Statistics Qualifying Exam

12:00 pm - 4:00 pm, Tuesday, August 20^{th}, 2019

1. Suppose that random variables X_1 and X_2 have the following joint probability mass function (pmf).

		x_2			
		0	1	2	3
x_1	0	$\frac{1}{8}$	$\frac{1}{8}$	0	0
	1	0	$\frac{2}{8}$	$\frac{2}{8}$	0
	2	0	0	$\frac{1}{8}$	$\frac{1}{8}$

   (a) Calculate the expectation of X_1, i.e., $E(X_1)$.

   (b) Let $Y = g(X_1, X_2) = X_2 - X_1$. Calculate the expectation of Y, i.e., $E(g(X_1, X_2))$.

   (c) Calculate the probability $P(X_1 \geq X_2)$.

2. Suppose that X_1 and X_2 are independent and identically distributed with the exponential distribution whose marginal probability density function (pdf) is given by

$$f_{X_j}(x_j) = \begin{cases} \frac{1}{2}\exp\left(-\frac{x_j}{2}\right) & x_j > 0, \\ 0 & \text{elsewhere,} \end{cases}$$

   for $j = 1, 2$. Define $Y_1 = \frac{1}{2}(X_1 - X_2)$ and $Y_2 = X_2$.

   (a) Find the joint pdf of Y_1 and Y_2.

   (b) Show that Y_1 follows the Laplace distribution (or the double exponential distribution) with zero mean.

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

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

   Let $Y_1 < Y_2 < \ldots < Y_n$ be the order statistics. Let $W = Y_1/Y_n$ and $Q = Y_n$.

   (a) Find the joint pdf of W and Q. Based on the obtained joint distribution funcition of W and Q, disucss if W and Q are independent random variables.

(b) Basu's Theorem says that if $T(\mathbf{X})$ is a complete and (minimal) sufficient statistic for θ, then $T(\mathbf{X})$ is independent of every ancillary statistic of θ. Based on Basu's Theorem, discuss if W and Q are independent random variables.

4. Let $X_1, \ldots, X_n$ be identically and idenpendently distriubted with the $N(0, \theta)$ pdf given as

$$f(x; \theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{x^2}{2\theta}\right), \quad \theta > 0, -\infty < x < +\infty.$$

(a) Derive the maximum likelihood estimator (mle) $\hat{\theta}$ for θ. Show its expectation.

(b) Show that the mle $\hat{\theta}$ obtained in (a) is also the minimum variance unbiased estimator (MVUE) for θ.

(c) We wish to test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Answer the following questions:

i. Show that the likelihood ratio test depends only on $S = \sum_{i=1}^{n} X_i^2/\theta_0$. That is, S is the test statistic used here.

ii. What is the sampling distribution of S under H_0?

iii. Explicitly state the decision rule of a size α test in the form "reject H_0 if $S \leq c_1$ or $S \geq c_2$" or "reject H_0 if $S \geq c_3$" with the constants c_1 and c_2 or c_3 clearly specified.

5. 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.

df	41	42	43	44	45
$t_{0.05;df}$	1.6829	1.6820	1.6811	1.6802	1.6794
$t_{0.025;df}$	2.0195	2.0181	2.0167	2.0154	2.0141

   (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	.	.	.

6. The following is part of 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. (No need to complete the 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 now know the least square estimate of β_1 is $b_1 = 3.57$. Construct a two-sided t-test of whether or not $\beta_1 = 3$. State the null and alternative hypotheses, the value of the test statistic, the sampling distribution under the null hypothesis and the decision rule. *[Note: You don't need to calculate the p-value or make the conclusion.]*

7. An experiment was conducted to determine the effects of four different pesticides on the yield of fruit from three different varieties of a citrus tree. Eight trees from each variety were available and the the four pesticides were then randomly assigned to *two* trees of each variety. Yields of fruit (in bushels per tree) were obtained after the test period. The mean yields for each combination of pesticide and variety are given below.

	Variety			
Pesticide	1	2	3	Pesticide Means
1	44	48	67	53.00
2	52.5	62.5	88.5	67.83
3	40.5	47.5	65.5	51.17
4	50.5	79	92	73.83
Variety Means	46.875	59.25	78.25	

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

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

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

Some SAS output is shown.

```
The GLM Procedure
Dependent Variable: yield

Source                  Sum of Squares
Model                      6680.458333
Error                       507.500000
Corrected Total            7187.958333

Source                    Type I SS
pesticide                2227.458333
variety                  3996.083333
pesticide*variety         456.916667
```

(a) What are the estimates of τ_3 and $(\tau\beta)_{23}$?

(b) Provide the degrees of freedom corresponding to each of the sums of squares in the output, which are marked by "???" below.

Source	Degrees of freedom
Model	???
Error	???
Corrected Total	???
pesticide	???
variety	???
pesticide * variety	???

(c) Do the effects of the pesticides on yield *dependent* on the variety of citrus tree? Conduct an appropriate test to answer this question. To get full credits, give hypotheses, a test statistic, determine its degrees of freedom, use an appropriate value from the table below, state the p-value (or its range) and give a conclusion using $\alpha = 0.05$.

(df_1, df_2)	$(5, 10)$	$(5, 11)$	$(5, 12)$	$(5, 13)$	(5,14)
$F_{0.05;df_1,df_2}$	3.3258	3.2039	3.1059	3.0254	2.9582
(df_1, df_2)	$(6, 10)$	$(6, 11)$	$(6, 12)$	$(6, 13)$	(6,14))
$F_{0.05;df_1,df_2}$	3.2172	3.0946	2.9961	2.9153	2.8477

(d) Use Tukey's method to perform a pairwise comparison for different *varieties*. Report the critical difference **and** report your results of comparison (using $\alpha = 0.05$). You can report the result as we have done in class by labeling significantly different combinations with different Latin letters.

(df_1, df_2)	$(2, 10)$	$(2, 11)$	$(2, 12)$	$(2, 13)$	$(2, 14)$
$q_{0.025;df_1,df_2}$	3.7247	3.6672	3.6204	3.5817	3.5491
$q_{0.05;df_1,df_2}$	3.1511	3.1127	3.0813	3.0552	3.0332
(df_1, df_2)	$(3, 10)$	$(3, 11)$	$(3, 12)$	$(3, 13)$	$(3, 14)$
$q_{0.025;df_1,df_2}$	4.4740	4.3913	4.3243	4.2687	4.2220
$q_{0.05;df_1,df_2}$	3.8768	3.8196	3.7729	3.7341	3.7014

(e) Suppose Pesticides 1 and 2 are sold by Company A and Pesticides 3 and 4 are sold by Company B. Conduct a test or construct a confidence interval that can be used to compare the effectively of Company A's pesticides to the effectiveness of Company B's pesticides. Show your work and provide a conclusion.

(df_1, df_2)	$(1, 10)$	$(1, 11)$	$(1, 12)$	$(1, 13)$	$(1, 14)$
$F_{0.05;df_1,df_2}$	4.9646	4.8443	4.7472	4.6672	4.6001
df	10	11	12	13	14
$t_{0.05;df}$	1.8125	1.7959	1.7823	1.7709	1.7613
$t_{0.025;df}$	2.2281	2.2010	2.1788	2.1604	2.1448

8. You have been asked to design an experiment to compare the effect of two types of plant food on tomato plant growth using sixteen plants. These sixteen plants will be grown on a greenhouse bench in two rows of eight as shown below (2 rows of 8 plants). Even though this is in a greenhouse, you are concerned about air temperature (a nuisance factor) affecting plant growth. For each item below,

   - describe **how you would allocate the treatments to these sixteen tomato plants**;
   - give the **model**, and write out the ANOVA table in terms of **sources** and **degrees of freedom**.

   (a) You are told that there is no temperature gradient.

   (b) You are told that there is a considerable temperature gradient running against the rows (↓).

   (c) You are told there is a moderate temperature gradient that runs along the bench (←).
