Statistics Qualifying Exam

12:00 pm - 4:00 pm, Tuesday, May 7^{th}, 2019

1. Suppose that a random variable Z follows the "standard" normal distribution and X is defined as $X = \mu + \sigma Z$ where μ and σ are unknown, but fixed constants.

   (a) Prove that the moment generating function (m.g.f.) of Z is $M_Z(t) = e^{t^2/2}$ where t is a real value.

   (b) Using the m.g.f. of Z in part (a), show that the m.g.f. of X is $M_X(t) = e^{\mu t + \sigma^2 t^2/2}$.

   (c) Find the mean and variance of X using the m.g.f. of X in part (b).

2. Let X be a random variable with the probability density function,

$$f_X(x) = \begin{cases} \theta x^{\theta-1}, & \text{if } 0 < x < 1; \\ 0, & \text{otherwise,} \end{cases}$$

   where $0 < \theta < \infty$.

   (a) Show that $f_X(x)$ is a "legitimate" probability density function.

   (b) Let $Y = -\log(X)$. Find the m.g.f. of Y, $M_Y(t)$.

   (c) Show that Y follows the exponential distribution with mean $1/\theta$, i.e., $M_Y(t)$ is the m.g.f. of the exponential distribution.

3. Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with the uniform distribution $U(0, \theta)$. That is the probability density function is

$$f_X(x) = \begin{cases} \frac{1}{\theta}, & \text{if } 0 < x < \theta; \\ 0, & \text{otherwise,} \end{cases}$$

   where $\theta > 0$. Let $Y_1 < Y_2 < \ldots < Y_n$ be the order statistics of the random sample.

   (a) A statistic $T(\mathbf{X})$ is a consistent estimator of θ if it converges in probability to θ, that is $T(\mathbf{X}) \xrightarrow{\mathcal{P}} \theta$, where $\mathbf{X}$ represents the random sample of size n. By this definition, show that the maximum order statistic Y_n is a consistent estimator for θ.

   (b) Derive the likelihood ratio Λ for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

   (c) When H_0 is true, show that $-2\log\Lambda$ has an exact $\chi^2(2)$ distribution (a chi-square distribution with 2 degrees of freedom).

4. Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with the mean parameter $\theta > 0$.

(a) It is known that the maximum likelihood estimator (mle) for θ is $\bar{X} = \sum_{i=1}^{n} X_i/n$. Determine the asymptotic distribution of the mle of θ.

(b) Obtain the mle of $\tau(\theta) = P(X \leq 1) = (1+\theta)e^{-\theta}$. Determine its asymptotic distribution.

(c) Find the unique minimum variance unbiased estimator (MVUE) of $\tau(\theta)$ as defined in Part (b). Clearly justify each of your steps.

5. The following data reflect information from 17 U.S. Naval hospitals at various sites. The regressors are workload variables, that is, items that result in the need for personnel in a hospital. A brief description of the variables is as follows.

Y=monthly labor-hours/1000

X_1=average daily patient load/100

X_2=monthly X-ray exposure/1000

X_3=monthly occupied bed-days/1000

X_4=eligible population in the area/1000

X_5=average length of patient's stay, in days

The goal is to produce an appropriate model that will estimate (or predict) personnel needs for Naval hospitals. Normal linear regression models are fitted to the data.

(a) Based on the SAS output in Figure 1 select the "best" model using the stepwise method. Use $\alpha = 0.05$.

(b) Use F test to compare the following two models (use $\alpha = 0.05$):

$$Y = \beta_0 + \beta_2 X_2 + \epsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

Here are the selected percentiles of F distribution, where $F_{\alpha;df_1,df_2}$ represents the $100(1-\alpha)$-percentile of the F distribution with the numerator degrees of freedom (df) as df_1 and the denominator df as df_2 and $Prob(F_{df_1,df_2} \geq F_{\alpha;df_1,df_2}) = \alpha$.

(df_1, df_2)	$(1,12)$	$(1,13)$	$(1,14)$	$(1,15)$	(1,16)
$F_{0.05;df_1,df_2}$	4.75	4.67	4.60	4.54	4.49
(df_1, df_2)	$(2,12)$	$(2,13)$	$(2,14)$	$(2,15)$	(2,16)
$F_{0.05;df_1,df_2}$	3.89	3.81	3.74	3.68	3.63

Number in Model	R-Square	Adjusted R-Square	SSE	Variables in Model
1	0.9722	0.9703	13.76231	x3
1	0.9715	0.9696	14.09985	x1
1	0.8934	0.8862	52.76006	x2
1	0.8843	0.8766	57.25318	x4
1	0.3348	0.2904	329.10538	x5
2	0.9867	0.9848	6.57238	x2 x3
2	0.9861	0.9841	6.86819	x1 x2
2	0.9848	0.9826	7.54209	x3 x5
2	0.9840	0.9817	7.90007	x1 x5
2	0.9754	0.9718	12.19065	x3 x4
2	0.9741	0.9704	12.79853	x1 x4
2	0.9725	0.9686	13.59958	x1 x3
2	0.9306	0.9207	34.33937	x2 x4
2	0.9239	0.9130	37.63959	x2 x5
2	0.9104	0.8976	44.31986	x4 x5
3	0.9901	0.9878	4.91340	x2 x3 x5
3	0.9894	0.9870	5.24179	x1 x2 x5
3	0.9873	0.9844	6.26972	x1 x2 x3
3	0.9868	0.9837	6.55484	x2 x3 x4
3	0.9861	0.9829	6.86734	x1 x2 x4
3	0.9850	0.9816	7.40779	x1 x3 x5
3	0.9850	0.9815	7.42033	x3 x4 x5
3	0.9847	0.9811	7.58999	x1 x4 x5
3	0.9785	0.9735	10.63777	x1 x3 x4
3	0.9523	0.9412	23.61614	x2 x4 x5
4	0.9908	0.9877	4.54592	x2 x3 x4 x5
4	0.9906	0.9875	4.64401	x1 x2 x4 x5
4	0.9905	0.9874	4.67752	x1 x2 x3 x5
4	0.9879	0.9838	5.99362	x1 x2 x3 x4
4	0.9851	0.9801	7.38889	x1 x3 x4 x5
5	0.9908	0.9867	4.53505	x1 x2 x3 x4 x5

Figure 1: SAS output for Problem 5.

6. The effect of five different ingredients (A, B, C, D, E) on the reaction time of a chemical process is being studied. Each batch of new material is only large enough to permit five runs to be made. Furthermore, each run requires approximately 1.5 hours, so only five runs can be made in one day. The experiment layout and results are given below.

	Day				
Batch	1	2	3	4	5
1	A=6	B=6	D=1	C=8	E=5
2	C=10	E=2	A=8	D=5	B=11
3	B=2	A=8	C=10	E=2	D=7
4	D=1	C=4	E=3	B=4	A=8
5	E=2	D=1	B=3	A=9	C=10

The grand mean $\bar{Y}_{...} = 5.44$. And the level means for batch, day and ingredient are given in Figure 2.

Level of batch	N	Mean	Level of day	N	Mean	Level of Ingredient	N	Mean
1	5	5.20	1	5	4.20	1	5	7.80
2	5	7.20	2	5	4.20	2	5	5.20
3	5	5.80	3	5	5.00	3	5	8.40
4	5	4.00	4	5	5.60	4	5	3.00
5	5	5.00	5	5	8.20	5	5	2.80

Figure 2: The means of the data in Problem 6.

Here are the selected percentiles of F distribution, where $F_{\alpha;df_1,df_2}$ represents the $100(1-\alpha)$-percentile of the F distribution with the numerator degrees of freedom (df) as df_1 and the denominator df as df_2 and $Prob(F_{df_1,df_2} \geq F_{\alpha;df_1,df_2}) = \alpha$.

(df_1, df_2)	$(4, 12)$	$(4, 24)$	$(5, 12)$	$(5, 24)$
$F_{0.05;df_1,df_2}$	3.26	2.78	3.11	2.62
$F_{0.025;df_1,df_2}$	4.12	3.38	3.89	3.15

Here are the selected percentiles of t distribution, where $t_{\alpha;df}$ represents the $100(1-\alpha)$-percentile of the t distribution with df degrees of freedom (df) and $Prob(T_{df} \geq t_{\alpha;df}) = \alpha$.

df	4	5	12	24
$t_{0.1;df}$	1.53	1.48	1.36	1.32
$t_{0.05;df}$	2.13	2.02	1.78	1.71
$t_{0.025;df}$	2.78	2.57	2.18	2.06
$t_{0.01;df}$	3.75	3.36	2.68	2.49
$t_{0.005;df}$	4.60	4.03	3.06	2.80

(a) What design is employed for this experiment?

(b) Write down a proper statistical model to analyze this dataset, and state the assumptions.

(c) Part of the SAS output for ANOVA is given below in Figure 3. Test if the five ingredients have different effect on the chemical time. To get full credits, give hypotheses, the test statistic, p-value, and your conclusion (use $\alpha = 0.05$).

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	218.880	18.2400	5.57	0.0029
Error	12	39.280	3.2733		
Co. Total	24	258.160			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
batch	4	27.760	6.940	2.12	0.1410
day	4	54.560	13.640	4.17	0.0241
ingredient	*	******	******	****	******

Figure 3: Part of ANOVA for Problem 6.

(d) Use Bonferroni method for treatment pairwise comparison. Calculate the critical difference **AND** draw your conclusion (use $\alpha = 0.1$).

(e) Assume that five operators are employed to conduct the experiment, and it is know that the operator can influence the experimental results. Derive an experimental plan that can be used to study the five ingredients using five batches, five operators, and in five days. Some useful squares are given below in Figure 4.

A B C D E	A B C D E	A B C D E	A B C D E
B C D E A	C D E A B	E A B C D	D E A B C
C D E A B	E A B C D	D E A B C	B C D E A
E A B C D	D E A B C	B C D E A	C D E A B
D E A B C	B C D E A	C D E A B	E A B C D

Figure 4: Some squared that may be useful in Problem 6.

7. Heat treating is often used to carbonize metal parts, such as gears. The thickness of the carbonized layer is a critical output variable from this process, and it is usually measured by performing a carbon analysis on the gear pitch (the top of the gear tooth). Six factors are to be studied: A=furnace temperature, B=cycle time, C=carbon concentration, D=duration of the carbonizing cycle, E=carbon concentration of the diffuse cycle, and F=duration of the diffuse cycle. Suppose a 2^{6-2} fractional factorial design will be used for the experiment. There are several ways to construct a 2^{6-2} design. The general strategy is as follows.

   First, A, B, C and D form a 24 full factorial design (basic design) . Second, alias E and F with some high order effects of the basic design. Let d_1 denote the design generated by $E = AB$ and $F = CD$; and d_2 the design generated by $E = ABC$ and $F = BCD$.

   (a) Derive the complete defining relation for d_1. What is the resolution of d_1? What is its wordlength pattern?

   (b) Derive the complete defining relation for d_2. What is the resolution of d_2? What is its wordlength pattern?

   (c) Which design will you choose for the experiment? Why?

   A qualify improvement team has chosen one of the above two designs for the experiment. The design matrix and output are given below in Figure 5.

standard order	run order	A	B	C	D	E	F	pitch
1	5	-	-	-	-	-	-	74
2	7	+	-	-	-	+	-	190
3	8	-	+	-	-	+	+	133
4	2	+	+	-	-	-	+	127
5	10	-	-	+	-	+	+	115
6	12	+	-	+	-	-	+	101
7	16	-	+	+	-	-	-	54
8	1	+	+	+	-	+	-	144
9	6	-	-	-	+	-	+	121
10	9	+	-	-	+	+	+	188
11	14	-	+	-	+	+	-	135
12	13	+	+	-	+	-	-	170
13	11	-	-	+	+	+	-	126
14	3	+	-	+	+	-	-	175
15	15	-	+	+	+	-	+	126
16	4	+	+	+	+	+	+	193

Figure 5: Design and data for Problem 7.

   (d) Which design has been used by the team?

   (e) Estimate the main effect of Factor A.

   (f) Explain in a couple sentences about how to identify potentially important effects.

8. A gardener is interested in studying the relationship between fertilizer and tomato yield. The gardener has two gardens (1 and 2). He divides each into 9 plots. Three fertilizer application rates (3, 5, and 7 units/acre) are assigned to the plots in garden 1 in a completely randomized fashion. The same three fertilizer application rates (3, 5, and 7 units/acre) are assigned to the plots in garden 2 in a completely randomized fashion. Thus there are three plots for each combination of garden and fertilizer application rate. After some initial analyses, the gardener decides to base his analysis on the following SAS code and output in Figure 6.

```
proc glm;
  class garden;
  model yield=garden rate garden*rate / solution;
run;

The GLM Procedure
Dependent Variable: yield
                              Sum of
Source            DF         Squares     Mean Square    F Value    Pr > F
Model              3     58.88888889     19.62962963      27.33    <.0001
Error             14     10.05555556      0.71825397
Corrected Total   17     68.94444444

Source            DF       Type I SS     Mean Square    F Value    Pr > F
garden             1      2.72222222      2.72222222       3.79    0.0719
rate               1     52.08333333     52.08333333      72.51    <.0001
rate*garden        1      4.08333333      4.08333333       5.69    0.0318

                                 Standard
Parameter          Estimate         Error     t Value     Pr > |t|
Intercept         -1.11 B      0.90993803      -1.22      0.2422
garden      1      3.69 B      1.28684670       2.87      0.0123
garden      2      0.00 B       .                .         .
rate               1.33 B      0.17299494       7.71      <.0001
rate*garden 1     -0.58 B      0.24465179      -2.38      0.0318
rate*garden 2      0.00 B       .                .         .
```

Figure 6: SAS code and output for Problem 8

(a) Note that rate was not included in the class statement. What would the Model and Error DF change to if rate were included in the class statement? That is, complete the following tables by filling in the missing values for df.

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

Source	DF	Type I SS
garden	???	
rate	???	
rate*garden	???	

(b) Estimate the equation of the regression line relating yield to fertilizer application rate in garden 1.

(c) Estimate the equation of the regression line relating yield to fertilizer application rate in garden 2.

(d) Is there a significant difference between the slopes of the two regression lines? To get full credits, give an appropriate test statistic, p-value, and conclusion, using $\alpha = 0.05$.

(e) Suppose the gardener were to apply 7 units of fertilizer per acre to all plots in both gardens. Which garden would have the higher expected yield?