Statistics Qualifying Examination

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

1. Let $X_i \sim$ Gamma(shape $= \alpha$, rate $= \lambda$) whose mean is α/λ, for $i = 1, \ldots, n$.

   (a) Find the moment generating function (mgf) of X_i.

   (b) Show that the variance of X_i is α/λ^2.

   (c) Find the maximum likelihood estimator (MLE) of λ when α is known.

2. Let X_i be a random variable with probability density function (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) (Rao-Cramér lower bound) 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,

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

   - Hint: Find the random variable $Z = \frac{\partial \log f(X_1;\theta)}{\partial\theta} + \frac{\partial \log f(X_2;\theta)}{\partial\theta}$, and use the fact that $Cov(Y,Z) = E(YZ) - E(Y)E(Z)$ and $Cov(Y,Z)^2 \leq V(Y)V(Z)$.

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

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

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

(a) Derive the cumulative distribution function 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.

4. Suppose $X_1, \ldots, X_n$ is a random sample from a continuous distribution with pdf

$$f(x;\theta) = \frac{1}{2}\theta^3 x^2 e^{-\theta x},$$

for $0 < x < \infty, 0 < \theta < \infty$, and zero elsewhere.

(a) It is known that the MLE for θ is $3n/\sum_{i=1}^n X_i$. Is the MLE an unbiased estimator for θ? Clearly justify your answer.

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

(c) Find the minimum variance unbiased estimator (MVUE) for θ.

(d) Show that the likelihood ratio Λ for testing $H_0 : \theta = 1$ vs $H_1 : \theta \neq 1$ is a function of $W = 2\sum_{i=1}^n X_i$.

(e) Specify the sampling distribution of W under H_0.

(f) Find the rejection region for an exact likelihood ratio test of size α.

5. The following is part of an ANOVA table for a simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where ε_i's are i.i.d. from $N(0, \sigma^2)$, $i = 1, \ldots, n$, and n is the number of observations. (You don't need to complete the ANOVA table.)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	***	******	252378	105.88	<.0001
Error	23	******	******		
Corrected Total	***	******			

(a) Compute the coefficient of determination R^2.

(b) Assume that we know the least square estimate of β_1 is $b_1 = 3.57$. Construct a two-sided t-test of whether or not $\beta_1 = 3$ ($\alpha = 0.05$). To get full credit, give the null and alternative hypotheses, the value of the test statistic, the sampling distribution under the null hypothesis, the (range of) p-value, and your conclusion.

$t_{0.025;23}$	$t_{0.025;24}$	$t_{0.025;25}$	$t_{0.05;23}$	$t_{0.05;24}$	$t_{0.05;25}$
2.0687	2.0639	2.0595	1.7139	1.7109	1.7081

6. For 50 high schools in Montana, researchers examined the relationship between

- Y, a school performance measure (based on test outcomes for a random group of recent graduates of the school), measured on a scale from -100 to 100 points

and predictors X_1 and X_2, where

- X_1 is a continuous variable that represents the quality of teachers in the school (centered such that $X_1 = 0$ for a school with teachers of average quality),
- X_2 is an indicator variable that refers to school size, where $X_2 = 1$ for a large school, and $X_2 = 0$ for a small school (based on number of students). Results of a fitted regression model with X_1, X_2 and an interaction between X_1 and X_2 are below. Some output from fitting this model in R is given below.

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \epsilon_i; i = 1, \ldots, n;$$

where $n = 50$. We assume $\epsilon_i \sim N(0, \sigma^2)$ independently.

```
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept)  0.2279   0.1093       2.086      0.0383 *
  X1         2.9695   0.2016      14.728      < 2e-16 ***
  X2        -1.1965   0.1519      -7.877      2.22e-13 ***
  X1:X2     -2.0134   0.2650      -7.599      1.19e-12 ***
```

(a) Write down the expression for the *fitted* regression equations for the two types of schools (use numbers from the output instead of $\hat{\beta}$'s), respectively.

(b) Interpret the estimate $\hat{\beta}_3$. Please use words to relate to the context of this data analysis instead of using notations in Y's or X's in your interpretation.

(c) If assumptions for the model above hold true, do we have evidence from this data analysis that large schools are doing worse than small schools with respect to students' test scores, regardless of teacher quality? You don't need to perform a formal hypothesis testing, but you need to explain your answer based on the output provided.

(d) Suppose that there is an alternative categorization system, in which schools are classified into three groups: very small, medium, and very large schools (thus some of the small and large schools from the former categorization end up in the medium category). Generalize the previous model for this new categorization system by defining indicator variable(s) and giving the model and assumptions. Be sure to clearly define your indicator variable(s) associated with schools of different sizes.

(e) How can the researchers decide if they should choose between the original model vs. the second in part (d)? Suggest and explain an exploratory approach (e.g., using residual plots) OR a quantitative approach (e.g., using a test or some other measure of model fit).

7. A manufacturer of commercial fishing nets produces the net material on a large number of machines. The company would like the machines to be as homogeneous as possible in order to produce netting that has a very uniform strength. The company is concerned that there is considerable amount of material that must be discarded, reworked, or downgraded to a lower quality product. This results in loss of revenue to the company. The process engineer, after examining the maintenance records of the machines, suspects that there may be a larger variation in material strength between machines relative to the usual variation in material produced from the same machine. The two sources of variation are measured by σ_M^2, the detectable variation between machines, referred to as the "Special Cause" variability, the variability due to Machine Differences. The second source of variability is measured by σ^2, the background noise or variability in the process, referred to as the "Common Cause" variability. The engineer decides to run a small pilot study by *randomly selecting four machines* and then selecting five samples of material from each of the selected machines for strength determinations. The 20 samples of material were then analyzed in random order. The results are given here.

	Strength Measurements: Y_{ij}					
Machines	1	2	3	4	5	Mean: $\bar{Y}_{i.}$
1	128	127	129	126	128	127.6
2	121	120	123	122	125	122.2
3	126	125	127	125	124	125.4
4	125	126	129	128	127	127.0

(a) What type of design is this? Write down the model with the model assumptions.

(b) Find the values for A, B, and C in the ANOVA table bellow. (You don't need to calculate D, E, or F when answering this part.)

Source	DF	Sum of Squares	Mean Square	F-Value	Pr $>$ F
Machine	A	87.75	D	E	F
Error	B	C	2.20		
Total	19	122.95			

(c) Is there significant variability due to Machine Differences ($\alpha = 0.05$)? To get full credit, give the null and alternative hypotheses, the value of the test statistic, the sampling distribution under the null hypothesis, the (range of) p-value, and your conclusion.

(df_1, df_2)	$(3, 16)$	$(4, 16)$	$(3, 19)$	$(4, 19)$
$F_{0.025;df_1,df_2}$	4.0768	3.7294	3.9034	3.5587
$F_{0.05;df_1,df_2}$	3.2389	3.0069	3.1274	2.8951

(d) Estimate all variance components.

(e) The overall goal of the company is to produce net with strength measurements mean of at least 123. Test if the overall mean strength measurements is *larger* than 123 ($\alpha = 0.05$).

$t_{0.025;3}$	$t_{0.025;16}$	$t_{0.025;19}$	$t_{0.05;3}$	$t_{0.05;16}$	$t_{0.05;19}$
3.1824	2.1199	2.0930	2.3534	1.7459	1.7291

8. An experiment is conducted to study the effects of loading frequencies (Frequency) and environmental conditions (Environment) on fatigue crack growth at a constant 22 MPa stress for a particular material. The data from this experiment are shown below (the response is crack growth rate):

	Environment		
Frequency	1: Air	2: H_2O	3: SaltH_2O
1	2.29, 2.47, 2.12	2.86, 3.03, 2.73	4.93, 4.75, 5.06
2	3.15, 2.88, 2.56	4.00, 4.44, 4.70	3.10, 3.24, 3.98
3	2.24, 2.71, 2.81	4.00, 4.30, 3.20	4.86, 4.26, 5.20

Some summary statistics are give on the next page.

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 ($j = 1, 2, 3$) are the main effects of frequency, the main effects of environment, respectively, and $(\tau\beta)_{ij}$ are their interactions. For parameter estimation, we impose the following constraints: $\sum_i \tau_i = \sum_j \beta_j = \sum_i (\tau\beta)_{ij} = \sum_j (\tau\beta)_{ij} = 0$.

(a) Calculate the estimate of $(\tau\beta)_{23}$.

(b) Calculate the sum of squares due to frequency.

(c) The ANOVA table from SAS is shown on the next page in which some quantities are removed. Test if the interaction of frequency and environment is significant ($\alpha = 5\%$). To get full credit, give the null and alternative hypotheses, the value of the test statistic, the sampling distribution under the null hypothesis, the (range of) p-value, and your conclusion. (You don't need to complete the ANOVA table.)

grand mean: 3.551

frequency	MEAN	environment	MEAN
1	3.360	1	2.581
2	3.561	2	3.696
3	3.731	3	4.376

frequency	environment	MEAN
1	1	2.293
1	2	2.873
1	3	4.913
2	1	2.863
2	2	4.380
2	3	3.440
3	1	2.587
3	2	3.833
3	3	4.773

Dependent Variable: crack

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	22.719	2.840	22.11	<.0001
Error	18	2.312	0.128		
Coed Total	26	25.031			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
frequency	*	*****	*****	****	*****
environment	2	14.773	7.387	57.51	<.0001
frequency*environment	*	*****	*****	****	*****

(df_1, df_2)	$(2, 18)$	$(4, 18)$	$(8, 18)$	$(2, 26)$	(4,26)	(8,26)
$F_{0.025;df_1,df_2}$	4.5597	3.6083	3.0053	4.2655	3.3289	2.7293
$F_{0.05;df_1,df_2}$	3.5546	2.9277	2.5102	3.3690	2.7426	2.3205

(d) Calculate the critical difference (CD) for comparing all the treatment pairwisely with Tukey's method ($\alpha = 0.05$).

(df_1, df_2)	$(8, 18)$	$(9, 18)$	$(8, 26)$	$(9, 26)$
$q_{0.025;df_1,df_2}$	5.3146	5.4476	5.0838	5.2049
$q_{0.05;df_1,df_2}$	4.8243	4.9552	4.6519	4.7733