**STAT 501** **FINAL EXAM** **Name _____**
**Spring 1999**

Instructions: Present your answers in the spaces provided on this examination paper. It is not necessary to exhibit all of your calculations, but clearly describe the formulas or methods you use. Credit will be given when the appropriate methods are used even though the final numerical answer is incorrect. Use the back of the page if you need more space, but clearly indicate where this is done.

1. Data were collected on seven soybean plants. Five characteristics were measured on each plant. The Euclidean distances between the vectors of measurements are displayed in the following table for all pairs of plants.

| | Plant | | | | | | |
|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| Plant 1 | -- | 33 | 37 | 24 | 31 | 36 | 39 |
| Plant 2 | | -- | 42 | 22 | 39 | 42 | 35 |
| Plant 3 | | | -- | 41 | 45 | 30 | 42 |
| Plant 4 | | | | -- | 41 | 32 | 40 |
| Plant 5 | | | | | -- | 46 | 48 |
| Plant 6 | | | | | | -- | 34 |
| Plant 7 | | | | | | | -- |

Use the complete linkage clustering procedure to make three clusters.

2. In a marketing survey, a random sample of n = 1200 consumers were asked to indicate the importance of each of five attributes of a new food product. The responses were recorded and the sample correlation matrix R was computed. The two largest eigenvalues of R are $\hat{l}_1 = 3.3$ and $\hat{l}_2 = 1.2$. The corresponding eigenvectors are presented in the following table.

| **Attribute** | **First eigenvector** | **Second eigenvector** |
|---|---|---|
| 1. Taste | .39 | .61 |
| 2. Appearance | .42 | .54 |
| 3. Good Source of Calcium | .48 | -.32 |
| 4. Good Source of Vitamins | .46 | -.29 |
| 5. Low in Calories | .47 | -.38 |

(a) Provide interpretations for the first and second principal components.

(b) What is the proportion of the total variance of the standardized variables that is accounted for by the first two principal components?

_____

(c) After seeing the two principal components what do you think the sample correlation matrix looks like? Which pairs of variables would have high positive correlations, which would have negative correlations, which would have moderate or small correlations?

(d) The estimated variance of the scores for the second principal component is

_____

(e) The estimated correlation between the value for taste and the score for second principal component is

_____

(f) With respect to the factors defined by the two principal components shown above, the communality for the low in calories attribute is

_____

(g) What is the primary objective of a varimax rotation of the first two principal components? Describe the two principal components that would be produced by a varimax rotation.

(h) Describe the computational procedure used by the iterated principal factor method to obtain factor loadings for the factor analysis of a correlation matrix.

(i) Describe the orthogonal factor model in which the measured traits and the observed factors have independent normal distributions. Show how this model is used to derive theformula for factor scores that Johnson and Wichern call the "Regression Method" in Chapter 9 of their book.

3. An experiment was performed to study how the effectiveness of a certain drug changes as the dosage changes. Independent samples of 24 male subjects and 24 female subjects were used in the study. Each subject used each of the $p = 4$ dosage levels for one month. At the end of the month the response was measured, so each subject provided $p = 4$ responses, one for each dosage level. The drug was administered in pill form and pills for the various dosage levels were made to look the same to prevent the subjects from knowing the order in which the dosage levels were given. One male and one female were randomly assigned to each of the 24 possible orderings of the presentation of the 4 drug levels.

   (a) Consider the null hypothesis that the mean responses for males are the same for all 4 drug levels in the male population and the mean responses for females are the same for all four drug levels. This null hypothesis does <u>not</u> require males to have the same mean response as females. the value for Wilks criterion is 0.12. Report the value of the corresponding F-test and its degrees of freedom.

   (b) Now consider the null hypothesis that the mean response curve for males is parallel to the mean response curve for females (i.e., there is no interaction between sex and dosage level). The value for Wilks criterion is 0.38. Report the value of the corresponding F-test and its degrees of freedom.

4. Two tests are to be used to determine whether a cow has a certain disease. For each test the cow gives a binary response: positive (+) or negative (-). For diseased cows, probabilities of the four possible outcomes are known to be

| Test Results | Probability |
|---|---|
| Both tests are positive | 0.90 |
| Test A is positive and Test B is negative | 0.01 |
| Test A is negative and Test B is positive | 0.05 |
| Both tests are negative | 0.04 |

For cows that do not have the disease, probabilities of the four possible outcomes are known to be

| Test Results | Probability |
|---|---|
| Both tests are positive | 0.04 |
| Test A is positive and Test B is negative | 0.06 |
| Test A is negative and Test B is positive | 0.08 |
| Both tests are negative | 0.82 |

The results of the two tests are to be used to classify a cow as diseased or not diseased. Construct a classification rule to minimize the expected cost of misclassification when the prior probability of observing a diseased cow is 0.1 and the cost of misclassifying a diseased cow as not diseased is twenty times greater than the cost of classifying a healthy cow as diseased.

5. Anyone who has applied for a loan or a credit card has supplied information on an application form about age, income, employment history, credit history, other pertinent information. Data bases containing such information are maintained by several credit information companies who try to obtain as much information as they can on individuals in their data bases. This information is sold to banks and other customers.

Suppose you are employed by a bank that had previously purchased a data file containing information on 400,000 individuals from one of these credit information companies and mailed an application for a new credit card to each individual in the data file. They received responses from 8,257 of these individuals. The bank wants you to use the information on these 400,000 individuals to construct a classification rule to determine which type of potential customer is most likely to respond to a mail offer to acquire a new credit card. This will enable the bank to save money by not sending mail to people who are unlikely to respond. The data file contains information on the following variables for each of the 400,000 individuals.

| **VARIABLE** | **DESCRIPTION** |
|---|---|
| ID | Identification code |
| Response | Coded 0 = did not respond<br>1 = responded to the mailing |
| $X_1$ | Age in years |
| $X_2$ | Sex coded 0 for female<br>1 for male |
| $X_3$ | Annual income in thousands of dollars |
| $X_4$ | Number of credit card transactions in the last three months |
| $X_5$ | Number of credit card transactions in the last six months |
| $X_6$ | Number of credit card transactions in the last 12 months |
| $X_7$ | Total number of active credit cards |
| $X_8$ | Total credit limit on all currently active credit cards |
| $X_9$ | Total outstanding balance on all currently active credit cards |
| $X_{10}$ | Total debt on credit cards, auto loans, home loans, and other loans |
| $X_{11}$ | Level of education coded<br>1 = did not graduate from high school<br>2 = high school graduate<br>3 = college graduate<br>4 = professional or advanced degree |
| $X_{12}$ | Home ownership, coded<br>1 owns a home with no mortgage<br>2 owns a home with a mortgage<br>3 rents a house or apartment<br>4 other |

| | |
|---|---|
| $X_{13}$ | Number of times a loan or credit card payment was more than 30 days overdue in the last 5 years |
| $X_{14}$ | Number of jobs held in the last five years |
| $X_{15}$ | Number of months of unemployment in the last 5 years |
| $X_{16}$ | Marital status: coded<br>1 for single<br>2 for married<br>3 for divorced<br>4 for other |
| $X_{17}$ | Number of children |

Outline the steps you would take to develop a classification rule from these data. Be sure to include some explanation of how you would assess the ability of your rule to correctly classify. (Use the back of this page if needed.)

**EXAM SCORE _____** **COURSE GRADE _____**