Statistics Qualifying Examination

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

1. Let X_1 and X_2 have the joint probability density function (pdf)

$$f_{X_1,X_2}(x_1,x_2) = \begin{cases} 15x_1^2 x_2 & 0 < x_1 < x_2 < 1, \\ 0 & \text{elsewhere.} \end{cases}$$

(a) Calculate the probability Pr$(X_1 + X_2 \leq 1)$.

(b) Find the marginal pdf of X_2, $f_{X_2}(x_2)$.

(c) Show that $f_{X_2}(x_2)$ is a (legitimate) probability density function.

(d) Calculate $E\left(\frac{1}{X_2}\right)$.

2. Suppose that the random variable X has a Poisson distribution such that

$$50 \cdot \text{Pr}(X = 1) = \text{Pr}(X = 2).$$

(a) Find the mean and the variance of X.

(b) Calculate the probability Pr$(X \geq 1)$.

(c) Calculate the probability Pr$(75 < X < 125)$ using the Chebyshev's inequality.

- Chebyshev's inequality: If X has a finite variance σ^2 with mean μ, Pr$(|X - \mu| \geq k\sigma) \leq 1/k^2$ for every $k > 0$.

3. Let $X_1, \ldots, X_n$ be a random sample from a Possion distribution with the probability mass function as $P(X = x) = \frac{e^{-\theta}\theta^x}{x!}$, $x = 0, 1, 2, \ldots$, and $0 < \theta < \infty$.

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

(b) Is the MLE $\hat{\theta}$ an efficient estimator of θ? Clearly justify your answer.

(c) Find the MLE $\hat{\tau}$ of $\tau = P(X \leq 1)$.

(d) Determine the limiting distribution of $\sqrt{n}(\hat{\tau} - \tau)$.

4. Let $X_1, \ldots, X_n$ be a random sample from a $Gamma(3, \theta)$ distribution, where $0 < \theta < \infty$.

(a) Find the exact likelihood ratio test of size α for testing the hypotheses $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$. Clearly state the likelihood ratio, the rejection region, and the decision rule.

(b) For $\theta_0 = 3$ and $n = 5$, specify the rejection region so that the test that rejects the null hypothesis has a significant level 0.05.

(c) Obtain the power function of the test in (a) with $\theta_0 = 3$, $n = 5$, and $\alpha = 0.05$.

5. Consider the normal error regression model as follows: $Y_i = \beta X_i + \epsilon_i$, $i = 1, \ldots, n$ where X_i's are known constants, ϵ_i's are $i.i.d.$ $N(0, \sigma^2)$, and β and σ^2 are parameters to be estimated.

(a) Derive the maximum likelihood estimators (MLE) of β and σ^2, say $\hat{\beta}$ and $\hat{\sigma}^2$, respectively.

(b) Show that $\hat{\beta}$ is a linear combination of Y_i's, i.e, $\hat{\beta} = \sum_{i=1}^{n} k_i Y_i$ and that $\hat{\beta}$ is an unbiased estimator (UE) of β .

(c) Show that $Var(\hat{\beta}) = \sigma^2 / \sum_{i=1}^{n} X_i^2$.

(d) Let $e_i = Y_i - \hat{Y}_i$, $i = 1, \ldots, n$. Show that $\sum_{i=1}^{n} X_i e_i = 0$.

6. The electric power consumed each month by a chemical plant is thought to be related to the average ambient temperature X_1, the number of days in the month X_2, the average product purity X_3, and the tons of product produced X_4. The past year's historical data are available. Consider all 4 variables as covariates in the multiple regression model with normal errors. Use the following information,with $n = 12$.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model		4957.24074			
Error					
Corrected Total		6656.25000			

Parameter Estimates

	Parameter	Standard			
Variable	DF	Estimate	Error	t Value	$Pr > \|t\|$
Intercept	1	-102.71324	207.85885	-0.49	0.6363
x1	1	0.60537	0.36890	1.64	0.1448
x2	1	8.92364	5.30052	1.68	0.1361
x3	1	1.43746	2.39162	0.60	0.5668
x4	1	0.01361	0.73382	0.02	0.9857

Number in Model	R-Square	C(p)	AIC	MSE	SSE	Variables in Model
1	0.6446	1.7471	67.4074	236.57679	2365.76786	x2
1	0.5647	3.9371	69.8397	289.73308	2897.33079	x1
1	0.0024	19.3586	79.7922	664.03676	6640.36765	x3
1	0.0001	19.4218	79.8198	665.56870	6655.68695	x4
2	0.7314	1.3665	66.0471	198.66239	1787.96148	x1 x2
2	0.6463	3.6989	69.3479	261.56262	2354.06360	x2 x3
2	0.6447	3.7437	69.4032	262.77130	2364.94169	x2 x4
2	0.6412	3.8385	69.5194	265.32872	2387.95845	x1 x3
3	0.7447	3.0003	67.4353	212.38659	1699.09274	x1 x2 x3
3	0.7316	3.3612	68.0386	223.33621	1786.68970	x1 x2 x4
3	0.6466	5.6930	71.3406	294.07956	2352.63647	x2 x3 x4
3	0.6414	5.8343	71.5143	298.36753	2386.94025	x1 x3 x4
4	0.7447	5.0000	69.4347	242.71561	1699.00926	x1 x2 x3 x4

(a) Predict power consumption for a month and compute a 95% confidence interval for the predicted power consumption when $X_1 = 75$, $X_2 = 24$, $X_3 = 90$, and $X_4 = 98$.

(b) Test the hypothesis $H_0 : \beta_1 = \beta_3 = 0$ v.s. H_1 : not H_0.

(c) Test the hypothesis $H_0 : \beta_1 = \beta_3 = 0$ v.s. $H_1 : \beta_1 = 0$.

(d) Perform a stepwise regression using a $\alpha = .05$ level of significance.

7. Disk drive substrates may affect the amplitude of the signal obtained during readback. A manufacturer compares four substrates: aluminum(A), nickel-plated aluminum (B), and two types of glass (C and D). Sixteen disk drives will be made, four using each of the substrates. The design responses (in microvolts) are given in the following table (data from Nelson 1993; Greek letters indicate day)

	Operator			
Machine	1	2	3	4
1	$A\alpha = 8$	$C\gamma = 11$	$D\delta = 2$	$B\beta = 8$
2	$C\delta = 7$	$A\beta = 5$	$B\alpha = 2$	$D\gamma = 4$
3	$D\beta = 3$	$B\delta = 9$	$A\gamma = 7$	$C\alpha = 9$
4	$B\gamma = 4$	$D\alpha = 5$	$C\beta = 9$	$A\delta = 3$

The grand mean is $\bar{Y}_{\cdots} = 6$, and the level means for the four substrates are:

$$A : 5.75 \quad B : 5.50 \quad C : 9.00 \quad D : 3.75$$

(a) What kind of design is used for the experiment? Give a model and state the assumptions.

(b) Calculate the estimates of the treatment (i.e. the substrates) effects.

(c) Part of the ANOVA table from SAS is given below. Test if the *substrates* are different from each other in terms of their effects on the response. To get full credits, state the hypotheses, obtain the test statistic, and draw your conclusion ($\alpha = 0.05$).

Source	DF	Squares	Mean Square	F Value	Pr > F
Model	12	100.5000000	8.3750000	1.17	0.5098
Error	3	21.5000000	7.1666667		
Corrected Total	15	122.0000000			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
row	??	21.50000000	7.16666667	1.00	0.5000
col	??	14.00000000	4.66666667	0.65	0.6335
substrate	??	**??**	**??**	**??**	**??**
greek	??	3.50000000	1.16666667	0.16	0.9149

(df_1, df_2)	$(3,3)$	$(3,12)$	(3,15)	$(4,3)$	$(4,12)$	(4,15)
$F_{0.025;df_1,df_2}$	15.4392	4.4742	4.1528	15.1010	4.1212	3.8043
$F_{0.05;df_1,df_2}$	9.2766	3.4903	3.2874	9.1172	3.2592	3.0556

(d) If the machine were not considered as blocks and not included in ANOVA model, will the test results in part (c) change? Justify your answer.

8. An engineer suspects that the surface finish of a metal part is influenced by the feed rate and the depth of cut. He selects three feed rates and four depths of cut. He then conducts a factorial experiment and obtains the following data.

Feed rate (in/min)	1	Depth of 2	Cut (in) 3	4
1	74, 64, 60	79, 68, 73	82, 88, 92	99, 104, 96
2	92, 86, 88	98, 104, 88	99, 108, 95	104, 110, 99
3	99, 98, 102	104, 99, 95	108, 110, 99	114, 111, 107

Some summary statistics are give below.

```
Grand mean: 94.333

feed        mean     |    depth   mean
1          81.583    |    1        84.778
2          97.583    |    2        89.778
3         103.833    |    3        97.889
                     |    4       104.889
------------------------------------------------
feed    depth       mean
1       1           66.000
1       2           73.333
1       3           87.333
1       4           99.667
2       1           88.667
2       2           96.667
2       3          100.667
2       4          104.333
3       1           99.667
3       2           99.333
3       3          105.667
3       4          110.667
```

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

The ANOVA of the data was done in SAS and the output is shown.

(a) What are the estimates of τ_1 and $(\tau\beta)_{22}$?

(b) Test if the interaction between feed rate and cut depth is significant. To get full credits, give a test statistic, the corresponding p-value, and your conclusion.

```
Dependent Variable: finish
                 Sum of
Source    DF    Squares      Mean Square    F Value  Pr > F
Model     11     5842.667     531.151515     18.49    <.0001
Error     24      689.333      28.722222
Co Total  35     6532.000

ource     DF     Type I SS     Mean Square  F Value    Pr > F
feed       2     3160.500      1580.250      55.02    <.0001
depth      3     2125.111       708.370      24.66    <.0001
feed*depth 6      557.056        92.843       3.23    0.0180
```

(c) If we plan to perform pairwise comparison for *all* treatment combinations, which procedure should we use? What is the corresponding critical difference (using $\alpha = 5\%$)? *Note:* No table of critical values from distributions is given for this part, and thus you don't need to calculate the final numerical answer. Instead, please give the formula of the critical difference and specify the components in the formula, including but not limited to the degrees of freedom, significance level, and the distribution.

(d) Use the Bonferroni method to compare the following treatments (i.e. these *four specific* level combinations of speed and depth): (2,3), (2,4), (3,3) and (3,4), *pairwisely* (Use $\alpha = 6\%$). Calculate the critical difference and report your results of comparison. You can report the result as we have seen in SAS output by labeling significantly different combinations with different Latin letters.

df	2	3	6	24
$t_{0.005;df}$	9.9248	5.8409	3.7074	2.7969
$t_{0.01;df}$	6.9646	4.5407	3.1427	2.4922
$t_{0.015;df}$	5.6428	3.8960	2.8289	2.3069
$t_{0.03;df}$	3.8964	2.9505	2.3133	1.9740