STAT 501 **Assignment 2** **NAME _____**
Spring 2005

Reading Assignment: Chapter 5, and Sections 6.1-6.3 in Johnson & Wichern.

Written Assignment: Due Monday, February 14, in class. You should be able to do the first four problems without a computer, but use computer packages in any way you desire to answer these questions.

1. The following data consist of measurement made on the levels of three liver enzymes (U/L): aspartate aminotransferase (X_1), alanine aminotransferase (X_2), and glutamate dehydrogenase (X_3) in n=10 patients diagnosed with aggressive chronic hepatitis.

Patient	X_1	X_2	X_3
1	31	63	4
2	32	56	6
3	50	59	9
4	56	72	7
5	39	87	9
6	46	95	8
7	29	57	5
8	40	50	3
9	29	44	4
10	24	42	3

These are part of a larger set of data reported by Plomteux (1980, Clin. Chem. 26, 1897-1899). Assuming these data were sampled from a bivariate normal population, evaluate the following quantities.

(a) Compute the maximum likelihood estimates for the mean vector $\underset{\sim}{\mu}' = (\mu_1, \mu_2)$ and the covariance matrix Σ.

$$\underset{\sim}{\bar{X}} = \begin{bmatrix} \\ \\ \end{bmatrix} \qquad \hat{\Sigma} = \begin{bmatrix} \\ \\ \end{bmatrix}$$

(b) Compute the unbiased estimate of the covariance matrix

$$S = \begin{bmatrix} \\ \\ \end{bmatrix}$$

(c) The maximum likelihood estimate of the correlation between X_1 and X_2 and the t-statistic for testing H_0: $\rho = 0$ are

r = _____ t = _____ df = _____

(d) Use the Fisher z-transformation to construct an approximate 95% confidence interval for ρ.

(e) The generalized sample variance is $|S|$ = _____.

(f) The total sample variance is _____.

(g) Use the Fisher z-transformation to construct an approximate 95% confidence interval for the correlation between X_1 and X_3.

lower limit = _____ upper limit = _____

(h) Use the Fisher z-transformation to construct an approximate 95% confidence interval for $\rho_{13\cdot 2}$

lower limit = _____ upper limit = _____

(i) In one sentence, how would you interpret the estimate of $\rho_{13\cdot 2}$ for these data?

(j) Test the null hypothesis H_0: $\rho_{13\cdot 2} = 0$ against the alternative H_A: $\rho_{13\cdot 2} > 0$. Report

t = _____ d.f. = _____ p-value = _____

State your conclusion.

2. Consider a random sample $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_n$ from a p-dimensional normal population with mean vector $\underset{\sim}{\mu}$ and covariance matrix Σ. The purpose of the problem is to construct the likelihood ratio test of the null hypothesis that the p attributes (or components of X) have the same variance σ^2 and are independent, that is, $\Sigma = \sigma^2 I_{p\times p}$.

(a) Write down the formula for the natural logarithm of the joint likelihood function, when the null hypothesis is true, by substituting $\sigma^2 I$ for

$\ell(\underset{\sim}{\mu}, \sigma^2) =$

(b) Give formulas for the m.ℓ.e.'s for $\underset{\sim}{\mu}$ and σ^2.

(c) For testing the null hypothesis $H_0: \Sigma = \sigma^2 I$ against the alternative that Σ is any covariance matrix, find a formula for $-2\log(\Lambda)$ in terms of p, n, $\overset{\sim}{\overline{X}}$, and the elements of the sample covariance matrix S.

(d) For large n, the statistic in part (c) can be compared to the quantiles of the chi square distribution with d.f. =_____.

(e) Use the statistic from Part (c) to test $H_0: \Sigma = \sigma^2 I$ for the data in Problem 1. In this case, ignore the possibility that n may be too small to accurately use the large sample chi-square approximation for the null distribution of $-2\log(\Lambda)$.

$-2\log(\Lambda)$ = _____ d.f. = _____ p-value = _____.

(f) Note that a test statistic that more nearly has a central chi-squared distribution when $H_0: \Sigma = \sigma^2 I$ is true is obtained by multiplying the statistic in part (c) by $[1-(2p^2 + p + 2)/(6pn)]$. Such multipliers are called Bartlett corrections. See Anderson, An Introduction to Multivariate Statistical Analysis, 2nd edition, Section 10.7. Does using this correction factor change the conclusion you reached in part (e)?

3. Neither the generalized variance $|\Sigma|$ nor the total variance trace (Σ) retain all of the information in a covariance matrix. Different covariance matrices can have the same value of the generalized variance or the same value of the total variance. To illustrate this, compute the correlation coefficient, generalized variance, and total variance for each of the following covariance matrices.

$$\begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix} \quad \begin{pmatrix} 5 & -1 \\ -1 & 2 \end{pmatrix} \quad \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} \quad \begin{pmatrix} 8 & 3.2 \\ 3.2 & 2 \end{pmatrix} \quad \begin{pmatrix} 8 & 4 \\ 4 & 2 \end{pmatrix}$$

Correlation	_____	_____	_____	_____	_____
Generalized variance	_____	_____	_____	_____	_____
Total variance	_____	_____	_____	_____	_____

4. Independent samples of sizes $n_1 = 45$ and $n_2 = 55$ were taken from populations of Wisconsin homeowners with and without air conditioning, respectively. Two measurements of electrical usage (in kilowatt hours) were made on each home: X_1, a measure of total on-peak consumption during July 1977, and X_2, a measure of total off-peak consumption during July 1977. The resulting summary statistics are:

Homes with air conditioning	$\underset{\sim}{\overline{X}}_1 = \begin{bmatrix} 204.4 \\ 556.6 \end{bmatrix}$	$S_1 = \begin{bmatrix} 13825.3 & 23823.4 \\ 23823.4 & 73107.4 \end{bmatrix}$	$n_1 = 45$
Homes without air conditioning	$\underset{\sim}{\overline{X}}_2 = \begin{bmatrix} 130.0 \\ 355.0 \end{bmatrix}$	$S_2 = \begin{bmatrix} 8632.0 & 19616.7 \\ 19616.7 & 55964.5 \end{bmatrix}$	$n_2 = 55$

(a) Assuming that the population covariance matrices, Σ_1 (for homes with air conditioning) and Σ_2 for homes without air conditioning), are the same, obtain the pooled estimate of the common covariance matrix. Report

$$S = \begin{bmatrix} & \\ & \\ & \end{bmatrix} \qquad \text{with d.f.} = \_\_\_\_\_$$

(b) Evaluate Bartlett's test of the null hypothesis $H_0: \Sigma_1 = \Sigma_2$ against the alternative $H_A: \Sigma_2 \neq \Sigma_2$. Report

$M = \_\_\_\_\_$ $C^{-1} = \_\_\_\_\_$ $MC^{-1} = \_\_\_\_\_$

d.f. = _____ _____ $\leq$ p-value $\leq$ _____

State your conclusion.

(c) Test the null hypothesis that the correlations between total on-peak and off-peak usage are the same for homes with and without air conditioning. Present the formula for your test statistic and state your conclusion.

(d) Using the pooled covariance matrix, compute the estimate of the squared Mahalanobis distance between $\underset{\sim}{\overline{X}}_1$ and $\underset{\sim}{\overline{X}}_2$:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} \left(\underset{\sim}{\overline{X}}_1 - \underset{\sim}{\overline{X}}_2 \right)' S^{-1} \left(\underset{\sim}{\overline{X}}_1 - \underset{\sim}{\overline{X}}_2 \right) = \_\_\_\_\_ .$$

(e) Test the hypothesis that homes with air conditioning have the same vector of means for on-peak and off-peak consumption as homes without air conditioning (use T^2 from part (d)).

F = _____ d.f. = _____ p-value = _____

State your conclusion.

(f) Compute the value of a test statistic that would be more appropriate than the statistic in parts (d) and (e) when the covariance matrices are not homogeneous. Present a formula for your test statistic and report the value of your test statistic, degrees of freedom and a p-value. Do you reach the same conclusion as you did in part (e)?

5. Systolic blood pressure measures were made on a sample of 85 subjects (Bland and Altman, 1999). For each subject, simultaneous measurements were made by each of two experienced observers using a sphygmomanometer and a third measurement was made by a semi-automatic blood pressure monitor. The data are posted on the course web page as systolic.OBrien.dat. The data file has one line for each subject and four numbers on each line arranged in the following order.

Subject	Identification number
X1	Systolic blood pressure measurement made by observer 1
X2	Systolic blood pressure measurement made by observer 2
X3	Systolic blood pressure measurement made by semi-automatic monitor

(a) Construct a scatterplot matrix of the sample data for X_1, X_2, X_3 . Are there any obvious outliers?

(b) Report the values of the Shapiro-Wilk statistic for each variable.

	X_1	X_2	X_3
Value of W	_____	_____	_____
p-value	_____	_____	_____

Also examine corresponding univariate normal probability plots. State your conclusions. Keep in mind that the data are discrete because of the limited accuracy of the measuring instruments. Consequently, the joint probability distribution of the three systolic blood pressure measurements cannot exactly be a multivariate normal distribution. We only want to know if the multivariate normal distribution is a good approximation.

(b) Examine the chi-square probability plot. What does it indicate about the fit of a three dimensional normal model?

(c) If you conclude that the distribution of any measurement is not reasonably well modeled by a normal distribution, look for a transformation to improve the fit of the normal model. List the transformations you think should be used for each measurement. Report 'none' if no transformation is required.

	X_1	X_2	X_3
Transformation:			

State any additional comments you would like to make on selecting transformations.

(d) Regardless of your results in part (c), **use the natural logarithm of each variable to complete the rest of this problem.** (This transformation is imposed simply to make it easier to grade this question). Compute the sample covariance matrix S and the sample correlation matrix R and report your results.

(e) Test the null hypothesis that correlations among the natural logarithms of systolic blood pressure measurements for the two expert observers and the semi-automatic monitor are all zero (without assuming homogeneous variances). State your conclusions. What do your results imply about agreement between measurements made by the two experts and the semi-automatic monitor.

(f) Test the null hypothesis that all of the correlations are equal and all of the variances are equal for the natural logarithms of systolic blood pressure measurements for the two expert observers and the semi-automatic monitor. Report the value of your test statistic and the corresponding p-value. State your conclusions.

(g) Examine the estimates of the correlation coefficients. Which population correlations are different? Give some statistical justification for your conclusions. What do your results imply about the reliability of the two experts and the semi-automated monitor?

(h) Examine the estimated variances. Which population variances are different? Give some statistical justification for your conclusions. What do your results imply about the reliability of the two experts and the semi-automated monitor? If the variance for $\log(X_3)$ is significantly larger than the value of the variance for $\log(X_2)$, for example, does that imply that the semi-automatic monitor is less accurate than the second expert.

(i) Use the Hotelling T^2 statistic to test the null hypothesis that the mean systolic blood pressure measurements are the same for the two experts and the semi-automatic monitor. Report the value of you test statistic, degrees of freedom and a p-value. State your conclusion.

(j) Do pairwise tests to determine which means are significantly different. Identify the tests you used and report values of your test statistics and degrees of freedom. State your conclusions.

Some additional problems you might consider are problems 5.1, 5.2, 5.5, 5.7, 5.9, 6.2, 6.4, 6.5, 3.6, 3.8, 3.10, 3.11, 3.12, in the fifth edition of the text. Do not submit answers to these problems. Answers may be distributed with answers to this assignment.