Statistics Qualifying Exam

12:00 pm - 4:00 pm, Monday, August 20^{th}, 2018

1. Suppose that $X_i, i = 1, \dots, m$, independently follows the binomial distribution $b(n_i, p)$ with the probability mass function,

$$f_{X_i}(x_i) = \binom{n_i}{x_i} p^{x_i}(1-p)^{n_i - x_i}, \quad x_i = 0, 1, \dots, n_i.$$

Note that $X_1, X_2, \dots, X_m$ are independent, but they are not identically distributed, i.e., the parameters n_i are different.

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

(b) Let $Y = \sum_{i=1}^m X_i$. Then, prove that Y follows the binomial distribution $b(t, p)$ where

$$t = \sum_{i=1}^m n_i.$$

2. Let X_1 and X_2 have the joint probability density function (pfd) as

$$f_{X_1,X_2}(x_1, x_2) = 2\exp(-x_1 - x_2), \quad 0 < x_1 < x_2 < \infty, \text{ zero elsewhere.}$$

(a) Find the marginal pdf of X_1.

(b) Let $Y_1 = 2X_1$ and $Y_2 = X_2 - X_1$. Find the joint pdf of the random vector (Y_1, Y_2).

(c) Find the conditional expectation of Y_1 given Y_2, that is, $\mathrm{E}(Y_1|Y_2)$.

3. Let $X_1, \dots, X_n$ be a random sample from a Gamma($\alpha = 3, \beta = \theta$) distribution, where θ is an unknown parameter with $0 < \theta < \infty$. The pdf of the Gamma($\alpha = 3, \beta = \theta$) distribution is specified as

$$f(x_i|\theta) = \theta^{-3} x_i^2 \exp\left(-\frac{x_i}{\theta}\right) / \Gamma(3), \quad 0 < x_i < \infty.$$

(a) Show that the **exact** Likelihood Ratio test of the hypotheses $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$ is based upon the statistic $W = \sum_{i=1}^n X_i$. Obtain the null distribution of $2W/\theta_0$.

(b) For $\theta_0 = 3$ and $n = 5$, find c_1 and c_2 so that the test that rejects H_0 has a significant level of 0.05.

(c) Obtain the power function of the test obtained in (a)-(b) with $\theta_0 = 3$ and $n = 5$.

4. Let $X_1, \dots, X_n$ be a random sample from the beta distribution Beta$(\theta, 1)$ with the pdf as

$$f(x_i|\theta) = \theta x_i^{(\theta-1)}, \quad 0 < x_i < 1, \ \theta > 0.$$

(a) Find the maximum likelihood estimator (MLE) of $1/\theta$. Is it unbiased?

(b) Calculate the information inequality lower bound of the MLE obtained in Part (a) and check whether it achieves the lower bound.

(c) Find a complete sufficient statistic for θ.

5. Consider the general linear regression model $\mathbf{y} = \mathbf{X}\beta + \epsilon$ where $\mathbf{y}$ ia a $n \times 1$ vector of response, $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})^T$ is a $p \times 1$ vector of parameters, $\mathbf{X}$ is a $n \times p$ full rank matrix of predictor variables and ϵ is a $n \times 1$ vector of independent normal random variables with its expectation $\mathbf{0}_{n\times 1}$ and variance-covariance matrix $\sigma^2\mathbf{I}_n$.

   (a) Derive the normal equation for the least square (LS) estimation of β and find the LS estimator, $\hat{\beta}$, of β.

   (b) Show that $E(\hat{\beta}) = \beta$ and $Var(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

   (c) Show that the residual, $\mathbf{e} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$ and its variance $\sigma^2(\mathbf{I}_n - \mathbf{H})$ where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

   (d) Show that, for $i \neq j$, e_i and e_j are NOT independent each other where $\mathbf{e} = (e_1, e_2, \ldots, e_n)^T$.

6. Consider the following ANOVA table for the multiple linear regression model

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

where the error term ϵ are independent and identical normal random variables with mean 0 and variance σ^2.

Source	SS	df	MS
Model	2176606	3	725535
X_1	136366	1	136366
$X_2 \mid X_1$	2033565	1	2033565
$X_3 \mid X_1, X_2$	6675	1	6675
Error	985530	48	20532
Total	3162136	51	

   (a) Find $SSR(X_2, X_3|X_1)$.

   (b) For the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$, test the following hypothesis:
   $H_0 : \beta_2 = \beta_3 = 0$ versus $H_1 :$ not both β_2 and $\beta_3 = 0$. Use $\alpha = 0.05$.

   (c) Compute the coefficient of partial determination between Y and X_2 given that X_1 is in the model,
   i.e., $R^2_{Y2,3|1}$. Interpret the result.

   (d) With the given information, can a **best** multiple linear regression model be selected using a forward selection procedure for this application? If yes, show your detailed procedure; if not, justify your answer.

Important t- and F-values for Problems 7-8:
$F_{0.05,2,12} = 3.89$, $F_{0.05,2,17} = 3.59$, $F_{0.05,1,12} = 4.75$, $F_{0.05,1,17} = 4.45$, $F_{0.05,1,16} = 4.49$, $F_{0.05,1,23} = 4.28$, $t_{0.05,12} = 2.179$, $t_{0.05,16} = 2.120$, $t_{0.01,16} = 2.921$, $t_{0.01,12} = 3.055$.

7. A textile company weaves a fabric on a large number of looms. It would like the looms to be homogeneous so that it obtains a fabric of uniform strength. The process engineer suspects that, in addition to the usual variation in strength within samples of fabric from the same loom, there may also be significant variations in strength between looms. To investigate this, she selects four looms at random and makes four strength determinations on the fabric manufactured on each loom. This experiment is run in random order, and the data obtained are shown in Table bellow:

	Observations: y_{ij}					
Looms	1	2	3	4	Mean: $\bar{y}_{i.}$	
1	98	97	99	96	390/4	
2	91	90	93	92	366/4	
3	96	95	97	95	383/4	
4	95	96	99	98	388/4	

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

(b) Find the values for A, B and C in the ANOVA table bellow.

Source	DF	Sum of Squares	Mean Square	F-Value	Pr > F
Model	3	C	B	D	E
Error	??	A	1.8958333		
Total	??	111.9375000			

(c) Is there significant variability due to Loom Differences (use $\alpha = 0.05$)?

(d) Estimate all variance components.

(e) If the overall goal of the company is to produce fabric with a mean of at least 95. Test if they can support this claim from study conducted.

8. An engineer is interested in the effects of cutting speed (A), tool geometry (B), and cutting angle (C) on the life (in hours) of a machine tool. Tow levels of each factor are chosen, and three replicates of a 2^3 factorial design are run. The results follow:

	factor		replicate		
A	B	C	I	II	III
-	-	-	22	31	25
+	-	-	32	43	29
-	+	-	35	34	50
+	+	-	55	47	46
-	-	+	44	45	38
+	-	+	40	37	36
-	+	+	60	50	54
+	+	+	39	41	47

(a) Write down the statistical model of the above experiment with the associated assumptions and constrains.

(b) Estimate the main effect of B.

(c) Please complete the missing values for the degrees of freedom (DF) and Type I SS in the following tables indicated by "???".

		Sum of
Source	DF	Squares
Model	???	???
Error	16	482.66666
Corrected Total	???	2195.333333

Source	Type I SS	DF
A	0.6666667	
B	770.6666667	
C	???	
AB	16.6666667	
AC	468.1666667	
BC	48.167	
ABC	???	

(d) Discuss the statistical significance of all the main effects and interactions based on the ANOVA table (including interaction effects and main effects when meaningful). Use $\alpha = 0.01$ in the hypothesis testing.

(e) Estimate a 98% confidence interval for $\mu_{1..} - \mu_{2..}$. Interpret it.