Multivariate Statistical Analysis Mid Term
November 12, 2010
2 pages, 12 problems, 100 points

1. Briefly describe the difference between a population mean vector and a sample mean vector. (8%)
2. Verify that $\lambda_1 = 8$, $\lambda_2 = 3$ and $\mathbf{e}_1' = [1 \quad 2]$, $\mathbf{e}_2' = [-2 \quad 1]$ are the eigenvalues and eigenvectors of $\begin{bmatrix} 4 & 2 \\ 2 & 7 \end{bmatrix}$. (8%)
3. Compute the major axes and directions formed by the ellipse $(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2$, where $\mathbf{x}' = [x_1 \quad x_2]$, $\boldsymbol{\mu}' = [1 \quad -1]$, $\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 2 \\ 2 & 7 \end{bmatrix}$, and c is a positive real constant. (8%)
4. Find the major axes and directions of the solid ellipse which contains 95% of the probability for bivariate normal distribution $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ have been defined in Problem 3. (8%)
5. Set up a null hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ and an alternative hypothesis $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, with $\boldsymbol{\mu}_0' = [1 \quad -1]$ for bivariate normal distribution $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Show that the Hotelling's T^2 test rejects the null hypothesis at the level of significance $\alpha = 0.05$. Suppose that the sample size is 30, the sample mean $\bar{\mathbf{x}}' = [0.5 \quad 0.5]$, and the sample covariance matrix $\mathbf{S} = \begin{bmatrix} 4 & 2 \\ 2 & 7 \end{bmatrix}$. (8%)
6. Find the major axes and directions of the Hotelling's 95% T^2 confidence region for Problem 5. (8%)
7. Determine the 95% simultaneous T^2 confidence intervals for Problem 5. (8%)
8. Determine the 95% Bonferroni simultaneous confidence intervals for Problem 5. Note that if your *t* corresponding to the desired α can not be found directly in the table, use the following linear interpolation formula to get an approximate value: $t = t_1 + m(\alpha - \alpha_1), \quad m = (t_2 - t_1)/(\alpha_2 - \alpha_1)$, where t_1 and t_2 are the two *t* values corresponding to α_1 and α_2, respectively, with α_1 and α_2 being the two α values in the table and closest to the desired α. (8%)
9. Explain briefly why the lengths of the one-at-a-time confidence intervals are often smaller than those of the corresponding Bonferroni simultaneous confidence intervals. (8%)
10. Verify the Bonferroni inequality,

$$P[\text{all } C_i \text{ true}] = 1 - P[\text{at least one } C_i \text{ false}]$$
$$\geq 1 - \sum_{i=1}^{m} P[C_i \text{ false}] \quad ,$$

Where C_i denote a confidence statement, $i = 1,2,\cdots m$. For this problem, consider only the special case $m = 3$. (8%)

11. At least three kinds of *t*-tests can be applied to data sets – (a) unpaired *t*-tests (two independent populations, also called independent *t*-tests), (b) paired *t*-tests (the treatment group and the control group are paired, also called the dependent *t*-tests), (c) unpaired *t*-tests on the before-and-after difference scores for treatment group and control group (the treatment group and the control group are not paired). The same ideas can also be generalized to multivariate cases using Hotelling's T^2 test. For the following designs, select the most appropriate type of tests, and explain briefly the reason:
    (1) Scores on this test before and after for students in this class review the text book. Suppose that all students did not do any other activities to prepare for the test.(2%)
    (2) Crossover trial, with joint count of patients with rheumatoid arthritis, each of whom undergoes (a) 6 weeks of treatment with a new medicine, and (b) 6 weeks with a placebo. Order is randomized. (2%)
    (3) School performance of only children versus children with one brother or sister. (2%)
    (4) School performance of younger versus older brother/sister in two-child families. (2%)
    (5) School performance of older brother/sister in one-parent versus two-parent families. (2%)
    (6) Average intelligence of older and younger siblings, raised apart and raised together. (2%)
12. Result 3.3 in p. 133 of the text book said that if the sample size is less than or equal to the number of variables, the determinant of the sample covariance matrix will be zero, and the inverse of the sample covariance matrix will not exist. This will make us difficult to use many conventional statistical techniques to infer information about the population. For example, the square of the statistical distance will not be available, and the inference using Hotelling's T^2 test can not be carried out. However, a hot modern research topic in multivariate statistical analysis is to deal with high-dimensional problems, that is, the cases that the number of variables is much larger than the sample size. Do you think it possible to develop new statistical methods to get some information about the population for such problems? Why? (8%)