STAT 530 - Fall 2003 - Homework 3
Due: Monday, October 6th

It is probably easiest to use R for problem 1. You may use either SAS or R for problems 2 and 3.

1) The value of the Hotelling's T^2 statistic is related to both how the means of the two samples relate, and how the variables are related. Consider the following two ways that the means of two samples (on two variables) can be related:

(M1) $\bar{x}_1 = \begin{pmatrix} 5 \\ 8 \end{pmatrix}$ $\bar{x}_2 = \begin{pmatrix} 6 \\ 9 \end{pmatrix}$ This is the **uniform case**, where one population is larger than the other on both variables

(M2) $\bar{x}_1 = \begin{pmatrix} 5 \\ 8 \end{pmatrix}$ $\bar{x}_2 = \begin{pmatrix} 4 \\ 9 \end{pmatrix}$ This is the **split case**, where each population is larger on one of the variables

And the following three ways the two variables can co-vary:

(C1) $C = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ where the variables have **positive covariances** (tend to behave similarly)

(C2) $C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ where the variables are **independent** (have no overlapping information)

(C3) $C = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$ where the variables have **negative covariances** (tend to behave oppositely)

We could use formula 4.5 on page 39 to calculate the T^2 value (let n_1=10 and n_2=10) for the 6 combinations and display them in a table:

Mean \ Covariance	C1 = positive association	C2=independence	C3=negative association
M1 = uniform case	T^2=	T^2=	T^2=
M2 = split case	T^2=	T^2=	T^2=

Since the independence case is the sum of the two separate t-statistics squared, this table will give us a general idea about when the Hotelling's T^2 should have more power (a larger test statistic value and thus smaller p-value) than the separate t-tests and when it should have less power.

Give the filled in table.

Based on the table values, comment on when the T^2 will be relatively more powerful and relatively less powerful. (e.g. "When two-samples differ uniformly from one another the T^2 will be _____ when the variables are positively correlated. On the other hand....")

2) This problem continues problem 3 on homework 2, and again uses the data set
`http://www.stat.sc.edu/~habing/courses/data/bears.txt`.
Use Hotelling's T^2 to test whether there is a difference between the male and female bears on variables 3-7.

a) Report your conclusion at an overall α level of 0.05.

b) Do the assumptions for performing this test seem to be met? Why or why not? (You do not need to perform a test of hypotheses to check them, visual inspection is adequate if you explain what you were looking for.)

c) Briefly compare the p-value you observed in part a of this problem to the p-values you for the individual t-tests found in homework 2. In light of what you found in question 1 is this result surprising? (Recall that the p-value will be large when the t statistics are near zero and vice-versa.)

3) The web page http://www.stat.sc.edu/~habing/courses/data/iris.txt contains a famous data set gathered by E. Anderson and discussed by R.A. Fisher. It concerns the measurements of 150 irises from 3 species. The four variables that the flowers are measured on are: sepal length, sepal width, petal length, and petal width. (The sepal is a leaf in the outer whorl of leaves that protect the flower.)

a) Calculate Mahalinobis distance (page 62-67) between the sample means of the three species. Which two species are most simlar?

b) Conduct a principal components analysis on this data.

(i) Report the loadings and give a descriptive name to each of the components (if it seems interpretable)
(ii) Report the percentage of the variation explained by each of the components
(iii) Plot the irises by their first two principal components, using a different symbol for each type of iris. Does the plot seem to confirm the results in part a?
(iv) How many of the principal components do you think it is worth using? Why?