Statistics Qualifying Exam

9:00 am - 1:00 pm, Monday, August 14, 2017

1. Consider two random variables X_1 and X_2 whose joint probability density function (pdf) is given by the following :

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

   (a) Find the conditional probability density function of X_2 given that $X_1 = x_1$, that is, $f_{X_2|X_1}(x_2|x_1)$. Derive the conditional expectation of X_2 given that $X_1 = x_1$, i.e., $E(X_2|X_1 = x_1)$.

   (b) Suppose two new random variables are defined as $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$. Find the joint pdf of (Y_1, Y_2), that is, $f_{Y_1,Y_2}(y_1, y_2)$.

   (c) Find the marginal pdfs of Y_1 and Y_2, that is, $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$, respectively.

2. Let the random variable X follow a normal distribution with a mean 0 and a variance θ, that is, $N(0, \theta), 0 < \theta < \infty$. And let $X_1, \dots, X_n$ be a random sample of X with a size n.

   (a) Find the maximum likelihood estimator (mle) of θ, $\hat{\theta}_{mle}$. What is the asymptotic distribution of $\sqrt{n}(\hat{\theta}_{mle} - \theta)$?

   (b) Derive an unbiased estimator of $\sqrt{\theta}$ (hint: find $E(|X|)$ first).

   (c) Determine the efficiency of the unbiased estimator derived in part (b).

3. Let $X_1, \dots, X_n$ denote a random sample of size $n > 2$ from a distribution with the pdf as

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

   where $0 < \theta < \infty$. Assume that it is of interest to test the following hypotheses

$$H_0 : \theta = 1 \quad \text{versus} \quad H_1 : \theta \neq 1.$$

   (a) Find an exact size α test for the above hypotheses.

   (b) The likelihood ratio is defined as $\Lambda = L(H_0)/L(H_1)$, where $L(H_*)$ represents the likelihood under H_*. Explicitly write out Λ for the above hypotheses. What is the asymptotic distribution of $-2 \log \Lambda$? Clearly specify the decision rule in the likelihood ratio test.

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

$$f(x; \theta) = \begin{cases} 1/(3\theta), & -\theta < x < 2\theta \\ 0, & \text{elsewhere,} \end{cases}$$

where $0 < \theta < \infty$.

(a) Find the mle $\hat{\theta}_{mle}$ of θ.

(b) Is $\hat{\theta}_{mle}$ a sufficient statistic for θ? Justify your answer.

(c) Find the unique MVUE of θ.

5. A hospital administrator wished to study the relation between patient satisfaction (Y) and patient's age (X_1, in years), severity of illness (X_2, an index), and anxiety level (X_3 , an index). The administrator randomly selected 36 patients and collected the data, where larger values of Y, X_2, and X_3 are, respectively, associated with more satisfaction, increased severity of illness, and more anxiety. Below is part of SAS output for all possible linear regression models fitted for this data. Based on the information, answer the following questions.

Number in Model	R-Square	MSE	SSE	Variables in Model
1	0.5656	0.04348	1.47843	x3
1	0.4045	0.05961	2.02660	x1
1	0.3354	0.06652	2.26177	x2
2	0.6145	0.03976	1.31195	x1 x3
2	0.5917	0.04211	1.38954	x2 x3
2	0.4219	0.05962	1.96735	x1 x2
3	0.6150	0.04095	1.31025	x1 x2 x3

(a) Use the adjusted R-square to compare the model $Y = \beta_0 + \beta_2 X_2 + \epsilon$ and the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$.

(b) For the full model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$, test the following using $\alpha = 0.05$.

$$H_0: \ \beta_1 = \beta_3 = 0, \text{ vs } \quad H_a: \text{ not both } \beta_1 \text{ and } \beta_3 \text{ are } 0.$$

(c) Use the stepwise selection method to select the "best" regression model (use $\alpha = 0.05$).

6. In an experiment, the amount of radon released in shower was investigated. Radon-enriched water was used, and five different orifice diameters were tested in shower heads. The data from the experiment are shown in the following table (the response is the percentage of radon released).

Orifice Diameter					Mean	Std.D
0.40	87	88	89	93	89.25	2.62995564
0.60	74	73	76	77	75.00	1.82574186
0.80	69	71	70	72	70.50	1.29099445
1.00	76	72	74	74	74.00	1.63299316
1.20	89	92	84	89	88.50	3.31662479

The ANOVA table obtained from SAS is given below.

```
                              Analysis of Variance

                                     Sum of              Mean
Source                   DF         Squares             Square        F Value

Model                    4         1230.200            307.550000         60.11
Error                    15        76.750              5.116667
Corrected Total          19        1306.950
```

(a) Give the model, and test if the diameter of orifice affects the mean percentage of radon released (use $\alpha = 0.05$). To get full credits, give the hypotheses, test statistic, the p-value and conclusion.

(b) Perform pairwise comparison using Bonferroni and Tukey's methods, respectively (use $\alpha = 0.05$). Summarize and compare the results, and comment on which method is preferred.

(c) Notice that the amount of released radon changes when the size of orifice varies from 0.40 to 1.20 in diameter. An analyst wants to study the functional relationship between the response and the diameter. She obtains the complete set of orthogonal contrasts from Table IX in Montgomery:

$C1$:	-2	-1	0	1	2
$C2$:	2	-1	-2	-1	2
$C3$:	-1	2	0	-2	1
$C4$:	1	-4	6	-4	1

The contrast sum of squares for $C1$, $C3$, and $C4$ and their testing results are given below,

Contrast	DF	Contrast SS	Mean Square	F Value	Pr>F
C1	1	2.500000	2.500000	0.49	0.4952
C2	*	********	********	****	******
C3	1	0.625000	0.625000	0.12	0.7316
C4	1	1.289286	1.289286	0.25	0.6230

Obtain the estimate of $C1$ and test if there exists a significant linear relationship (use $\alpha = 0.05$).

(d) Note that the contrast SS, Mean square, F Value and Pr $>$ F for $C2$ are missing. Recover these values and test if $C2$ is significant (Use $\alpha = 0.05$).

7. An engineer is studying the mileage performance characteristics of 5 types of gasoline additives. In the road test he wishes to use cars as blocks; however, because of a time constraint, he must use an incomplete block design. He runs the balanced design with the five blocks that follow.

	car				
additive	1	2	3	4	5
1		17	14	13	12
2	14	14		13	10
3	12		13	12	9
4	13	11	11	12	
5	11	12	10		8

(a) Verify that this is a balanced incomplete block design and give the parameters of this design.

(b) The ANOVA table from SAS is as follows (here the SS is Type I SS of the block and the treatment). Test if there is a difference between the five additives using $\alpha = 5\%$. To get full credits, please give the hypotheses, test statistic, p-value and your conclusion.

```
                          Sum of
Source             DF     Squares         Mean Square   F Value  Pr > F
car                4      31.20000000     7.80000000    8.57     0.0022
trt                4      35.73333333     8.93333333    9.81     0.0012
Error              11     10.01666667     0.91060606
Corrected Total    19     76.95000000
```

(c) The overall mean is $\bar{y}_{..} = 12.05$. Calculate the estimates of treatment means (i.e., the least square means).

(d) Calculate the standard error of the difference between two treatment mean estimates (i.e. the standard error of $\tau_i - \tau_j$). Also, calculate the critical difference for Tukey's pairwise comparisons and draw the conclusions.

(e) Suppose the engineer wants to know whether the combination of additives 1 and 2 has the same characteristics as the combination of additives 4 and 5. Use a proper contrast to address this issue and offer your answer.

8. Two students at Napier University in Scotland performed an experiment to investigate the effects of several factors on the distance (in yards) a golf ball travels. The factors and data are given below. The experiment was run as a completely randomized design with one observation at each treatment combination.

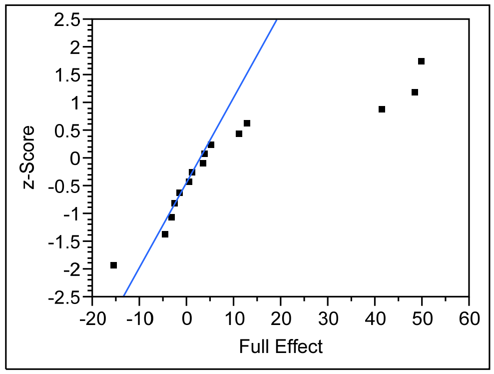
Factor	-1	$+1$
A: Ability (handicap)	Lower (8)	Higher (4)
B: Teeing	no tee	tee
C: Club	wood	metal
D: Ground	soft	hard

A Ability	B Teeing	C Club	D Ground	Y Distance
-1	-1	-1	-1	153
$+1$	-1	-1	-1	195
-1	$+1$	-1	-1	211
$+1$	$+1$	-1	-1	232
-1	-1	$+1$	-1	160
$+1$	-1	$+1$	-1	204
-1	$+1$	$+1$	-1	222
$+1$	$+1$	$+1$	-1	236
-1	-1	-1	$+1$	183
$+1$	-1	-1	$+1$	260
-1	$+1$	-1	$+1$	242
$+1$	$+1$	-1	$+1$	285
-1	-1	$+1$	$+1$	200
$+1$	-1	$+1$	$+1$	264
-1	$+1$	$+1$	$+1$	276
$+1$	$+1$	$+1$	$+1$	301

(a) Please calculate the effect of interaction AB, and its corresponding Sum of Square.

(b) Below is a table of estimated effects (except AB). The QQ plot of the estimated effects is also given. Please indicate the potentially important effects. Are there effects that can be dropped? If so, which ones? Explain briefly.

Name	Full Effect	Name	Full Effect	Name	Full Effect	Name	Full Effect
Mean	226.50	C	12.75	D	49.75	CD	5.00
A	41.25	AC	-4.50	AD	11.00	ACD	-3.25
B	48.25	BC	3.50	BD	1.00	BCD	3.75
AB		ABC	-1.75	ABD	-2.75	ABCD	0.50



(c) The students include all main effects and two-way interactions in the model statement in SAS. Below is the partial ANOVA table. According to this output, what effects and interactions are statistically significant? Also complete the degrees of freedom for SSM, $SSTO$ and SSE, indicated by "???". (*No need to fill in "******"*).

Source	DF	Sum of Squares
Model	???	
Error	???	
Corrected Total	???	

Source	DF	Sum of Squares	F Ratio	Prob > F
A	1	6806.2500	239.6567	$< .0001$
B	1	9312.2500	327.8961	$< .0001$
A*B	*****	*****	*****	0.0021
C	1	650.2500	22.8961	0.0049
A*C	1	81.0000	2.8521	0.1521
B*C	1	49.0000	1.7254	0.2460
D	1	9900.2500	348.6004	$< .0001$
A*D	1	484.0000	17.0423	0.0091
B*D	1	4.0000	0.1408	0.7228
C*D	1	100.0000	3.5211	0.1194

(d) Using only those factors and interactions that are statistically significant in part (c), (i) give the prediction equation for distance and (ii) use this equation to predict the distance for a higher ability golfer using a metal club and a tee on hard ground.