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 the random variable Y has the probability mass function (pmf),

$$f_Y(y) = p(1-p)^y, y = 0, 1, 2, \dots,$$

where $0 < p < 1$.

(a) Identify the name of the distribution that Y follows, and explain what p and y denote.

(b) Find the moment generating function (mgf) of Y (for t near zero).

(c) Find the mean and the variance of Y using the mgf.

2. Let X be a random variable with finite expectation on the support $S_x \subset (a, b)$.

(a) Suppose that g is a convex function on the open interval (a, b). Prove the Jensen's inequality, i.e., show that

$$g[E(X)] \leq E[g(X)].$$

(b) Let X be a positive random variable with finite expectation. Determine which is greater between $1/E(X)$ and $E(1/X)$.

(c) Check your answer in (b) with the random variable $X \sim \Gamma(\alpha, \beta)$ with $E(X) = \alpha\beta$, i.e.,

$$f_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}} \text{ for } x > 0.$$

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

$$f(x; \theta) = \theta^{-2} x \exp\Big(-\frac{x}{\theta}\Big),$$

where $x > 0$ and $\theta > 0$.

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

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

(c) Find the MLE for $Var(X_i)$, $i = 1, \dots, n$.

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

(e) Find the limiting distribution of $\sqrt{n}(\hat{\theta}^2 - \theta^2)$.

4. Let $X_1, X_2, \dots, X_n$ be identically and independently distributed random variables with the pdf

$$f(x; \theta) = \theta x^{\theta - 1},$$

where $0 < x < 1$ and $\theta > 0$. It is known that the MLE of θ is $-\frac{n}{\sum_{i=1}^{n} \log X_i}$.

(a) Show that the likelihood ratio Λ for testing $H_0 : \theta = 1$ vs $H_1 : \theta \neq 1$ depends only on $W = -2 \sum_{i=1}^{n} \log X_i$.

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

(c) Find the rejection region for a test of size α.

(d) Specify the sampling distribution of W under the general alternative hypothesis with $\theta \neq 1$.

(e) Derive the power function of the test.

5. A study collected data on the starting salaries of 25 college teachers. The average salaries for the 10 female college teachers is $\bar{Y}_1$ =3.31 (10 thousands of dollars) with the sample standard deviation as $S_1 = 0.50$ (10 thousands of dollars); The average salaries for the 15 male college teachers is $\bar{Y}_2$ =3.93 (10 thousands of dollars) with the sample standard deviation as $S_2 = 0.45$ (10 thousands of dollars). The sample mean and the sample standard deviation are defined as $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$ and $S_i = \sqrt{\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2/(n_i - 1)}$, respectively, where $i = 1$ for the female teachers group, $i = 2$ for the male teachers group, $n_1 = 10$ and $n_2 = 15$.

Assume the population variances in the female and male teachers are equal, that is, $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Assume normality in the salaries and independence within and between the female and male college teacher groups.

Below is a list of selected percentage points for the standard normal distribution with Z_α implying that $P(Z > Z_\alpha) = \alpha$ and the t-distribution with v degrees of freedom, where $t_{v,\alpha}$ implies $P(T_v > t_{v,\alpha}) = \alpha$. You may need to use some of them when answering parts (a) through (c).

$$\begin{array}{lll}
Z_{0.10} = 1.2816; & Z_{0.05} = 1.6449; & Z_{0.025} = 1.9600 \\
t_{25,\,0.10} = 1.3163; & t_{25,\,0.05} = 1.7082; & t_{25,\,0.025} = 2.0595 \\
t_{24,\,0.10} = 1.3178; & t_{24,\,0.05} = 1.7109; & t_{24,\,0.025} = 2.0639 \\
t_{23,\,0.10} = 1.3195; & t_{23,\,0.05} = 1.7139; & t_{23,\,0.025} = 2.0687 \\
t_{22,\,0.10} = 1.3212; & t_{22,\,0.05} = 1.7171; & t_{22,\,0.025} = 2.0739 \\
t_{21,\,0.10} = 1.3232; & t_{21,\,0.05} = 1.7207; & t_{21,\,0.025} = 2.0800
\end{array}$$

(a) Carry out a two-sample t-test to examine if there is evidence to support the claim that the starting salaries for female college teachers are lower than that of the male college teachers. Please clearly state the null and alternative hypotheses, the test statistic and its observed value, the decision rule used and the context specific conclusion. Use $\alpha = 0.05$.

(b) Now we construct a simple linear regression model to investigate the relationship between gender and the starting salaries for college teachers. The model is specified as below:

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

where Y_i is the starting salaries of the i^{th} college teachers, $i = 1, \ldots, 25$. The covariate $X_i = 1$ for female college teachers and $X_i = 0$ for male college teachers. The random error terms ϵ_i's are identically and independently distributed as a normal distribution with mean 0 and variance τ^2, that is, $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \tau^2)$.

Find the least squares estimate $\hat{\beta}_1$ of the slope coefficient β_1. Please clearly show all your steps in the derivation.

(c) If it is known that $\hat{\beta}_1 = -0.62$ and the standard error for $\hat{\beta}_1$ is 0.19, carry out the hypothesis testing for $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 < 0$. Please clearly state the construction of the test statistic and its observed value, the decision rule used, and the context specific conclusion. Use $\alpha = 0.05$.

6. Measurements on nine infants were taken with the goal to arrive at a suitable estimating equation relating the length of an infant (Y, in centimeters (cm)) to all or a subset of independent variables, including age in days (X_1), length at birth in cm (X_2), weight at birth in kilograms (X_3), and chest size at birth in cm (X_4). The normal linear regression model is used. Results on a collection of models are included below, where MSE is the error mean square and R^2 represents the coefficient of determination.

Below is a list of percentage points for the F-distribution with v_1 and v_2 degrees of freedom, where $F_{v_1,v_2,\alpha}$ implies $P(F_{v_1,v_2} > F_{v_1,v_2,\alpha}) = \alpha$. You may need to use some of them when answering parts (a) through (c).

$$\begin{array}{lll}
F_{2,3,0.05} = 9.55; & F_{2,4,0.05} = 6.94; & F_{2,5,0.05} = 5.78 \\
F_{2,6,0.05} = 5.14; & F_{2,7,0.05} = 4.73; & F_{2,8,0.05} = 4.46
\end{array}$$

Model index	MSE	R^2	variables in the model
1	4.728	0.897	X_1
2	15.134	0.670	X_2
3	19.305	0.579	X_3
4	31.483	0.314	X_4
5	1.537	0.971	X_1 X_2
6	0.631	0.988	X_1 X_3
7	3.213	0.940	X_1 X_4
8	0.506	0.991	X_2 X_3
9	7.347	0.863	X_2 X_4
10	22.334	0.583	X_3 X_4
11	0.606	0.991	X_2 X_3 X_4
12	0.720	0.989	X_1 X_3 X_4
13	1.844	0.971	X_1 X_2 X_4
14	0.598	0.991	X_1 X_2 X_3
15	0.741	0.991	X_1 X_2 X_3 X_4

(a) For the full model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$, test the following hypotheses. Use $\alpha = 0.05$.

$$H_0: \beta_1 = \beta_3 = 0$$
$$H_1: \text{ not both } \beta_1 \text{ and } \beta_3 \text{ are equal to 0.}$$

(b) Assume the model $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \epsilon$ is fitted. Compute the coefficient of partial determination between Y and X_3 given that X_1 is in the model, i.e., find $R^2_{Y3|1}$. Interpret the obtained value.

(c) The Akaike Information Criterion for a regression model with p regression coefficients (AIC_p) is defined as $AIC_p = n \ln SSE_p - n \ln n + 2p$, where SSE_p represents the error sum of squares of the corresponding regression model. Select the best model using AIC_p.

7. An experiment was conducted to assess the relative resistance to abrasion of four grades of leather (A,B,C,D). A machine was used in which the samples of leather could be tested in any one of four machine positions. Since different runs (replications) are known to yield variable results, it was decided to make four runs. There may be variation in readings depending on which position in the machine and the particular run of the machine. Thus, we have two blocking variables: Machine Position and Run. The following design is used for the experiment and the experiment outcome is also included.

	Position			
	1	2	3	4
1	C=31	D=43	A=67	B=36
Run 2	D=39	A=96	B=40	C=48
3	B=57	C=33	D=40	A=84
4	A=85	B=46	C=48	D=50

The grand mean $\bar{Y}_{\dots} = 52.6875$, and the level means for the four methods are

```
A: 83  B: 44.75 C: 40 D: 43
```

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

b. Please fill in the missing values indicated below by "?".

Source	DF	SS
Treatment	?	?
Run	?	408.1875
Position	?	88.6875
Error	?	515.875
Total	?	5959.4375

c. Perform the Bonferroni pairwise treatment comparison test. Please report the critical difference (CD) *and comparison results*.

$t_{6,\,0.05} = 1.943;$ $t_{6,\,0.025} = 2.447;$ $t_{6,\,0.05/6} = 3.287;$ $t_{6,\,0.05/12} = 3.863$
$t_{12,\,0.05} = 1.782;$ $t_{12,\,0.025} = 2.178;$ $t_{12,\,0.05/6} = 2.779;$ $t_{12,\,0.05/12} = 3.153$

8. A chemical production process consists of a first reaction with an alcohol and a second reaction with a base. A factorial experiment with three alcohols and two bases was conducted with three replicate reactions conducted in a completely randomized design. The collected data were percent yield.

Base	Alcohol		
	1	2	3
1	91, 90, 91	89, 88, 90	87, 88, 90
2	87, 88, 91	91, 92, 95	90, 92, 93

Here are some summary statistics:

```
Grand mean: 90.1667
----------------------
Base mean
1 89.3333
2 91.0000
-----------------------------
Alcohol  mean
1 89.6667
2 90.8333
3 90.1667
------------------------------------------
Base alcohol mean
1 1 90.6667
1 2 89.0000
1 3 88.3333
2 1 88.6667
2 2 92.6667
2 3 91.6667
```

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

a. Estimate the main effect τ_1 (base=1) and interaction effect $(\tau\beta)_{23}$ (base=2, alcohol=3).

b. Complete the following ANOVA table from SAS by filling in the missing values indicated by "??".

```
Source      DF          Squares     Mean Square    F Value    Pr > F
Model       ??      47.16666667      9.43333333       3.86    0.0257
Error       ??           ??          2.44444444
Total       ??      76.50000000

Source            DF  Type I SS     Mean Square    F Value    Pr > F
base              ??     ??            ??
alcohol           ??  4.33333333       ??
base⋆alcohol      ??     ??            ??
```

c. Do the effects of base on the production process *dependent* on alcohol 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 (or range of p-value), and give your conclusion using $\alpha = 0.05$.

(df_1, df_2)	$(2, 12)$	$(2, 13)$	$(2, 14)$	$(3, 12)$	(3,13)	(3,14)
$F_{df_1,df_2;0.025}$	5.0959	4.9653	4.8567	4.4742	4.3472	4.2417
$F_{df_1,df_2;0.05}$	3.8853	3.8056	3.7389	3.4903	3.4105	3.3439

d. Use Tukey's method to perform a pairwise comparison for *all treatments* with $\alpha = 0.05$. Report the critical difference (CD). *Note: You don't need to report comparison results.*

(df_1, df_2)	$(3, 12)$	$(3, 13)$	$(3, 14)$	$(6, 12)$	$(6, 13)$	$(6, 14)$
$q_{df_1,df_2;0.025}$	4.3243	4.2687	4.2220	5.3319	5.2481	5.1776
$q_{df_1,df_2;0.05}$	3.7729	3.7341	3.7014	4.7502	4.6897	4.6385