Statistics Qualifying Examination

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

1. Let a random variable Y have the cumulative distribution function

$$F_Y(y) = \begin{cases} 1 - \frac{1}{y^2}, & \text{if } 1 \leq y < \infty; \\ 0, & \text{if } y < 1. \end{cases}$$

   (a) Show that $F_Y(y)$ is a "legitimate" cumulative distribution function (cdf) of a continuous random variable, i.e., it satisfies the conditions to be defined as a cdf.

   (b) Find the probability density function of Y.

   (c) Let $Z = 10(Y - 1)$. Find the cdf of Z, i.e., $F_Z(z)$.

2. Let X follow the standard normal distribution, i.e., with mean 0 and standard deviation 1. Define a new random variable $Y = |X|$.

   (a) Find the probability density function of Y.

   (b) Find the mean of Y.

   (c) Find the variance of Y.

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

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

   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 $X_1, \ldots, X_n$ be identically and independently distributed Poisson random variables with the probability mass function as $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$, $x = 0, 1, 2, \ldots$ and $0 < \lambda < \infty$.

(a) Show that $\sum_{i=1}^{n} X_i$ is a sufficient statistic for λ using the Neyman Factorization Theorem.

(b) Consider the statistics $W = I(X_1 = 1)$, which is defined as $W = 1$, if $X_1 = 1$ and $W = 0$, elsewhere. Show that W is an unbiased estimator for $\theta = \lambda \exp(-\lambda)$.

(c) Find the Rao-Blackwellized version of the estimator W for θ. Clearly justify all your steps.

5. Consider the general linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$ where $\mathbf{y}$ is an $n \times 1$ vector of response, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^T$ is a $p \times 1$ vector of parameters, $\mathbf{X}$ is an $n \times p$ full rank matrix of predictor variables, and ϵ is an $n \times 1$ vector of independent normal random variables with its expectation $\mathbf{0}_{n\times 1}$ and variance-covariance matrix $\sigma^2\mathbf{I}_n$. Here, $\mathbf{I}_n$ denotes the $n \times n$ identity matrix, and $\mathbf{A}^T$ denote the transpose of the matrix $\mathbf{A}$.

(a) Derive the normal equation for the least square (LS) estimation of $\boldsymbol{\beta}$ and find the LS estimator $\hat{\boldsymbol{\beta}}$, of $\boldsymbol{\beta}$.

(b) Show that $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

(c) The residual is defined as $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$. Show that $\mathbf{e} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$ and its variance is equal to $\sigma^2(\mathbf{I}_n - \mathbf{H})$ where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

(d) Show that, for $i \neq j$, e_i and e_j are NOT independent of each other in general, where $\mathbf{e} = (e_1, e_2, \ldots, e_n)^T$.

6. A commercial real estate company evaluates vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data are taken from 81 suburban commercial properties that are the newest, best located, most attractive, and expensive for five specific geographic areas. Shown here are the age (X_1), operating expenses and taxes (X_2), vacancy rates (X_3), total square footage (X_4), and rental rates (Y). Use the SAS code and its partial results in the handout.

(a) Obtain the regression sum of squares (SS) and the following extra SS: extra SS associated with X_4; with X_1, given X_4; with X_2, given X_1 and X_4; and with X_3, given X_1, X_2 and X_4. Verify that the the regression sum of squares (also called model sum of squares) can be decomposed into these extra SS.

(b) Test whether X_3 can be dropped from the regression model given that X_1, X_2 and X_4 are retained. Use the F test statistic and level of significance $\alpha = 0.01$. State the alternatives, decision rule, and conclusion.

(c) Test whether both X_2 and X_3 can be dropped from the regression model given that X_1 and X_4 are retained. Use $\alpha = 0.01$. State the alternatives, decision rule, and conclusion.

(d) Find the best regression models based on each of following measures: R^2, Mallows' C_p, Akaike Information Criterion (AIC) and MSE.

(e) Perform a stepwise regression selection with the level of significance at $\alpha = 0.05$.

(f) In the case that you have to pick one regression model as the best model to predict rental rates, what would be your recommendation? Justify your answer.

7. An experiment was conducted to assess the yield of a manufacturing process in a chemical factory. Seven 2-level factors (denoted by A, B, ..., G) are considered in the experiment. **Suppose that it is reasonable to assume that factorial effects of order 3 or higher are negligible**. The research decide to design a 2^{7-3} fractional factorial, defined by $E = BCD$, $F = ACD$, $G = ABD$.

(a) How many runs (observations) are needed for this experiment? Can we have observation for $(A, B, C, D, E, F, G) = (-, +, +, -, -, +, -)$?

(b) What is the complete defining relation of this design, and what is the resolution of this experiment?

(c) Please find out the **non-negligible** factorial effects that is aliased with main effect A, and **non-negligible** factorial effects that is aliased with interaction effect AB.

(d) Assume the effect corresponding to factorial effect AB has estimated value 12.8. Please interpret this estimation.

8. The percentage of hardwood concentration (HC) in raw pulp and the vat pressure are being investigated for their effects on the strength of paper. Three levels of hardwood concentration and three levels of pressure are selected. A factorial experiment with three replicates is conducted, and the following data are obtained:

	Pressure		
HC Percentage	400	500	600
2	24.9, 26.7, 23.2	27.6, 29.3, 26.3	54.3, 52.5, 55.6
4	31.5, 28.8, 25.6	37.0, 41.4, 44.0	34.0, 35.4, 42.8
6	20.4, 25.1, 26.1	35.0, 38.0, 27.0	49.6, 43.6, 53.0

grand mean: 35.51

HC Percent	MEAN	Pressure	MEAN
1	35.60	1	25.81
2	35.61	2	33.96
3	35.31	3	46.76

HC Percent	Pressure	MEAN
1	1	24.93
1	2	27.73
1	3	54.13
2	1	28.63
2	2	40.80
2	3	37.40
3	1	23.87
3	2	33.33
3	3	48.73

Some summary statistics are give above. Notes that for HC percentage level, 2, 4, 6 are coded as 1, 2, 3, respectively in SAS, while Pressure level 400, 500, 600 as 1, 2, 3, respectivly.

Suppose the following statistical model is considered:

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

where τ_i ($i = 1, 2, 3$), β_j ($j = 1, 2, 3$) and $(\tau\beta)_{ij}$ are the main effects of HC percentage, the main effects of pressure, and their interactions, respectively, which satisfy the additional constraints as in STAT 6032 lecture notes: $\sum_i \tau_i = \sum_j \beta_j = \sum_i (\tau\beta)_{ij} = \sum_j (\tau\beta)_{ij} = 0$.

Part of the ANOVA from SAS is given below.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	2739.531852	342.441481	26.66	<.0001
Error	18	231.206667	12.844815		
Cor.Total	26	2970.738519			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
percent	2	0.520741	0.260370	0.02	0.9800
pressure	*	*******	********	*****	******
percent*pressure	*	*******	********	*****	******

(a) Calculate the sum of squares due to the main effects of pressure.

(b) Test if the interaction between HC percentage and pressure is significant ($\alpha = 0.05$).

(c) Use the Bonferroni method to compare the following treatments, i.e., level combinations of the two factors, (HC percentage, pressure): (2,1), (2,2), (2,3) and (3,2), *pairwisely*. Calculate the critical difference and report the comparison results ($\alpha = 6\%$).