Questions from the 2008 STA 437/1005 mid-term test

1. A researcher is interested in what type of people go out to movies in Toronto. Each day for a week, he randomly chooses a cinema in Toronto, and then interviews every fourth person who goes into this cinema. He asks each person their age, their income, how long they have lived in Toronto, and number of movies they go out to each year. (For purposes of this question, ignore the possibility that some people may not be willing to answer some or all of these questions.)

   The researcher ends up with data from 1517 interviews. He would like to regard this data as a random sample from people who go to the movies in Toronto (with more weight on people who go more often). Discuss below whatever problems you can think of with this — ie, in what ways might this data not be a suitable random sample?

2. In a psychological experiment, four subjects are tested on how long they take to perform two tasks. The variables X_1 and X_2 are the times in minutes for the two tasks. Here is the data:

X_1	1	2	4	1
X_2	5	1	2	4

   a) Find the sample means of the two variables.

   b) Find the sample covariance matrix for the two variables.

   c) Find the sample mean and sample variance of the difference $X_1 - X_2$, using your answers to parts (a) and (b) — ie, do **not** compute the actual differences for the four observations.

   e) Suppose that the experimenter decides that it would be better to express the times as seconds rather than minutes. Find the new sample means and the new sample covariance matrix, using your answers to parts (a) and (b) — ie, do **not** re-compute the mean and covariance after converting the four observations to seconds.

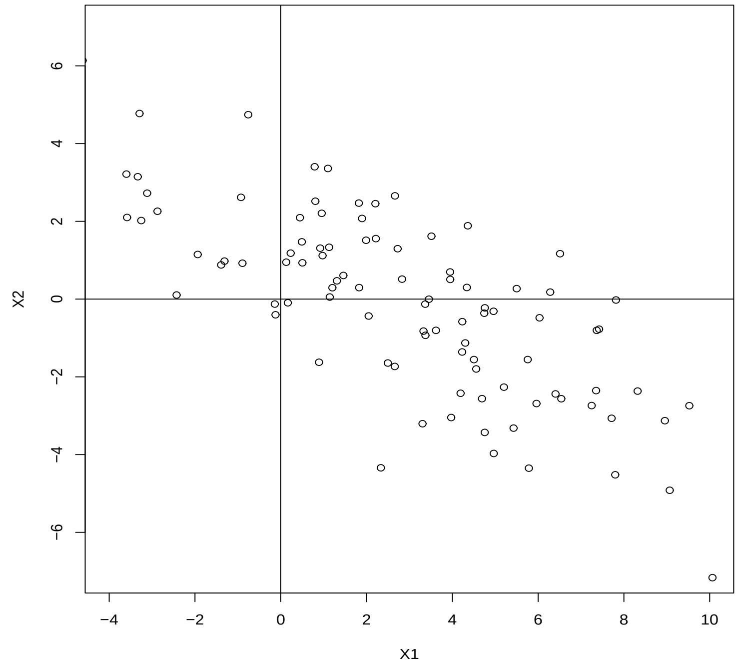
3. This question has three loosely related parts.

   a) Prove that if $\mathbf{A}$ and $\mathbf{B}$ are both $k \times k$ symmetric positive definite matrices, then $\mathbf{C} = \mathbf{A} + \mathbf{B}$ is also a symmetric positive definite matrix.

   b) Prove that if $\mathbf{e}$ is an eigenvector of a $k \times k$ $\mathbf{A}$ with eigenvalue λ, then $\mathbf{e}$ is also an eigenvector of the matrix $\mathbf{A} + \mathbf{I}$, where $\mathbf{I}$ is the $k \times k$ identity matrix, and find the associated eigenvalue.

   c) Suppose that $\mathbf{A}$ is a symmetric $k \times k$ matrix that is not positive definite. Consider the matrix $\mathbf{B} = \mathbf{A} + c\mathbf{I}$, where $\mathbf{I}$ is the $k \times k$ identity matrix and c is some real number. Describe the set of values values for c that will make $\mathbf{B}$ be positive definite, in terms of attributes of the matrix $\mathbf{A}$. Explain your answer.

4. Here is a scatter plot of a sample of $n = 100$ values for two variables:



Judge by eye what the eigenvectors and eigenvalues of the sample covariance matrix for these variables is, choosing one of the following:

A) $\mathbf{e_1} = [\,0.65 \;\; 0.75\,]'$, $\lambda_1 = 41.2$, $\mathbf{e_2} = [\,0.75 \;\; -0.65\,]'$, $\lambda_2 = 0.9$.
B) $\mathbf{e_1} = [\,0.65 \;\; 0.75\,]'$, $\lambda_1 = 17.2$, $\mathbf{e_2} = [\,0.75 \;\; -0.65\,]'$, $\lambda_2 = 1.9$.
C) $\mathbf{e_1} = [\,0.87 \;\; -0.50\,]'$, $\lambda_1 = 3.7$, $\mathbf{e_2} = [\,0.50 \;\; 0.87\,]'$, $\lambda_2 = 0.9$.
D) $\mathbf{e_1} = [\,0.87 \;\; -0.50\,]'$, $\lambda_1 = 18.3$, $\mathbf{e_2} = [\,0.50 \;\; 0.87\,]'$, $\lambda_2 = 1.9$.

Enter your choice here:

Explain your choice below.

5. Let X_1, X_2, X_3, and X_4 be independent random variables that each have the normal distribution with mean zero and variance one. Define the random variables Y_1, Y_2, and Y_3 as follows:

$$\begin{aligned} Y_1 &= (X_1 + X_2)\,/\,2 \\ Y_2 &= X_3 + X_4 \\ Y_3 &= X_1 + X_4 \end{aligned}$$

Let the random vector Y be $[Y_1 \; Y_2 \; Y_3]'$.

a) Find the covariance matrix for Y.
b) Give a full description of the conditional distribution for Y_3 given $Y_1 = -2$ and $Y_2 = 3$.

6. Let $\mathbf{X}_1, \ldots, \mathbf{X}_2$ be independent p-dimensional random vectors, each with the $N_p(\mu, \mathbf{\Sigma})$ distribution, with μ and $\mathbf{\Sigma}$ unknown. Consider the null hypothesis $H_0 : \mu = \mathbf{0}$.

   a) If $p = 9$ and $n = 13$, how big would the T^2 statistic computed from this data have to be for H_0 to be rejected at the $\alpha = 0.05$ level?

   b) If $p = 3000$ and $n = 20000$, and H_0 is true, approximately what value would you expect the T^2 statistic to have? Explain.

7. When we have n observations, each of p variables, we looked at three ways of computing confidence intervals for the means of each of these p variables — intervals from univariate t tests, simultaneous intervals from univariate t tests with Bonferroni correction, and simultaneous intervals from the confidence ellipse found using the T^2 test. The T^2 confidence ellipse can also be used to find simultaneous confidence intervals for any set of linear combinations of the variables. For each of the situations on the next page, discuss which way of computing confidence intervals would be most appropriate.

   (a) A detergent company is trying to find a better detergent. It has $p = 6$ new formulations to test. Each of these is tested on $n = 8$ dirty garments, to see how long it takes to remove grime. For each new formulation, a confidence interval for the time needed is computed. Based on these confidence intervals, the company either selects one of the detergent formulations as best, or it decides that more testing is needed.

   (b) Agricultural researchers have tested the yields of $p = 5$ varieties of wheat for each of the past $n = 12$ years. They would now like to recommend how much of each variety should be planted. Although several factors influence this decision, one important factor is the total expected yield when a certain amount of each variety is planted. They would like to have a confidence interval for this for any set of amounts that they are considering recommending.

   (c) A forensic laboratory tests blood samples for alcohol content. Today, they have $p = 3$ samples from different people to test. They always run the samples through $n = 4$ different machines, so they can be sure of how accurate the results are. The confidence intervals they obtain are used to decide which people are charged with impaired driving.