STAT 501 **Assignment 5** **Name _____**
Spring 2005

Reading Assignment: Johnson & Wichern, Chapter 9 and Chapter 10

Written Assignment: Due Friday, April 11, in class. Problems 2, 3, and most of Problem 4 can be done without use of a computer.

1. Consider the data from problem 3 on assignment 4. These data are from 52 census tracts in the vicinity of El Paso, Texas, and 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.) Use the standardized values for X1-X8 to answer the following questions.

   a. Report the estimated loadings for a varimax rotation of the first 3 principal components, and give a one sentence interpretation of each of those components.

   b. Report the estimated loadings for a promax rotation of the first 3 principal components, and give a one sentence interpretation of each of those components. Also report the correlations among the factor scores.

2. The correlation matrix for chicken-bone measurements (see example 9.14 in J&W) is given in problem 9.10 in Johnson and Wichern's book. Maximum likelihood estimates of factor loadings were obtained for $m=2$ factors and the varimax rotated estimated factor loadings are reported in the following table.

Variable	Rotated Factor Loadings F_1^2	F_2^*	Communalities	Specific Variances
Skull length	.484	.411		
Skull breadth	.375	.319		
Femur length	.603	.717		
Tibia length	.519	.855		
Humerus length	.861	.499		
Ulna length	.744	.594		

a. Report estimates of communalities and specific variances for these two factors.

b. Report the proportion of total standardized sample variation accounted for by each rotated factor, and give a one sentence interpretation of each rotated factor.

c. How do the loadings for the rotated factors correspond to patterns in the correlation matrix?

d. Construct the residual matrix $r - \hat{L}_z \hat{L}_z' - \hat{\varphi}_z$.

e. How would the answers to parts a and b change if you used unrotated factors?

f. Make a plot of the loadings for the two rotated factors (sketch this by hand), and draw in what you anticipate the factors for a promax rotation would look like. (You can check your intuition by running the SAS program posted in the file *chbones.sas*.)

3. Consider a toxicology experiment where pregnant mice are exposed to a toxic substance in their diet. Measurements $\underset{\sim}{X}_j = \left(X_{1j}, X_{2j}, X_{3j}, X_{4j}\right)'$ of the serum level of the toxic substance are taken a two days after birth from of the four largest pups born to the j-th pregnant female. If the results for the four littermates are equally correlated, then the correlation matrix is

$$\begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

a. Show that $\underset{\sim}{e}_1 = (1/2, 1/2, 1/2, 1/2)'$ is an eigenvector for this correlation matrix. Obtain a formula for the corresponding eigenvalue as a function of ρ?

b. What are the eigenvalues for the eigenvectors

$$\underset{\sim}{e}_2 = \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0, 0\right)', \qquad \underset{\sim}{e}_3 = \left(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{-2}{\sqrt{6}}, 0\right)' \quad \text{and}$$

$$\underset{\sim}{e}_4 = \left(\frac{1}{2\sqrt{3}}, \frac{1}{2\sqrt{3}}, \frac{1}{2\sqrt{3}}, \frac{-3}{2\sqrt{3}}\right)' \quad ?$$

c. Show that this correlation matrix can be written in orthogonal factor model form with just one factor. Report the factor loadings and the specific variances.

d. If a correlation matrix can be exactly written in orthogonal factor model form, using just one factor, do the correlations all have to be equal? Explain.

4. The early development of factor analysis was due to Charles Spearman (1904). He studied correlations between scores for various types of tests and noted that many observed correlations could be described by a simple model. For example, Spearman obtained the following matrix of correlations for 33 boys in a preparatory school for their scores on tests in Classics (C), French (F), English (E), Mathematics (M), Discrimination of pitch (D), and Music (Mu).

	C	F	E	M	D	Mu
C	1.00	0.83	0.78	0.70	0.66	0.63
F	0.83	1.00	0.67	0.67	0.65	0.57
E	0.78	0.67	1.00	0.64	0.54	0.51
M	0.70	0.67	0.64	1.00	0.45	0.51
D	0.66	0.65	0.54	0.45	1.00	0.40
Mu	0.63	0.57	0.51	0.51	0.40	1.00

For this matrix Spearman noted that any two rows are nearly proportional if the diagonals are ignored. For example, for rows C and E the ratios of correlations are:

$$\frac{0.83}{0.67} \cong \frac{0.70}{0.64} \cong \frac{0.66}{0.54} \cong \frac{0.63}{0.51} \cong 1.2$$

This caused Spearman to propose the model $X_i = a_i F + \varepsilon_i$ where X_i is the standardized score for the i-th test (with a mean of zero and a standard deviation of one), a_i is a constant, F is a random 'factor' that corresponds to the boy's overall ability (with a mean of zero and standard deviation of one), and ε_i is the part of X_i that is specific to the i-th test only.

a. Show that a constant ratio between each pair of rows in a correlation matrix is a consequence of the model $X_i = a_i F + \varepsilon_i$.

b. How are the factor loadings $(a_1, a_2, a_3, a_4, a_5, a_6)$ related to the variances of the ε_i's?

c. Use the test on pages 488-490 in Chapter 8 of your textbook to test the null hypothesis that all correlations are equal. Report values for

test statistic = _____ d.f. = _____ p-value = _____

d. Use PROC FACTOR to analyze this correlation matrix. Use the maximum likelihood estimation method. SAS code is posted in the file *spearman.sas*. This shows you how to use a correlation matrix with PROC FACTOR.

(i) Does a one factor model appear to be adequate? Report values for

Chi-squared test value = _____ d.f. = _____ p-value = _____

(ii) Is your answer for part (i) consistent with your result in part c.? Explain.

5. The data for this example consist of scores for 24 psychological tests given to 145 seventh and eighth grade school children in a suburb of Chicago. These data were collected by Karl Holzingen and Frances Swineford (1937), Psych. 2, 42-54) and they have become a classic example in the factor analysis literature. Descriptions of the 24 test are given below. These were taken from Holzinger, Karl L. & Harry H. Harman, *Factor Analysis*, Univ. of Chicago Press, 1941.

1. *Visual Perception Test.* A nonlanguage multiple-choice test composed of items selected from Spearman's Visual Perception Test, Part III. Testing time: 19 minutes.
2. *Cubes.* A simplification of Brigham's test of spatial relations. Testing time: 8 minutes.
3. *Paper Form Board.* A revised multiple-choice test of spatial imagery, with dissected squares, triangles, hexagons, and trapezoids. Testing time: 8 minutes.
4. *Flags.* Adapted from a test by Thurstone. Requires visual imagery in two or three dimensions. Testing time: 5 ½ minutes.
5. *General Information.* A multiple-choice test of a wide variety of simple scientific and social facts. Testing time: 18 minutes.

6. *Paragraph Comprehension.* Part III of Traxler Silent Reading Test, Form 1, for Grades VII-X. Comprehension measured by completion and multiple-choice questions. Testing time: 20 minutes.
7. *Sentence Completion.* A multiple-choice test in which "correct" answers reflect good judgment on the part of the subject. Testing time: 6 minutes.
8. *Word Classification.* Arranged by M.A. Wenger. Sets of five words one of which is to be indicated as not belonging with the other four. Testing time: 10 minutes.
9. *Word Meaning.* Part II of Traxler Silent Reading Test. A multiple-choice vocabulary test. Testing time: 14 minutes.
10. *Add.* Speed of adding pairs of one-digit numbers. Testing time: 2 minutes.
11. *Code.* A simple code of three characters is presented and exercise therein given to measure perceptual speed. Testing time: 2 minutes.
12. *Counting Groups of Dots.* Four to seven dots, arranged in random patterns, to be counted by subject. A test of perceptual speed. Testing time: 4 minutes.
13. *Straight and Curved Capitals.* A series of capital letters. The subject is required to distinguish between those composed of straight lines only and those containing curved lines. A test of perceptual speed. Testing time: 3 minutes.
14. *Word Recognition.* Twenty-five four-letter words are studied for three minutes. These words are then to be checked from memory on a hundred-word list. Testing time: 5 minutes. (Score includes two forms.)
15. *Number Recognition.* Similar to Test 14. Fifteen three-digit numbers.
16. *Figure Recognition.* Similar to Test 14. Fifteen geometric designs.
17. *Object*-Number. Twenty pairs of names of familiar objects and two-digit numbers are studied for three minutes. The words only are then presented to the subject, who is required to supply the proper numbers. Testing time: 5 minutes. (Score includes two forms.)
18. *Number-Figure*. Similar to Test 17. Ten pairs of numbers and geometric figures.
19. *Figure-Word.* Similar to Test 17. Ten pairs of geometric figures and words studied for one minute.
20. *Deduction.* Logical deduction test using the symbols) and (and the letters A, B, C, and D. Testing time: 24 minutes.
21. *Numerical Puzzles.* A numerical deduction test, the object being to supply four numbers which will produce four given answers employing the operations of addition, multiplication, or division. Testing time: 14 minutes.
22. *Problem Reasoning.* A reasoning test in completion form. Each problem lists the steps in obtaining a required amount of water using two or three vessels of given capacity. Testing time: 14 minutes.
23. *Series Completion.* From a series of five numbers the subject is supposed to deduce the rule of procedure from one number to the next, and thus supply the sixth number in the series. Testing time: 14 minutes.
24. *Woody-McCall Mixed Fundamentals: Form I.* A series of 35 arithmetic problems, graduated for difficulty, is included. Testing time: 20 minutes.

Means and standard deviations for the scores are presented below for each of the 24 tests.

Basic Statistics for Twenty-Four Psychological Tests

Test X_j	Mean $\overline{X}_j$	Standard Deviation s_j
1. Visual Perception	29.60	6.90
2. Cubes	24.84	4.50
3. Paper Form Board	15.65	3.07
4. Flags	36.31	8.38
5. General Information	44.92	11.75
6. Paragraph Comprehension	9.95	3.36
7. Sentence Completion	18.79	4.63
8. Word Classification	28.18	5.34
9. Word Meaning	17.24	7.89
10. Addition	90.16	23.60
11. Code	68.41	16.84
12. Counting Dots	109.83	21.04
13. Straight-Curved Capitals	191.81	37.03
14. Word Recognition	176.14	10.72
15. Number Recognition	89.45	7.57
16. Figure Recognition	103.43	6.74
17. Object-Number	7.15	4.57
18. Number-Figure	9.44	4.49
19. Figure-Word	15.24	3.58
20. Deduction	30.38	19.76
21. Numerical Puzzles	14.46	4.82
22. Problem Reasoning	27.73	9.77
23. Series Completion	18.82	9.35
24. Arithmetic Problems	25.83	4.70

The 24x24 sample correlation matrix has been posted in the file *tests.cor*. There are two rows in the data file for each row of the correlation matrix. This file also contains sample sizes, means and standard deviations in the first six lines of data. PROC FACTOR in SAS can read this file as a TYPE=CORR data set which can be accomplished with the code posted in the file *tests.sat*. Use this correlation matrix to answer the following questions.

a. Perform a principal component analysis with a varimax rotation of the first 5 components and give a one sentence interpretation of each of the rotated components. Also report the percentage of the total variance of the standardized test scores explained for each of the five rotated components

b. Perform a factor analysis with the iterated principal factor method using 5 factors. Do a varimax rotation and give a one sentence interpretation of each of the rotated factors.

c. Obtain maximum likelihood estimates for factor loadings for the first five factors and perform a varimax rotation. Make comparisons with the results to parts a. and b.

d. Examine the data more thoroughly to determine how many factors are needed. Give some justification for your answer. If your answer is not 5 factors, compute maximum likelihood estimates for factor loadings for the appropriate number of factors (if possible), perform a varimax rotation, and give interpretations of the rotated factors.

e. Try a PROMAX rotation for the number of factors selected in part d. Make comparisons with the factors obtained from varimax rotation.

f. Try a QUARTIMAX rotation for the same number of factors and make comparisons with results in parts d. and e.

6. The counts in the following table were obtained from a survey of customers of overnight delivery services. Each respondent was classified with respect to the favorite shipper and size of the respondent's business (row categories) and also with respect to the most important reason for choosing their favorite shipper (column category). These data are posted on the course web page as shippers.dat.

Columns

Row	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Totals
1	4.00	5.00	5.00	20.00	1.00	21.00	2.00	1.00	22.00	8.00	3.00	13.00	1.00	4.00	9.00	119.00
2	2.00	5.00	7.00	13.00	1.00	16.00	4.00	2.00	16.00	10.00	6.00	10.00	7.00	6.00	5.00	110.00
3	9.00	9.00	8.00	12.00	2.00	17.00	5.00	3.00	16.00	11.00	6.00	13.00	10.00	7.00	11.00	139.00
4	5.00	11.00	10.00	20.00	1.00	15.00	3.00	2.00	25.00	10.00	6.00	14.00	10.00	6.00	8.00	146.00
5	25.00	25.00	19.00	19.00	20.00	11.00	11.00	8.00	19.00	19.00	13.00	2.00	15.00	15.00	10.00	231.00
6	25.00	25.00	21.00	20.00	19.00	14.00	12.00	10.00	18.00	20.00	15.00	3.00	15.00	18.00	14.00	249.00
7	22.00	23.00	18.00	15.00	20.00	12.00	14.00	8.00	15.00	22.00	19.00	2.00	14.00	16.00	11.00	231.00
8	17.00	17.00	18.00	11.00	11.00	10.00	11.00	3.00	10.00	14.00	12.00	1.00	13.00	11.00	12.00	171.00
9	7.00	11.00	8.00	7.00	3.00	7.00	3.00	2.00	10.00	7.00	8.00	13.00	6.00	3.00	5.00	100.00
10	12.00	13.00	13.00	9.00	11.00	8.00	12.00	9.00	9.00	14.00	10.00	1.00	10.00	11.00	7.00	149.00
Totals	128.00	144.00	127.00	146.00	89.00	131.00	77.00	48.00	160.00	135.00	98.00	72.00	101.00	97.00	92.00	1645.00

Descriptions of the row and column categories are given below.

Rows (Supplier and buyer site category)	Columns (Supplier attributes)
1. Alpha – very small (SSA)	1. Almost always delivers the package by the promised time
2. Alpha – small (SA)	2. Packages are almost always delivered in good condition
3. Alpha – large (LA)	3. Almost never late in pickup of packages to be shipped
4. Alpha – very large (LLA)	4. Little paperwork is required in preparing packages for shipment
5. Gamma – very small (SSG)	5. Will make special trips to pick up packages if needed
6. Gamma – small (SG)	6. Easy to calculate the shipping costs
7. Gammas – large (LG)	7. Easy to trace a package that has gone astray
8. Gamma – very large (LLG)	8. Sensitive to customer needs in settling claims
9. Beta – (BTA)	9. Capable of shipping packages almost anywhere I want
10. Delta – (DLT)	10. Friendly and congenial employees
	11. Extremely efficient in all their business dealings
	12. Less expensive than most
	13. Responsible and dependable in customer relations
	14. Uses the most advanced technology
	15. Really interested in the smaller customer

Use the code posted as shippers.sas or shippers.R to do a correspondence analysis of this two-way table of counts.

(a) Report the value of the Pearson chi-square test for independence and its degrees of freedom and p-value. State your conclusion.

(b) Inspect the results from the correspondence analysis and describe what it reveals about relationships between supplier attributes (column categories) and the combined supplier and customer size categories (row categories).

7. (Optional, do not submit written answers.) Do Problems 9.1, 9.2, 9.3, 9.4, 9.7, 9.8 and 9.9 in the text. Answers for these problems will be provided.