STAT 501 **Assignment #4** **Name ____________**
Spring 2005

Reading Assignment: Johnson &Wichern, Chapter 8 and 9

Written Assignment: Due in class on Monday, March 28. We did not leave enough space for you to present your answers on this assignment. Please present your answers on other paper. You can do problem 1 without a computer.

1. In a study of the effects on future development of the size of apple trees at the time of transplanting, four measurements of size were made on each of 54 trees when they were transplanted. The four measured traits are:

X_1 = tree weight in kg.

X_2 = the square of the trunk circumference in cm^2.

X_3 = the length of the laterals in cm. (The total length of the branches.)

X_4 = the length of the central leader in cm. (A measure of the height of the tree.)

The sample correlation matrix is:

$$R = \begin{bmatrix} 1.0000 & & & \\ .7571 & 1.0000 & & \\ .7792 & .6684 & 1.0000 & \\ .5500 & .6017 & .3207 & 1.0000 \end{bmatrix}$$

The eigenvalues and eigenvectors associated with R are

eigenvalues: 2.855 0.721 0.253 $\hat{\lambda}_4$

$$\text{eigenvectors:} \quad \begin{bmatrix} .55 \\ .54 \\ .49 \\ .42 \end{bmatrix} \begin{bmatrix} -.16 \\ .06 \\ -.56 \\ .81 \end{bmatrix} \begin{bmatrix} .35 \\ -.84 \\ .26 \\ .31 \end{bmatrix} \begin{bmatrix} .74 \\ .02 \\ -.61 \\ -.28 \end{bmatrix}$$

A. Find the value of $\hat{\lambda}_4$

B. Write down the formula for the first principal component.

C. What is the percentage of total variance in the standardized variables accounted for by the first principal component?

D. Give a one-sentence interpretation of the first principal component.

E. Give a one sentence interpretation of the second principal component.

F. Estimate the correlation between the second principal component and the standardized length of the laterals.

G. Estimate the correlation between the scores for the first and second principal components.

H. Describe the information provided by third and fourth principal components.

2. Ramsey, et al. (1994), Case Studies in Biometry) report habitat data from a study of the northern spotted owl (*strix occidentalis*). The U. S. Fish and Wildlife Service has declared that the northern spotted owl is a threatened species, and under the Endangered Species Act the owl's survival must take priority over other uses of its habitat, "old growth" forest. Environmental activists have tried to use this law to prevent companies from cutting down the remaining virgin forest in the Pacific Northwest region of the United States.

   Ensuring survival requires an understanding of how much old growth forest is required by the spotted owl. Studies of such questions are called habitat preference or habitat association studies. The data posted in the file *spotowls.dat* on the course web page were obtained from a study where percentage of mature forest was recorded for seven concentric rings with outer diameters of 0.91, 1.18, 1.40, 1.60, 1.77, 2.41, 3.38 km, respectively, around each of 30 spotted owl nesting sites in a 7100 $(\text{km})^2$ region of National Forest in Western Oregon. Thirty other sites where selected from the same region at random coordinates, and percentages of mature forest were ascertained in seven concentric rings of the same diameters at each of those sites.

   There is one line of data in the file for each site. Values in the first column are coded "N" for a nesting site and "R" for a random site. The remaining seven columns contain percentages of mature forest within consecutive rings with outer diameters of 0.91, 1.18, 1.40, 1.60, 1.77, 2.41, 3.38 km, respectively.

   a. Compute principal components for the percentages of mature forest for the seven rings. Use the sample correlation matrix for all 60 sites. Write down the formulas for the first two principal components. Give an interpretation of each of these components.

   b. Describe how the first two principal components correspond to patterns in the correlation matrix.

   d. What proportion of the total sample variance of the standardized variables is accounted by the two principal components described in part a?

d. How many components are needed to adequately describe the variation in the standardized percentages of old growth forest at these 60 sites. Explain.

e. Plot scores for the important principal components in a scatter plot matrix. Use different symbols or colors for nesting sites (N) and randomly selected sites (R). What does this plot reveal?

f. Examine these data with the XGOBI or GGOBI package. Describe what this analysis revealed to you.

g. Test the null hypothesis that the vectors of mean percentage of "old growth" forest are the same for nesting sites and random sites. Give a formula for your test statistic, degrees of freedom, and a p-value. State your conclusions.

h. Compute univariate t-tests to compare mean percentage of old growth forest between nesting and random sites within each ring size. Also, compute the two sample Wilcoxon tests for each ring. State your conclusions.

3. Data from 52 census tracts in the vicinity of El Paso, Texas, were derived from the 1970 U. S. Census of Population and Housing. These data are posted under *census.dat* on the course web page. Each line of the file provides data for a different census tract. The nine variables on each line are

ID	Identification code for census tracts
X1	Percentage of population age 16-21
X2	Median family income (dollars)
X3	Percentage of population over age 25 who are high school graduates
X4	Percentages of population with Spanish surnames
X5	Percentages of housing units with more than 1 person per room
X6	Unemployment rate
X7	Percentage of housing units occupied by owners
X8	Percentage of population under age 16

(The U. S. Census Bureau divides the United States into thousands of non-overlapping areas called Census Tracts.) Perform a principal component analysis of the standardized values for X1-X8, and answer the following questions.

a. The proportion of the total variance of the standardized variables accounted for by the first three principal components is _________

b. Give one sentence interpretations of the first three principal components

c. Explain why it is better to perform the principal component analysis on the sample correlation matrix rather than the sample covariance matrix for these data.

d. Is there any reason to use more or less than 3 principal components? Explain.

e. Examine these data with the XGOBI or GGOBI package. Does this graphical analysis reveal any additional interesting aspects of these data? Explain.

4. The data posted in the file broota.dat were obtained from an experiment where two groups of subjects (standard training (ST) or enhanced training (ET)) were required to perform two tasks (task 1 and task 2) under three different cue conditions (normal (N), congruent (C), and incongruent (I)). The measured responses were the times (in hundredths of a second) required to complete each task under the three different cue conditions, so each subject provided 6 response times. Subjects were randomly assigned to the ST and ET groups with 12 subjects in each group. The subjects received individual training, they never met each other, and they were tested at different times. Hence, it is reasonable to assume that each subject responded independently of any other subject.

A. Complete the ANOVA table for the following model:

$$X_{ijkl} = \mu + \gamma_k + \eta_{kl} + \alpha_i + \alpha\gamma_{ik} + \delta_{ikl} + \beta_j + \beta\gamma_{jk} + \psi_{jkl} + \alpha\beta_{ij} + \alpha\beta\gamma_{ijk} + \varepsilon_{ijkl}$$

where

γ_k is a training type effect ($\gamma_2 = 0$)

α_i is a task effect ($\alpha_2 = 0$)

β_j is a cue condition effect ($\beta_3 = 0$)

$\eta_{kl} \sim NID(0, \sigma^2_\eta)$ $\psi_{jkl} \sim NID(0, \sigma^2_\psi)$

$\delta_{ikl} \sim NID(0, \sigma^2_\delta)$ $\varepsilon_{ijkl} \sim NID(0, \sigma^2_\varepsilon)$

and each random term is independent of any other random term. Code for completing the ANOVA table is posted in the files broota05.sas and broota05.R.

Source of variation	df	SS	MS	F	p-value
Training groups					
Subjects (groups)					
Tasks					
Groups × Task					
error (b)					
Cues					
Group × Cues					
error (c)					
Tasks × Cues					
Group × Task × Cues					
error (d)					
Corrected total					

B. Describe what the model in part A implies about the covariance structure for observations taken on the same subject.

C. What can you conclude from this analysis?

5. Consider a MANOVA approach to the data in problem 4 using the model $X = A\beta + \varepsilon$ where

$$\beta = \begin{bmatrix} \mu_{11} & \mu_{12} & \mu_{13} & \mu_{21} & \mu_{22} & \mu_{23} \\ \gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{21} & \gamma_{22} & \gamma_{23} \end{bmatrix}$$

The fist three columns of β correspond to mean completion times for task 1 under the three different cue conditions and the last three columns of β correspond to mean completion times for task 2 under corresponding cue conditions. With respect to the parameters in the model given in part A of Problem 4,

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij}$$

represents how the mean completion times for the ST treatment (group 2) change across tasks and cues and

$$\gamma_{ij} = \gamma_1 + (\alpha\gamma)_{i1} + (\beta\gamma)_{j1} + (\alpha\beta\gamma)_{ij1}$$

represents how differences between mean completion times for the ET and ST training change across tasks and cues. Show how to write each of the following null hypotheses in the form

$$H_0:\ C_{k\times r}\, B_{r\times p}\, M_{p\times u} = O_{k\times u}$$

by listing appropriate choices for C and M. Also, compute a value for Wilks criterion (Λ) and the corresponding F-value, its degrees of freedom, and a p-value. State your conclusion.

A. No difference between training groups with respect to the mean completion time of any task under any cue condition.

B. For each type of training there is no difference between mean completion times for the two tasks, averaging across cue conditions.

C. For each type of training, there are no differences among mean completion times for the three cue conditions, averaging across tasks.

D. No interaction between type of training and type of task, averaging across cue conditions.

E. No interaction between type of training and cue condition, averaging across tasks.

F. No three factor interaction between type of training, type of task, and cue conditions.

G. No interaction between type of task and cue conditions, averaging across training groups.