	******	******	****	******

Figure 2: Part of ANOVA for Problem 8.

A B C D E	A B C D E	A B C D E	A B C D E
B C D E A	C D E A B	E A B C D	D E A B C
C D E A B	E A B C D	D E A B C	B C D E A
E A B C D	D E A B C	B C D E A	C D E A B
D E A B C	B C D E A	C D E A B	E A B C D

Figure 3: Some squared that may be useful in Problem 8.