**Reading Assignment**: Johnson & Wichern, Chapters 5 and 6.

**Written Assignment**: Due, Monday, February 17 in class.

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.

<table>
<thead>
<tr>
<th>Patient</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31</td>
<td>63</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
<td>56</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>59</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>56</td>
<td>72</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>39</td>
<td>87</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>46</td>
<td>95</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>29</td>
<td>57</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>40</td>
<td>50</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>29</td>
<td>44</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>24</td>
<td>42</td>
<td>3</td>
</tr>
</tbody>
</table>

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

(a) Compute the maximum likelihood estimates for the mean vector $\mu = (\mu_1, \mu_2, \mu_3)'$ and the covariance matrix $\Sigma$.

(b) Compute the unbiased estimate of the covariance matrix, $S$.

(c) Compute $r_{12}$, maximum likelihood estimate of the correlation between $X_1$ and $X_2$, and the value of the t-statistic for testing $H_0 : \rho_{12} = 0$ versus $H_0 : \rho_{12} \neq 0$. The formula for the statistic is $t = \frac{r_{12} \sqrt{n-2}}{\sqrt{1-r_{12}^2}}$ with n-2 degrees of freedom. Report the p-value for your test and state your conclusion.
(d) Use the Fisher z-transformation to construct an approximate 95% confidence interval for \( \rho_{12} \). The lower and upper limits of the confidence interval are

\[
\left( \frac{e^{2Z_{\text{lower}}} - 1}{e^{2Z_{\text{lower}}} + 1}, \frac{e^{2Z_{\text{upper}}} - 1}{e^{2Z_{\text{upper}}} + 1} \right)
\]

where

\[
Z_{\text{lower}} = \frac{1}{2} \log \left( \frac{1 + r}{1 - r} \right) - z_{\alpha/2} \sqrt{\frac{1}{n - 3}}
\]

and

\[
Z_{\text{upper}} = \frac{1}{2} \log \left( \frac{1 + r}{1 - r} \right) + z_{\alpha/2} \sqrt{\frac{1}{n - 3}}
\]

(e) Compute the generalized sample variance, \( |S| \).

(f) Compute the total sample variance, trace(S).

2. Consider a random sample \( X_1, X_2, ..., X_n \) from a p-dimensional normal population with mean vector \( \mu \) and covariance matrix \( \Sigma \). 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 \( \sigma^2 \) and are uncorrelated, that is, \( \Sigma = \sigma^2 I_p \). This is called a spherical covariance matrix, because contours of constant density are concentric p-dimensional spheres centered at the mean vector.

(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 \( \Sigma \). This joint likelihood is a function of \( p+1 \) parameters, the \( p \) components of the population mean vector \( \mu \) and the common population \( \sigma^2 \).

(b) Give formulas for the m.l.e.'s for \( \mu \) and \( \sigma^2 \), as functions of the components of the sample mean vector \( \bar{X} \) and the sample covariance matrix \( S \).

(c) For testing the null hypothesis \( H_0: \Sigma = \sigma^2 I \) against the alternative that \( \Sigma \) is any covariance matrix, find a formula for \( -2 \log(\Lambda) \) in terms of \( p, n, \bar{X}, \) 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 a chi square distribution. What are the degrees of freedom?

(e) Use the statistic from part (c) to test \( \Sigma = \sigma^2 I \) 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 \( -2 \log(\Lambda) \). Report

\[
-2 \log(\Lambda) = \underline{\quad} \quad \text{d.f.} = \underline{\quad} \quad \text{p-value} = \underline{\quad}.
\]
(f) Note that a test statistic that more nearly has a central chi-squared distribution when $H_0$ is true is obtained by multiplying the statistic in Part (c) by $[1-(2p^2 + p + 2)/(6pn)]$. Such multipliers are called Bartlett corrections and they are obtained by expanding the text statistic on a Taylor series expansion and matching the moments of the leading terms in the expansion to the moments of a chi-square distribution. See Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd edition, Section 10.7. Would using this correction factor in Part (e) affect your conclusion?

3. Neither the generalized variance $|\Sigma|$ nor the total variance trace ($\Sigma$) retains all of the information in a covariance matrix. To illustrate this compute the correlation coefficient, the generalized variance, and total variance for each of the following covariance matrices.

$$\Sigma_1 = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix} \quad \Sigma_2 = \begin{pmatrix} 5 & -1 \\ -1 & 2 \end{pmatrix} \quad \Sigma_3 = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} \quad \Sigma_4 = \begin{pmatrix} 8 & 3.2 \\ 3.2 & 2 \end{pmatrix} \quad \Sigma_5 = \begin{pmatrix} 8 & 4 \\ 4 & 2 \end{pmatrix}$$

(a) With respect to the generalized variance, which populations have the same level of overall variation.

(b) With respect to the total variance, which populations have the same level of overall variation.

4. The College test data in table 5.2 of Johnson & Wichern have been stored in the file `college.dat` in the data folder on the course web page. There is one line of data for each individual. SAS code for performing the computations needed for this problem is posted as `college.sas` in the SAS code folder, and R-code is posted as `college.R` in the R code folder. Do these data appear to have been sampled from a tri-variate normal distribution?

(a) To begin to address this question, report the value of the Shapiro-Wilk test statistics for each of the three test scores and report the related p-value, i.e., report

```
Social Science ($X_1$): $W =$ \hspace{1cm} p-value =

Verbal Ability ($X_2$): $W =$ \hspace{1cm} p-value =

Science ($X_3$): $W =$ \hspace{1cm} p-value =
```

State your conclusion.

(b) Make a chi-square probability plot for these data (do not submit the plot) and compute the value of the squared correlation test for the points on the plot. Using simulation, compute an approximate p-value for this test of multivariate normality. Report the value of the test statistic, the approximate p-value, and state your conclusion.
(c) Construct a scatterplot matrix for $X_1$, $X_2$, and $X_3$. What does this scatterplot matrix reveal?

1. Partition the data into subsets using the intervals of the verbal ($X_2$) scores shown below and report an estimate of the correlation between social science ($X_1$) and science ($X_3$) scores within each interval:

   Intervals of verbal scores  
   (25-34)  (35-44)  (45-54)  (55-64)  (65-75)

   Considering that there are approximations to partial correlations between social science and science scores conditioning on verbal scores, what pattern should these estimated correlations exhibit if the data were actually sampled from a tri-variate normal distribution?

2. Assuming multivariate normality, estimate the mean vector and covariance matrix of the condition distribution of $(X_1, X_3)$ given $X_2 = 50$. Use the estimated covariance matrix for this conditional distribution to estimate the partial correlation between social science ($X_1$) and science ($X_3$) scores controlling for (or conditioning on) verbal ability $X_2$. How does this partial correlation compare to the marginal correlation between social science ($X_1$) and science ($X_3$) scores? Is this what you expected?

5. Independent samples of sizes $n_1 = 45$ and $n_2 = 55$ were taken from populations of Wisconsin homeowners, those with air conditioning installed in their homes and those without air conditioning installed in their homes. 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. Summary statistics are shown in the following table.

<table>
<thead>
<tr>
<th>Homes with air conditioning</th>
<th>$\overline{X}_1 = \begin{bmatrix} 204.4 \ 556.6 \end{bmatrix}$</th>
<th>$S_1 = \begin{bmatrix} 13825.3 &amp; 23823.4 \ 23823.4 &amp; 73107.4 \end{bmatrix}$</th>
<th>$n_1 = 45$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homes without air conditioning</td>
<td>$\overline{X}_2 = \begin{bmatrix} 130.0 \ 355.0 \end{bmatrix}$</td>
<td>$S_2 = \begin{bmatrix} 8632.0 &amp; 19616.7 \ 19616.7 &amp; 55964.5 \end{bmatrix}$</td>
<td>$n_2 = 55$</td>
</tr>
</tbody>
</table>
(a) Assuming that the population covariance matrices, $\Sigma_1$ for homes with air conditioning and $\Sigma_2$ for homes without air conditioning, are the same, obtain the spooled estimate of the common covariance matrix. Also report the associated degrees of freedom.

(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 values for $M$, $C^{-1}$, $MC^{-1}$ and report the corresponding df and the p-value. 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. (We did not discuss this in lecture, but you can use anything you learned in this course or on this assignment).

(d) Using the pooled covariance matrix, compute the estimate of the squared Mahalanobis distance between $\overline{X}_1$ and $\overline{X}_2$, i.e., evaluate

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} \left( \overline{X}_1 - \overline{X}_2 \right) S^{-1} \left( \overline{X}_1 - \overline{X}_2 \right)$$

(e) Test the null 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 the $T^2$ value from part (d)). Report the value of the related F-statistic, its degrees of freedom, and the associated p-value. State your conclusion.

(f) Compute the value of a test statistic that would be more appropriate than the statistic in part (d) when the covariance matrices are not homogeneous. Present a formula for your test statistic and report the value of the test statistic and its p-value. State your conclusion.

6. This problem is concerned with the identification of forged bank notes. The file bnotes.dat, available from the data folder on the course web page, has data from six measurements which roughly quantify the size and the position of the printed image on 1000-franc Swiss bank notes. Many other traits could be considered, but these six measurements are easily measured with automatic scanning equipment. The file contains data for a sample of 100 genuine 1000-franc bank notes. (Flury and Riedwyl, 1987). There is one line of data for each bank note in the sample with the data arranged in the following order:

<table>
<thead>
<tr>
<th>Note</th>
<th>Identification number</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>Length of the bill (mm)</td>
</tr>
<tr>
<td>X2</td>
<td>Width of the bill on the left side (mm)</td>
</tr>
<tr>
<td>X3</td>
<td>Width of the bill on the right side (mm)</td>
</tr>
<tr>
<td>X4</td>
<td>Width of the margin at the bottom (mm)</td>
</tr>
<tr>
<td>X5</td>
<td>Width of the margin at the top (mm)</td>
</tr>
</tbody>
</table>
It is relatively easy to obtain a sample from the population of genuine notes, but it would be difficult to sample from the population of forged notes for obvious reasons. In this case information is available only for the genuine notes, and no information is available for forged notes. The identification problem consists of using the data available from the sample of genuine notes to develop a procedure for deciding whether notes of uncertain origin are genuine or forged. We will first try to determine if the variation in the six measurements on genuine notes can be approximately described by a normal distribution. If so, we will use the information in the sample of 100 genuine notes to determine whether or not a “suspect” note was likely to come from the population of genuine notes. The files bnotes.sas and bnotes.R, also available from the course web page, contain SAS and R code, respectively, for entering the data and completing some of the requested computations.

(a) Report the values of the Shapiro-Wilk statistic and associated p-values for each of the six variables. 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 probability distribution of the six measurements on genuine notes cannot be exactly multivariate normal. 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 suitability of a 6-dimensional normal model?

(c) If you conclude that the distribution of any measurement is not reasonably well modeled by a normal distribution, examine the possibility of finding transformations 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.

(d) Identification analysis: Suppose the measurements $\sim 0 \sim = (X_{01}, X_{02}, \ldots, X_{06})'$ are made on a note of uncertain origin. Under the null hypothesis that a vector $\sim 0 \sim$ is a random observation from the population of genuine bills, give formulas for the mean vector and covariance matrix for $\sim 0 \sim - \sim \overline{X}$, where $\sim \overline{X}$ is the sample mean vector for a sample of $n$ genuine bills. Assume $\sim 0 \sim$ is sampled independently of $\sim \overline{X}$. You formulas should be expressed in term of $\sim \overline{X}$, the mean vector for the population of genuine notes, and $\Sigma$, the covariance matrix of the six measurements for the population of genuine notes.

$$E(\sim 0 \sim - \sim \overline{X}) = $$
$$V(\sim 0 \sim - \sim \overline{X}) = $$
(e) Using your formula for \( V(X_0 - \overline{X}) \) from part (d), present a formula for \( d^2 \), the squared Mahalanobis distance of \( \sim X_0 \) from \( \sim \overline{X} \). Assuming normality and assuming that \( \sim X_0 \) is a set of measurements on a genuine note, \( d^2 \) is approximately distributed as a chi-square random variable with 6 d.f.

(f) The measurements on a note of unknown origin are

\[
\sim X_0 = (214.9, 130.5, 130.2, 8.4, 11.6, 138.4)'.
\]

Would you conclude that the bill is genuine or forged? Explain.

Some additional problems you might consider are problems 3.8, 3.10, 3.11, 3.12, 4.14, 4.15, 4.16, 5.1, 5.2, 5.5, 5.7, 5.9 in Johnson & Wichern. Do not submit answers for these problems. Some solutions will be distributed later.