Statistics Qualifying Exam

9:00 am - 1:00 pm, Monday, August 17, 2015

1. Let X and Y have a joint probability density function (pdf)
$$f(x,y) = 2$$
for $0 < x < 1$ and $x < y < x + \frac{1}{2}$.

   (a) Find the covariance and correlation of X and Y.

   (b) Find the variance of $Z = 2X + 4Y$.

   (c) Find the conditional pdf of $Y|X = x$.

2. Let $X_1 \sim N(0,4)$, $X_2 \sim N(3,3^2)$ are normally distributed independent random variables
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right].$$

   (a) Find $P(3X_1 < X_2)$ (find the probabilities as a function of standardized normal distribution's cdf).

   (b) If $Y_1 = X_1 + X_2$, $Y_2 = X_1$, and $Y_3 = 3X_1 - X_2$, find the joint pdf of Y_1, Y_2, and Y_3. Are Y_1, Y_2, and Y_3 independent random variables?

3. Let $Y_1 < Y_2 < Y_3 < Y_4$ be the order statistics of a random sample of size $n = 4$ from a distribution with pdf $f(x) = 2x$, $0 < x < 1$, zero elsewhere.

   (a) Find the joint pdf of Y_3 and Y_4.

   (b) Find the conditional pdf of Y_3, given $Y_4 = y_4$.

   (c) Evaluate $E(Y_3|y_4)$.

4. Let $\overline{X}_n$ denote the mean of a random sample of size n from a Poisson distribution with parameter $\mu = 1$.

   (a) Show that the mgf of $Y_n = \sqrt{n}(\overline{X}_n - 1)$ is given by $\exp[-t\sqrt{n} + n(e^{t/\sqrt{n}} - 1)]$.

   (b) Investigate the limiting distribution of Y_n as $n \to \infty$.

   (c) Find the limiting distribution of $\sqrt{n}(\sqrt{\overline{X}_n} - 1)$.

5. Let $\overline{X}$ be the mean of a random sample of size n from $N(\theta, \sigma^2)$ distribution, $-\infty < \theta < \infty$, $\sigma^2 > 0$. Assume that σ^2 is known. Show that $\overline{X}^2 - \frac{\sigma^2}{n}$ is an unbiased estimator of θ^2 and find its efficiency.

6. A study is conducted to estimate the average difference in computation efficiency of analyzing data using two different statistical packages. To do so, 12 data sets are used. Each data set is analyzed by each of the two packages. The minutes needed for the analysis are recorded as shown in the following table

	j												
i	1	2	3	4	5	6	7	8	9	10	11	12	
Package I	23	25	21	22	21	22	20	23	19	22	19	21	$\bar{Y}_{1\cdot} = 21.5$
Package II	28	27	27	29	26	29	27	30	28	27	26	29	$\bar{Y}_{2\cdot} = 27.75$
													$\bar{Y}_{\cdot\cdot} = 24.625$

Note: $\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}$ represent the sample means for the statistical packages "Package I" and "Package II", respectively. $\bar{Y}_{\cdot\cdot}$ represent the grand sample mean.

(a) Use an appropriate two-sample t-test to test $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$, where μ_1 and μ_2 represent the average computation time for "Package I" and "Package II", respectively. Use $\alpha = 0.05$. Please clearly specify the assumptions made in the procedure.

(b) Assume an ANOVA model $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, where ϵ_{ij} are independent and identical random variables with $\epsilon_{ij} \sim N(0, \sigma^2)$, $i = 1, \ldots, a$, and $j = 1, \ldots, n$. Can this model be applied to the above data with $a = 2$ and $n = 12$? If yes, please conduct the F test for the equality of factor level means μ_1 and μ_2. If no, please explain.

7. 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 hypothesis. Use $\alpha = 0.05$.
$H_0 : \beta_1 = \beta_3 = 0$
H_1 : not both β_1 and β_3 are equal to 0.

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

8. An engineer is designing a battery for use in a device that will be subject to some extreme variations in temperatures. She decides to test three plate materials at three temperature levels. Because there are two factors at three levels, this design is sometimes called a 3^2 factorial design. Four batteries are tested at each combination of plate materials and temperature, and all 36 tests are run in random order. The experiment and the resulting observed battery life data are given in the table below. A longer life is preferred. The overall mean battery life of the sample is 105.53.

Material Type	Temperature($^\circ F$)		
	15	70	125
1	130, 155 74, 180	34, 40 80, 75	20, 70 82, 58
2	150, 188 159, 126	136, 122 106, 115	25, 70 58, 45
3	138, 110 168, 160	174, 120 150, 139	96, 104 82, 60

Part of SAS output is included from a two-factor fixed effects ANOVA analysis.

(a) Construct the ANOVA table based on the output from SAS.

i. Clearly specify the sources of sum of squares, the degrees of freedom, the mean squares, and the values of F statistics.

ii. State your findings. Use $\alpha = 0.05$ for each F test.

(b) Now assume it is given that $Temperature = 70^\circ F$, Carry out Tukey multiple comparison on the material types effect. Use $\alpha = 0.05$. If the exact critical value needed cannot be found in the statistical tables provided, please choose the most appropriate approximate value and briefly discuss how the approximation affects the decision.

Problem 8 SAS OUTPUT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	***	59416.22222	7427.02778	***	<.0001
Error	***	***	***		
Corrected Total	***	77646.97222			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
temp	***	39118.72222	***	***	<.0001
type	***	10683.72222	***	***	0.0020
temp*type	***	****	***	***	0.0186

Level of temp	N	life Mean	life Std Dev
15	12	144.833333	31.6940870
70	12	107.583333	42.8834750
125	12	64.166667	25.6721757

Level of type	N	life Mean	life Std Dev
1	12	83.166667	48.5888751
2	12	108.333333	49.4723676
3	12	125.083333	35.7655455

Level of temp	Level of type	N	life Mean	life Std Dev
15	1	4	134.750000	45.3532432
15	2	4	155.750000	25.6173769
15	3	4	144.000000	25.9743463
70	1	4	57.250000	23.5990819
70	2	4	119.750000	12.6589889
70	3	4	145.750000	22.5444006
125	1	4	57.500000	26.8514432
125	2	4	49.500000	19.2613603
125	3	4	85.500000	19.2786583

9. The article "The New Mantra: MVT" (*Forbes*, March 11, 1996, by Koselka, Rita) provides an interesting example on experimental design for a movie theater. Imagine the owner of the movie theater would like to maximize her weekly profit. Her options include: (A) Jack up the ticket price by a buck (B) Take out bigger ads in the local paper (C) Give away the popcorn.

   (a) The owner would like to test these three options in isolation. As a statistician, you need to persuade her to test all three at once. Please list your arguments.

   (b) Assume the experiment is carried out as you suggested (all three at once). The data obtained is as follows.

	A	B	C	
Test	Raise ticket price	Advertise	Give out free popcorn	Profit($ thousand)
1	NO	NO	NO	10
2	NO	NO	YES	15
3	NO	YES	NO	5
4	NO	YES	YES	10
5	YES	NO	NO	12
6	YES	NO	YES	20
7	YES	YES	NO	7
8	YES	YES	YES	15

   Please prepare a mini-report for this data to address the following two questions

   i. What are the estimated effects for the three options (A) Jack up the ticket price by a buck (B) Take out bigger ads in the local paper (C) Give away the popcorn?

   ii. What would you recommend to the owner of the movie theater? Note that since your report is going to be the reviewed by a statistician, simply list the seemingly "best" option (for example, raise ticket price + give out free popcorn) is not going to earn you any credit. Your report needs to include clear description of your statistical model and inference procedures along with sufficient statistical evidence to support your recommendation. In case that you need to calculate the sum of squares (SS) for a factor, the formula is $(contrast)^2/(2^k \cdot n)$, where k is the number of main factors and n is the number of replicates.