Multivariate Analysis Prelim Exam

Friday, August 14, 2020, 1:30 PM - 4:00 PM

1. In a study to compare two geographical regions (R), namely Northeast and Southwest, in terms of quality of health care of diabetic patients, three hospitals (H) were chosen at random from each of Northeast and Southwest regions.

   From a sample of 30 diabetic patients, five were randomly assigned to each hospital's diabetes clinic. Reduction in glucose level of the patients after taking a standard medication was recorded. Assume that the patients were all similar in terms of their baseline characteristics.

   Here, a hospital (H) is nested within a region (R). Let y_{ijk} be the measurement on k-th patient in j-th hospital from i-th region. Assume a standard nested model for y_{ijk} given by

   $$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{k(ij)}$$

   where α_i represents the region (R) effect, and $\beta_{j(i)}$ represents the hospital within region H(R) effect, and $\epsilon_{k(ij)}$'s are with iid normal errors.

   (a) State the usual assumptions about the terms in the model.

   (b) The sum of squares (SS) corresponding to H(R) is given by $\sum_i \sum_j \sum_k (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot})^2$. Find the expected mean square (EMS) for H(R) by using the algebraic method, (also known as the "brute force" method, not the EMS rule). Show work.

2. A study was conducted to compare two types of machines (M) used to wind coils. Four operators (O) were chosen at random from a large pool of operators. Three coils were wound for each of the eight machine-operator combination, in a completely randomized design. Let y_{ijk} be the measurement on k-th coil wound using i-th machine by j-th operator.

   (a) Write down the appropriate statistical model for y_{ijk} with the main effects and interaction, and state the assumptions about the terms in your model.

   (b) Find the mean and variance of $\bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot}$.

   (c) A student used the data from the experiment and calculated the following summary values:
   $\bar{y}_{1\cdot\cdot} = 8, \bar{y}_{2\cdot\cdot} = 3$, sum of squares(SS) for each main effect was 60, SS for each interaction was 30, and the SS for errors was 80.
   Give the ANOVA table (Source, DF, SS, MS, EMS). No need to derive the expressions for EMS.

   (d) The investigator wants to determine if there is a significant difference between the two types of machines. Give the value of the test statistic, and state the distribution you would use to get the critical value.

3. An experiment was conducted to compare the yield of three different varieties of corns (C) and four different levels of manure (M). The experimental area was divided into 6 blocks. Each of these was then

subdivided into 3 large plots. The varieties of corn were sown on the large plots according to a randomized complete block design (so that every variety appeared in every block exactly once). Each large plot was then divided into 4 small plots, and the levels of manure were applied to the small plots according to a randomized complete block design (so that every level of M appeared in every large plot exactly once).

A SAS code using

```
PROC GLM ;
MODEL Y = C M C*M Block Block*C ;
```

gave the following SS:
SS(Total)=51985, SS(C)=1786, SS(Block)=15875, SS(M)= 20020, SS(C*Block)=6013, SS(C*M)=321.

Sample means of the three corn varieties 1-3 were, 97.0, 104.0 and 109.0.

(a) Identify the design and write down an appropriate statistical model for this experimental data. Be sure to state the assumptions.

(b) Give the ANOVA table with SS, DF, MS and F-values that you would use to answer the questions below.

(c) Do the varieties of corns and the levels of manure have significant interaction effects? Write down the hypotheses you would test, and give the test statistic and what distribution you would use to get the critical value.

(d) Find the 95% confidence intervals for the difference in the mean response for corn varieties 1 and 2. You may leave the answer in terms of percentile of known distribution.

4. A study was conducted to compare the effectiveness of a drug (to reduce blood pressure) given at four different dose levels, 10, 20, 30 and 40 mgs. Twelve patients were randomly selected and they were randomly assigned to each dose, so that each dose is assigned to three patients. For each patient in the study, blood pressure level y was measured on each of three consecutive days (at noon); Day 1, 2, 3, after the dose was administered.

   The data was used in SAS software to fit model indicated by the code below.

```
proc glm; class dose;
model y1 y2 y3 = dose/nouni;
repeated time / printe;
run;
```

   Here, `y1`, `y2`, and `y3` denote the measurement taken on a patient at Day 1, 2, and 3, respectively.

   A partial output from SAS is attached.

   For the purpose of writing an equivalent statistical model, let y_{ijk} be the response from j-th patient receiving i-th Dose, at Day k, $i = 1, 2, 3, 4; j = 1, 2, 3; k = 1, 2, 3$; and let $Y_{ij} = (y_{ij1}, y_{ij2}, y_{ij3})'$ be the response vector for i-th patient receiving Dose j.

(a) Write down the multivariate version of the statistical model fitted to the response vector Y_{ij}, identify the parameters in the model and explain their meaning, and state any assumptions you make.

(b) Univariate Analysis

i. SAS output includes Univariate Tests of Hypotheses for Within Subject Effects at the bottom of the output. State an assumption under which the univariate analysis is valid, and write down the null and alternative hypotheses you would use to test this assumption. Use the output to give the p-value and conclusion.

ii. Use the univariate analysis output to test for significance of the time effect. Give the value of the test statistic, and conclusion.

(c) Multivariate Analysis

i. SAS output provides a test of hypotheses for Between Subjects Effects. State the null and alternative hypotheses being tested here.
State the degrees of freedoms associated with the test statistic. In simple terms, interpret the conclusion of the test.

ii. SAS output also includes results of a MANOVA test for "no time effect."
State the null and alternative hypotheses tested here in terms of the parameters in the statistical model, and state the conclusion in simple terms relating to the current context.

SAS OUPUT to accompany Multivariate Analysis Prelim (2 pages)

Sphericity Tests				
Variables	**DF**	**Mauchly's Criterion**	**Chi-Square**	**Pr > ChiSq**
Transformed Variates	2	0.5500635	4.1840508	0.1234
Orthogonal Components	2	0.5534131	4.1415532	0.1261

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of no time Effect
H = Type III SSCP Matrix for time
E = Error SSCP Matrix

S=1 M=0 N=2.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.02198787	155.68	2	7	<.0001

MANOVA Test Criteria and F Approximations for the Hypothesis of no time*dose Effect
H = Type III SSCP Matrix for time*drug
E = Error SSCP Matrix

S=2 M=0 N=2.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.14665710	3.76	6	14	0.0193

NOTE: F Statistic for Wilks' Lambda is exact.

The GLM Procedure
Repeated Measures Analysis of Variance
Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
dose	3	139.7327548	46.5775849	16.03	0.0010
Error		23.2465985	2.9058248		

The GLM Procedure
Repeated Measures Analysis of Variance
Univariate Tests of Hypotheses for Within Subject Effects

						Adj Pr > F	
Source	**DF**	**Type III SS**	**Mean Square**	**F Value**	**Pr > F**	**G - G**	**H-F-L**
time		107.6869832			<.0001	<.0001	<.0001
time*dose		13.5586881	2.2597813	2.57	0.0614	0.0953	0.0827
Error(time)		14.0600905	0.8787557				