Family Name: *Given Name:* *Student Number:*

UNIVERSITY OF TORONTO
Faculty of Arts and Science

DECEMBER EXAMINATIONS 2009

STA 437H1 F (plus STA 1005)

Duration - 3 hours

1	
2	
3	
4	
5	
6	
T	

No books, notes, or calculators are allowed.

The six questions are worth equal amounts.

Answer in the space provided; if you run out, use the back of a page (and point to where).

Except as noted, when the answer is a number, you must provide an actual number (eg, 1.5 or 3/2), not just a formula that could be evaluated to give this number.

Except as noted, you must explain how you obtained your answer to obtain full credit.

The following formulas may (or may not) be useful

Covariance of transformed random vector: $\text{Cov}(\mathbf{CX}) = \mathbf{C\Sigma_X C}'$

Probability density function for multivariate normal:

$$f(\mathbf{x}) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp(-(\mathbf{x}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)/2)$$

Conditional mean and covariance for multivariate normal:

$$\text{Mean of } \mathbf{X}_1 \text{ given } \mathbf{X}_2 = \mathbf{x}_2 \quad = \quad \mu_1 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \mu_2)$$

$$\text{Covariance of } \mathbf{X}_1 \text{ given } \mathbf{X}_2 = \mathbf{x}_2 \quad = \quad \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$$

T^2 statistic for one sample: $T^2 = n(\overline{\mathbf{X}} - \mu_0)'\mathbf{S}^{-1}(\overline{\mathbf{X}} - \mu_0)$

The distribution of T^2 under the null hypothesis is $[(n-1)p/(n-p)]F_{p,n-p}$, which is approximately χ^2_p when $n-p$ and n/p are both large.

T^2 statistic for two samples, using pooled covariance estimate:

$$T^2 = ((\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) - \delta_0)' \left[(1/n_1 + 1/n_2)\,\mathbf{S}_{\text{pooled}}\right]^{-1} ((\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) - \delta_0)$$

Here, $\mathbf{S}_{\text{pooled}} = ((n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2) \,/\, (n_1+n_2-2)$. The distribution of T^2 is $[(n_1+n_2-2)p/(n_1+n_2-p-1)]F_{p,n_1+n_2-p-1}$ under the null hypothesis that $\mu_1 - \mu_2 = \delta_0$. This distribution is approximately χ^2_p when n_1+n_2-p and $(n_1+n_2)\,/\,p$ are both large.

The factor analysis model: $\mathbf{X} = \mu + \mathbf{LF} + \epsilon$
$\mathbf{F} \sim N(0, \mathbf{I})$ and independently $\epsilon \sim N(0, \mathbf{\Psi})$, with $\mathbf{\Psi}$ diagonal.

1. Indicate whether or not each of the following statements is true. No explanation is required.

   (a) Every non-zero vector of length k is an eigenvector of the matrix $5I$, where I is the $k \times k$ identity matrix.

   (b) If a symmetric matrix is positive definite, then its determinant will be positive.

   (c) If the trace of a symmetric matrix is positive, then the matrix must be positive definite.

   (d) If the random vector X has mean μ_X and covariance matrix Σ_X, and the random vector Y has mean μ_Y and covariance matrix Σ_Y, then $X - Y$ will have mean $\mu_X - \mu_Y$ and covariance matrix $\Sigma_X - \Sigma_Y$.

   (e) If Hotelling's T^2 test for of $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ gives a p-value of 0.7, we can safely conclude that μ is more likely to be zero than it is to be non-zero (assuming the assumptions of independence and normality hold).

   (f) If H_0 is true, there is a 90% chance that the p-value for a test of H_0 versus some alternative, H_1, will be less than 0.9, regardless of how large or small a sample the test is based on.

   (g) The multivariate confidence intervals based on the T^2 statistic will be valid even if the observations are not independent, as long as the observations have a multivariate normal distribution.

   (h) Suppose a factor analysis model with one common factor is fit by maximum likelihood (without rescaling variables). If the estimates of the specific variances (uniquenesses), ψ_i, are all equal, then the factor loadings found will be equal to the first principal component of the covariance matrix times some scalar.

   (i) When factor analysis is performed on observations of p variables, if the variables are actually independent, the factor analysis model will need to have p common factors in order to fit the data well.

   (j) Suppose we are classifying observations as coming from either population A or population B. If in both populations the distribution is multivariate normal, and the eigenvectors of the covariance matrix for population A are the same as the eigenvectors of the covariance matrix for population B, then the boundary between observations that should be classified in population A versus the population B will be a hyperplane.

2. We measure the performance in four athletic events of a sample of 100 male college students registered in a physical education programme. The purpose is to see what is the average level of athletic performance in this population, and to see how performance in one event relates to performance in the other events.

   The measurements are as follows:

   x_{i1} Time in seconds for student i to run 400 metres.

   x_{i2} Time in seconds for student i to run 800 metres.

   x_{i3} Distance in metres that student i throws the shotput.

   x_{i4} Distance in metres that student i throws the javelin.

   Each student does all of these things on one day, in the order above, and is given an hour to rest after each event before doing the next. The measurements are done on five consecutive days, 20 students per day.

   Suppose the sample mean vector computed from the data is $\bar{x} = [54\ 120\ 17\ 60]'$. The sample covariance matrix computed from the data is as follows:

$$S = \begin{bmatrix} 121 & 110 & -6 & 10 \\ 110 & 400 & -24 & 27 \\ -6 & -24 & 36 & 3 \\ 10 & 27 & 3 & 25 \end{bmatrix}$$

   (a) List some reasons that these measurements might not constitute a random sample from the distribution of interest, and comment on how important you think each possible problem is likely to be.

(b) Suppose that the measurements for the 400m and 800m runs were converted so that the time was in minutes rather than seconds. What would the sample mean vector and sample covariance matrix be for the data after this conversion?

(c) What is the sample correlation between the measured time for the 400m run and the measured time for the 800m run?

(d) Suppose that the time measurements for the 400m and 800m runs are subject to independent errors with means of zero and standard deviations of 1 second. How would you estimate the correlation of the true times (without measurement error) for the 400m and 800m runs? For this answer, you may give a formula containing numbers only (no symbols), without finding its numerical value.

3. Answer the following questions about matrices.

   (a) A symmetric 3×3 matrix, $\mathbf{A}$, has the following eigenvectors (which are not necessarily of length one) and corresponding eigenvalues:

$$\begin{aligned} \mathbf{e}_1 &= [\ 3\ 0\ 4\ ]', \quad \lambda_1 = 100 \\ \mathbf{e}_2 &= [\ 0\ 1\ 0\ ]', \quad \lambda_2 = 80 \\ \mathbf{e}_3 &= [\ -4\ 0\ 3\ ]', \quad \lambda_3 = 25 \end{aligned}$$

   Write the matrix A below, giving the actual numerical values of its elements:

(b) Prove that if $\mathbf{A}$ has an eigenvector $\mathbf{e}$ with eigenvalue λ, then $\mathbf{e}$ is also an eigenvector of $\mathbf{A}^{-1}$, and find the eigenvalue of $\mathbf{e}$ as an eigenvector of $\mathbf{A}^{-1}$.

(c) Prove that if $\mathbf{A}$ is a symmetric positive definite matrix, then $\mathbf{A}^{-1}$ is also positive definite. You may assume without proof that $\mathbf{A}^{-1}$ exists. Your proof should be based directly on the definition of positive definiteness, and may not use any theorems that mention positive definiteness. You may use any other well-known theorems about matrices, however.

4. A group of researchers wish to investigate differences in development of male and female children, and whether such differences vary according to the social status of the family. They characterize social status in terms of income and education of the parents, which they use to defined two groups, a low status group and a high status group. They wish to measure three aspects of development — verbal, social, and physical — which they do using tests that produce a score for each child between 0 and 100.

   To reduce the influence of variation from one family to another, the researchers looked at twins in which one child was male and the other female. They recruited 100 two-parent families with such twins for their study, of which 50 families were of low status and 50 were of high status. For each family, they tested both of the twins at age 5 years and at age 8 years. For each of the 100 families, the following variables are therefore available (with a name for reference given at the front):

status	Social status (low or high)
MV5	Score on verbal test of the male child at age 5
MS5	Score on social test of the male child at age 5
MP5	Score on physical test of the male child at age 5
MV8	Score on verbal test of the male child at age 8
MS8	Score on social test of the male child at age 8
MP8	Score on physical test of the male child at age 8
FV5	Score on verbal test of the female child at age 5
FS5	Score on social test of the female child at age 5
FP5	Score on physical test of the female child at age 5
FV8	Score on verbal test of the female child at age 8
FS8	Score on social test of the female child at age 8
FP8	Score on physical test of the female child at age 8

   A number of possible questions could be addressed with this data. Here are some:

   A) For both the low-status and the high-status populations that these subjects were drawn from, what are the average levels of verbal, social, and physical development (as measured by scores on the tests used) for male and female children, at age 5 and at age 8. (Note: this is a total of $2 \times 3 \times 2 \times 2 = 24$ questions.)

   B) For both the low-status and the high-status populations, is there a difference between male and female children in the average level of verbal, social, or physical development, either at age 5 or at age 8? (Note: this is a total of $2 \times 3 \times 2 = 12$ questions.)

   C) Do the low status and high status populations differ with respect to the difference between male and female childen in average level of verbal, social, or physical development, either at age 5 or at age 8? (Note: this is a total of $3 \times 2 = 6$ questions.)

   Answer the questions that follow regarding how the researchers should analyse this data, depending on which of the questions above they are interested in. Assume that the researchers have ensured that the data on each family is independent of the data on other families. Also assume that preliminary analysis has not revealed any outliers, and that the 12 measurements above for both the low-status and the high-status subjects appear to have close to a multivariate normal distribution. Mention any aspects of the analysis that would need to be decided on after after examining the data further.

(a) If the researchers are interested only in question (A) on page 7, how should they analyse the data? Note that they wish both an estimate of the averate scores in the population, and an indication of how uncertain these estimates are. Explain your answer.

(b) If the researchers are interested only in question (B) on page 7, how should they analyse the data? Explain your answer.

(c) If the researchers are interested only in question (C) on page 7, how should they analyse the data? Explain your answer.

5. Suppose that we wish to classify items as coming from either population A or population B, based on observing the values of two variables measured for each item. We know that for both populations these measurements have bivariate normal distributions. Suppose also that we know the mean for population A is $\mu_A = [0\ 0]'$ and the mean for population B is $\mu_B = [0\ 1]'$.

   (a) Suppose that the cost of misclassifying an item from population A as being from population B is **the same** as the cost of misclassifying an item from population B as being from population A. Suppose also that the covariance matrices for populations A and B are the same, with

$$\Sigma_A = \Sigma_B = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$$

   What is the optimal method of classifying item $[x_1\ x_2]$? Give a simple, explicit formula for doing the classification. Also, give a simple formula for the conditional probability that the item is from population A given the values of x_1 and x_2.

(b) Suppose that the cost of misclassifying an item from population A as being from population B is **twice as large** as the cost of misclassifying an item from population B as being from population A. Suppose also that the covariance matrices for populations A and B are the same, with

$$\Sigma_A = \Sigma_B = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}$$

What is the optimal method of classifying item $[x_1\ x_2]$? Give a simple, explicit formula for doing the classification. Also, give a simple formula for the conditional probability that the item is from population A given the values of x_1 and x_2. (Your formulas may contain constants, like $\sqrt{19}$, that you haven't found the numerical values of.)

6. Consider a factor analysis model for six variables, $X = [\, X_1 \; X_2 \; X_3 \; X_4 \; X_5 \; X_6 \,]'$, using two common factors, $F = [\, F_1 \; F_2 \,]'$. Suppose that the mean vector is

$$\mu \;=\; \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{bmatrix}$$

the factor loadings matrix is

$$L \;=\; \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

and the diagonal matrix of specific variances (uniquenesses) is

$$\Psi \;=\; \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 \end{bmatrix}$$

(a) What is the covariance matrix of X?

(b) What is the conditional distribution of X_5 given $X_3 = 2$? Describe the distribution fully.

(c) Give values for the missing elements marked A, B, C, D, and E in the matrix L^* below so that the distribution for X produced by the factor analysis model using L^* as the loadings matrix will be the same as the distribution produced using L above (assume μ and Ψ stay the same). Explain how you obtained your answer.

$$L^* = \begin{bmatrix} 0 & -1 \\ 0 & A \\ -1 & B \\ -1 & C \\ D & 0 \\ E & 0 \end{bmatrix}$$