Family name: *Given names:* *Student ID:*

STA 437/1005 — Mid-term Test — 2008-10-26

For all questions, show enough of your work to indicate how you obtained your answer. No books or notes are allowed. For all questions that have a numerical answer, give your answer as an actual number (eg, 5/4 or 1.25).

The questions are worth equal amounts; they may not be equally difficult.

1	
2	
3	
4	
5	
T	

Here are some of the formulas relating to the material we covered. These formulas might or might not be relevant to some of the questions on the mid-term test.

Sample covariance:

$$\mathbf{S} = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{X}_j - \overline{\mathbf{X}})(\mathbf{X}_j - \overline{\mathbf{X}})'$$

Covariance of a random vector:

$$\mathrm{Cov}(\mathbf{X}) = E[(\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))']$$

Covariance of transformed random vector:

$$\mathrm{Cov}(\mathbf{CX}) = \mathbf{C}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{C}'$$

Spectral decomposition:

$$\mathbf{A} = \lambda_1\mathbf{e}_1\mathbf{e}_1' + \cdots + \lambda_k\mathbf{e}_k\mathbf{e}_k'$$

Probability density function for multivariate normal:

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

Conditional mean and covariance for multivariate normal:

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

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

Hotelling's T^2 statistic:

$$T^2 = n(\overline{\mathbf{X}} - \mu_0)'\mathbf{S}^{-1}(\overline{\mathbf{X}} - \mu_0)$$

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.

1. Consider the following five observations of two variables:

95	210
100	200
105	190
95	195
105	205

a) Find the sample mean vector for this data set.

b) Find the sample covariance matrix for this data. Use the definition in which the divisor is the number of data points minus one.

c) Find the square of the statistical distance of the first data point from the sample mean, using the sample covariance to define the distance.

2. Recall the spectral decomposition theorem: If A is a $k \times k$ symmetric real matrix, it is possible to find a set of k eigenvectors of A that are orthogonal and have length one, and if $e_1, \ldots, e_k$ are any such set of eigenvectors, with eigenvalues $\lambda_1, \ldots, \lambda_k$, then $A = \lambda_1 e_1 e_1' + \cdots + \lambda_k e_k e_k'$.

   a) Use the spectral decomposition theorem to prove that if the eigenvalues of a symmetric real matrix are all positive, then the matrix is positive definite. Your proof must use the spectral decomposition theorem in an essential way. You may use any other well-known theorems about matrices, as long as they don't mention positive definiteness. The only thing about positive definiteness that you may use is its definition.

   b) Use the spectral decomposition theorem to prove that if A and B are $k \times k$ real matrices that are symmetric and positive definite, and A and B have the same eigenvectors (but not necessarily the same eigenvalues), then AB is also symmetric and positive definite. Your proof must use the spectral decomposition theorem in an essential way. You may use any other well-known theorems about matrices, as long as they don't mention positive definiteness. The only things about positive definiteness that you may use are its definition and part (a) of this question, which you can use even if you didn't succeed in proving it.

3. We have n observations on p variables. We are interested in the mean vector, μ, of the distribution from which these observations came. Specifically, we wish to test the null hypothesis, $H_0 : \mu = 0$ versus the alternative hypothesis $H_1 : \mu \neq 0$. We plan to use Hotelling's T^2 test to do this.

   a) Which of the following best describes the issue of independence of the n observations?

      A) The observations must be independent for the T^2 test to be valid. A small amount of dependence might not affect the results much, but any substantial amount of dependence will be a problem, regardless of the values of n and p.

      B) The observations must be independent for the T^2 test to be exactly valid, but if n is large (greater than about 30), the results will be close to correct even if there is a moderate amount of dependence.

      C) The T^2 test is valid regardless of whether or not the observations are independent.

   b) Which of the following best describes the issue of normality of the n observations?

      A) The observations must come from a multivariate normal distribution for the T^2 test to be valid. A small departure from normality might not affect the results much, but any substantial non-normality will be a problem, regardless of the values of n and p.

      B) The observations must come from a multivariate normal distribution for the T^2 test to be exactly valid, but if n is large (greater than about 30), the results will be close to correct even if there is a moderate degree of non-normality.

      C) The T^2 test is valid regardless of whether or not the observations come from a multivariate normal distribution.

   c) Which of the following best describes the distribution of the p-values found using the T^2 test?

      A) The p-value is an unbiased estimate of the probability that H_0 is true, and will converge to the true probability as n goes to infinity.

      B) The p-value is uniformly distributed between 0 and 1 if H_0 is true. If H_1 is true, the distribution of the p-value will instead favour values near zero, more strongly as n increases.

      C) The distribution of p-values will be concentrated below 0.05 if H_1 is true, and above 0.05 if H_0 is true. If H_0 is true, the probability of getting a p-value less than 0.05 goes to zero as n goes to infinity.

d) Which of the following best describes the effect on the T^2 test of changing the units in which variables are measured (ie, rescaling each variable separately), or of performing some other non-singular linear transformation on the observed data vectors (ie, replacing each observed vector x by Ax, for some non-singular matrix A)?

A) Changing the units used, or performing any other non-singular linear transformation, can affect the results of the T^2 test. In order to make the results of the test interpretable, the variables should be standardized to have mean zero and variance one before doing the test.

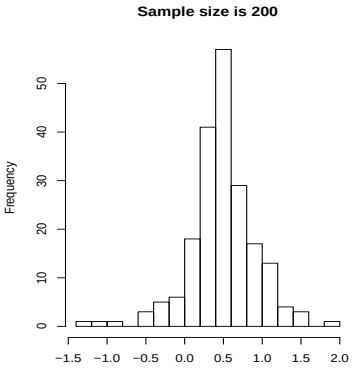
B) Changing the units of measurement will not affect the results of the T^2 test, but other non-singular linear transformations (ie, in which the matrix A is not diagonal) can affect the results.

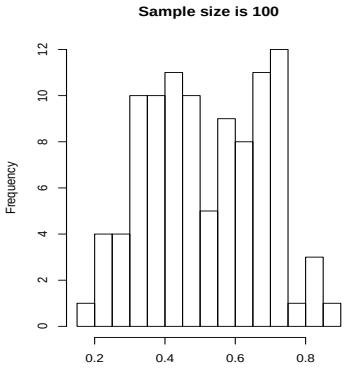
C) The results of the T^2 test are not affected by applying any non-singular linear transformation to the data.

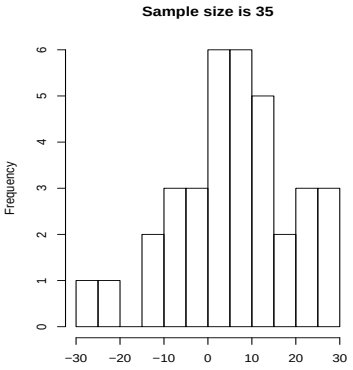
Below, prove that your answer to part (d) above is correct:

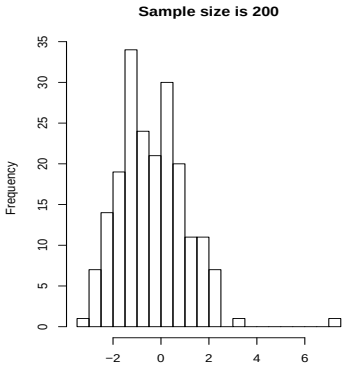
4. Below are histograms from samples for four variables. Below each histogram, write the following:

   - Which of the six normal quantile-quantile plots on the next page is for the same data as is shown in the histogram.
   - Whether there are one or more points (and if so which) that appear to be "outliers", that don't follow the distribution of the rest of the data.
   - Whether there is good reason to think that the variable is not normally distributed (ignoring any outlier points that you identified). If you think the data is not normally distributed, describe in what way it is different from a normal distribution.

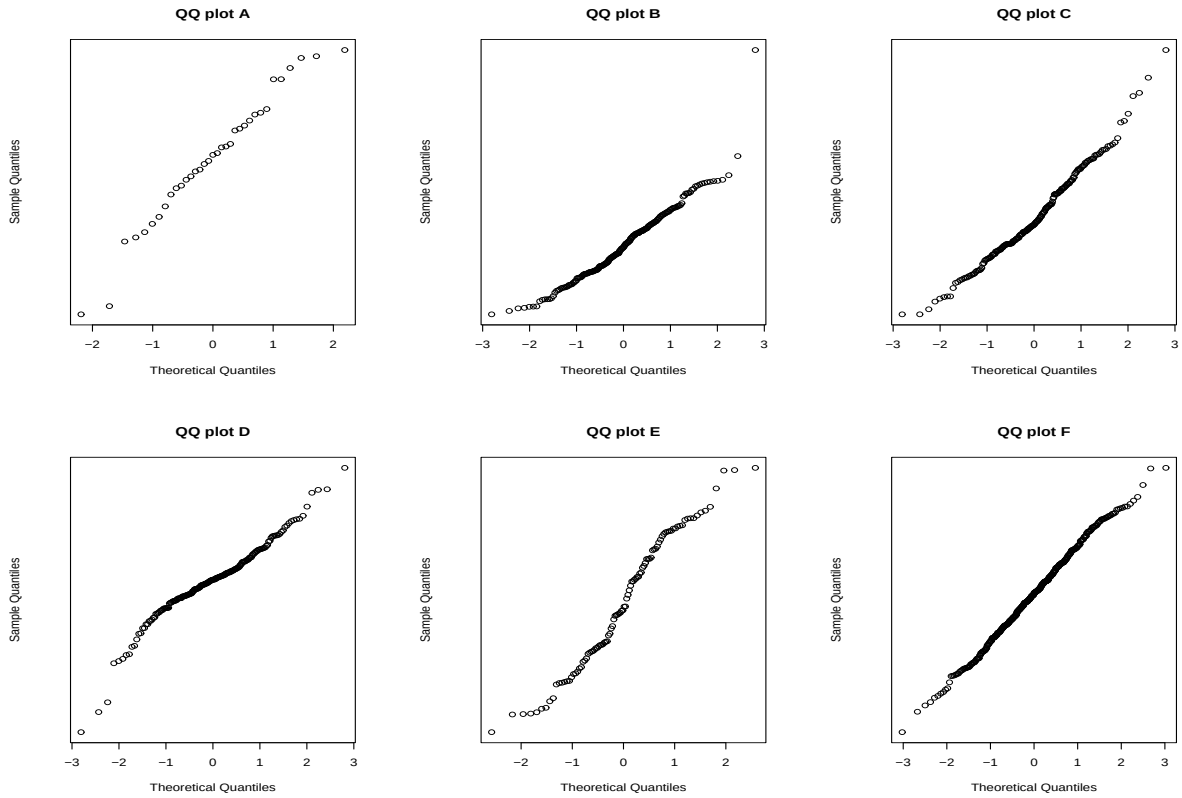




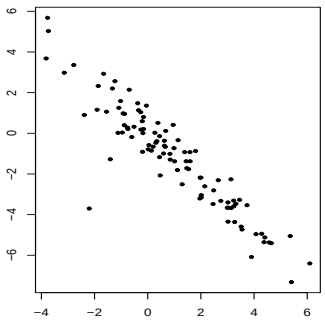


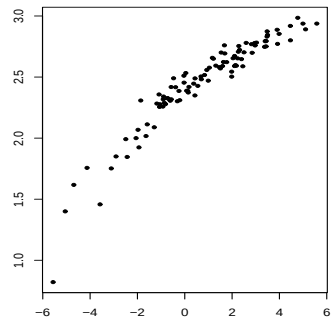


Here are the normal QQ plots to refer to (by letter, A to F) in your answers on the previous page. Note that the scale for the sample quantiles that is normally present on the vertical axis has been omitted here.



Below are scatterplots from two different samples of bivariate data. For each scatterplot, write whether there are one or more points that appear to be outliers (circle these points in the plots), and whether there is good reason to think that the data is not normally distributed, ignoring any outliers that you have identified.





5. Suppose that X_1, X_2, X_3, and X_4 are independent random variables with the $N(0,1)$ distribution. Define the following random variables:

$$\begin{aligned} Y_1 &= X_1 + \epsilon_1 \\ Y_2 &= X_1 + X_2 + \epsilon_2 \\ Y_3 &= X_1 + X_2 + X_3 + \epsilon_3 \\ Y_4 &= X_1 + X_2 + X_3 + X_4 + \epsilon_4 \end{aligned}$$

where ϵ_1, ϵ_2, ϵ_3, and ϵ_4 have the $N(0, 2^2)$ distribution, and are independent of each other and of X_1, X_2, X_3, and X_4.

a) Find the covariance matrix for the vector $[Y_1\ Y_2\ Y_3\ Y_4]'$.

b) Find the variance of $\overline{Y} = (Y_1 + Y_2 + Y_3 + Y_4)/4$.

c) Find the conditional distribution of Y_1 given that $Y_4 = y_4$.