STAT 445/645
Final Examination

Time: 15:30 to 18:30

Student ID#: ________________ **Student Name:** ________________________

Instructions: *Answer all questions in the spaces provided. You are permitted to bring a copy of the text and any personal notes of your choosing. Calculators will also be permitted, but no computers or electronic devices with communication capabilities will be allowed. Except for writing equipment (pencils, eraser, and ruler recommended), the text, personal notes, and a non-communicating calculator, you must leave all materials (including coats, hats, backpacks and other containers) at the front of the room for the duration of the test or examination. Any student exhibiting questionable behaviour may be required to sit the rest of the examination in isolation.*

Take a moment to organize your thoughts before composing an answer. Marks will be deducted for disorganized work, and no credit will be given for collections of possibly relevant remarks that do not show evidence of an ability to work toward a solution to the problem posed.

1. Is the following matrix an orthogonal matrix? Justify your answer.

$$M = \begin{pmatrix} 0.577 & -0.707 & 0.500 \\ 0.577 & 0.000 & -0.707 \\ 0.577 & 0.707 & 0.500 \end{pmatrix}$$

2. Following are eigenvalues and eigenvectors, with minor adjustments, for the correlation matrix for the iris data discussed in class.

Eigenvalues	2.9185	0.9140	0.1479	
Eigenvectors	1	2	3	4
Sepal Length	0.5211	-0.3774	0.7196	0.2613
Sepal Width	-0.2693	-0.9233	-0.2444	-0.1235
Petal Length	0.5804	-0.0245	-0.1421	-0.8014
Petal Width	0.5649	-0.0669	-0.6343	0.5236

a. Evaluate the missing entry in the upper right-hand corner of the above table.

b. What is the usual name for the multivariate analysis technique that is based on such an eigenvalue decomposition?

c. Provide verbal descriptions of what you can learn in terms of the variation in this dataset from the coefficients of the first two eigenvectors. Restrict your answer to at most 3 sentences per vector.

d. How many of these eigenvectors appear to be useful for summarizing the bulk of the variation in this dataset? Construct a relevant plot, and justify your answer using (i) the information in this plot and (ii) some other standard criterion for making such a recommendation.

e. Following is a similar summary, but for the eigenvalues of the variance-covariance matrix. The eigenvalues and eigenvectors are different, but not overwhelmingly so. Why should they be different, but not substantially so? Explain in at most two sentences.

Eigenvalues	4.2282	0.2427	0.0782	0.0238
Eigenvectors	1	2	3	4
Sepal Length	0.3614	0.6566	-0.5820	0.3155
Sepal Width	-0.0845	0.7302	0.5979	-0.3197
Petal Length	0.8567	-0.1734	0.0762	-0.4798
Petal Width	0.3583	-0.0755	0.5458	0.7537

3. Hierarchical clustering methods:

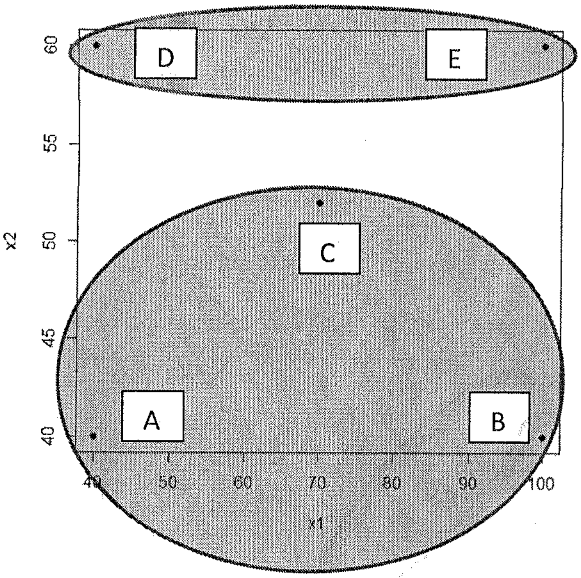
a. Sketch a segment of a dendrogram that contains an inversion.

b. Name one commonly considered hierarchical clustering method that is particularly prone to producing inversions?

c. Explain in words (using at most 4 sentences) how an inversion can be generated when using this method.

4. *K*-means clustering: Following are bivariate observations on five points. They have been tentatively clustered as shown in the accompanying graph. Perform the calculations for deciding whether or not Point C should be moved from its existing cluster to the other one in a *K*-means clustering algorithm with *K* = 2, and draw the appropriate conclusion.

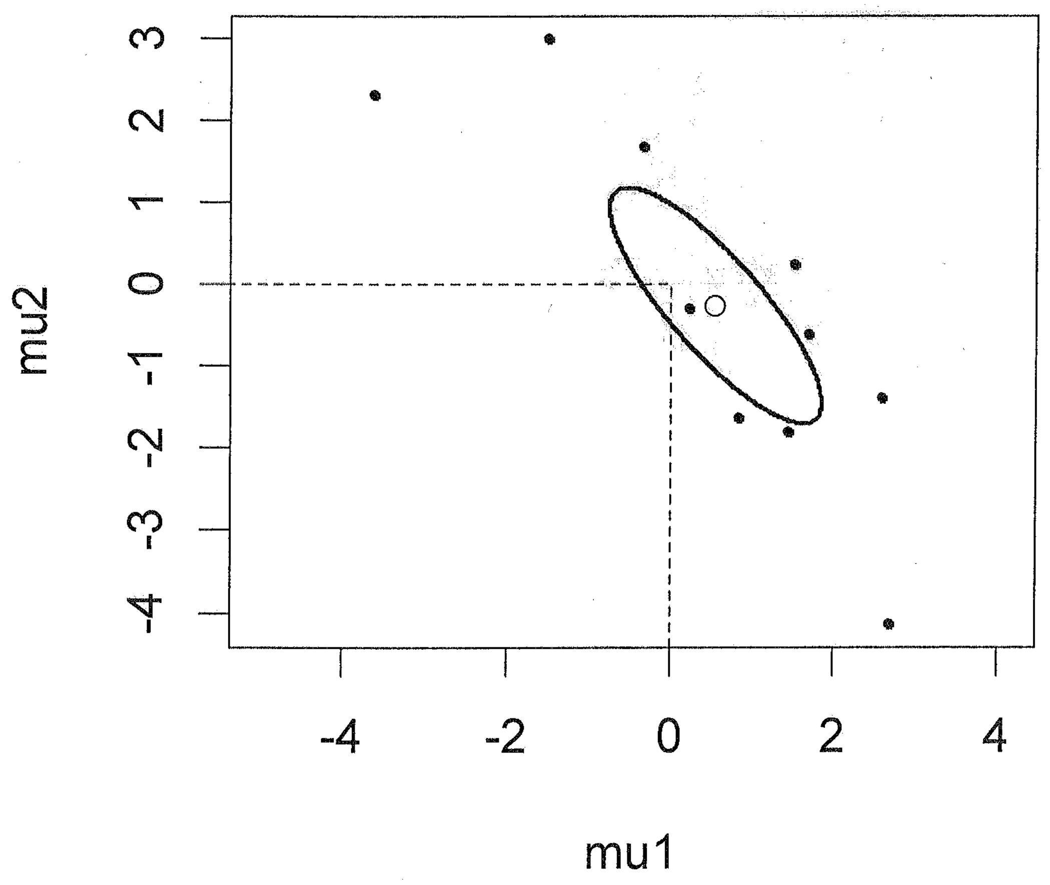
Point	A	B	C	D	E
	40	100	70	40	100
	40	40	52	60	60



5. Following are the mean vector and variance-covariance matrix for a set of $n = 10$ bivariate observations. The accompanying figure also provides a scatterplot of the data along with (i) an open circle indicating the location of the sample mean, and (ii) a 90% confidence ellipse for the underlying mean.

$\overline{x} = \begin{pmatrix} 0.541 \\ -0.255 \end{pmatrix}$, and = $\begin{pmatrix} 3.786 & -3.487 \\ -3.487 & 4.659 \end{pmatrix}$.

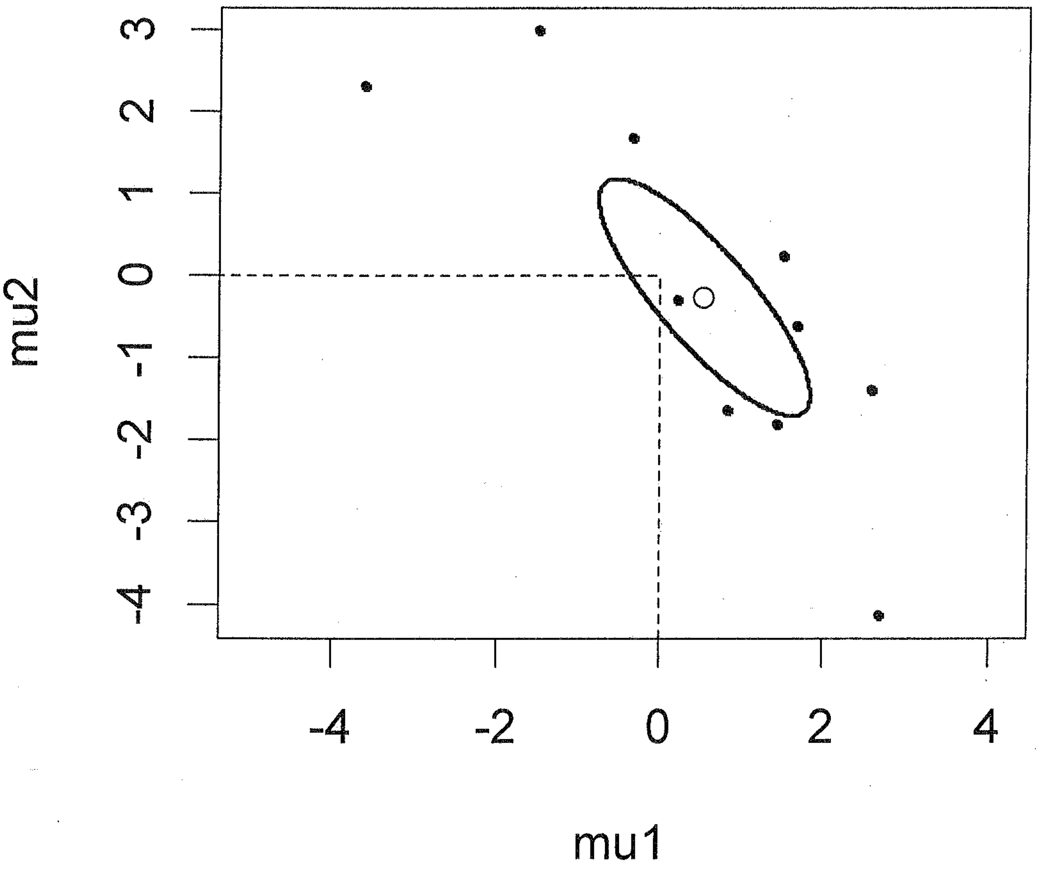
These statistics might not match the graph! Redo both if using again.



a. What does the above confidence ellipse permit you to conclude about the *p*-value for Hotelling`s T^2 test for testing $H_0: \boldsymbol{\mu} = \mathbf{0}$, vs. $H_a: \boldsymbol{\mu} \neq \mathbf{0}$? Explain your reasoning in one sentence.

b. Is it of concern that the 90% confidence ellipse contains only 10% of the data? Explain your reasoning in at most two sentences.

c. Use the graph to construct approximate 90% simultaneous confidence intervals of the Scheffé-type for each of the two univariate means. Try to provide two significant digits of accuracy, but you won't be penalized if your second digit is inaccurate. Focus on describing your method. You can draw on the following copy of the graph if that would help.



d. Find the most appropriate 90% confidence intervals for each of these means using some other method that is appropriate when these are the only two confidence intervals that are to be constructed.

e. How would you adjust the above confidence intervals if you were to adopt Fisher`s approach to simultaneous inferences? (You need not actually do the calculations, just describe how they would differ.)

f. Would, in this particular instance, Fisher's approach provide appropriate protection against either one of these two confidence intervals not containing 0 when in fact both underlying means were 0? Explain your answer in at most 3 sentences.

6. R Code:

   a. What does the following line of R code calculate?

      ```
      hc.complete <- hclust(dist(my.data), method="complete")
      ```

   b. What does the segment, 'method="complete" ' specify?

   c. What does the function, "dist", perform?

d. Name one other option for the 'method', and state how it differs from the option, "complete". Be specific in terms of the calculations that the computer will make.

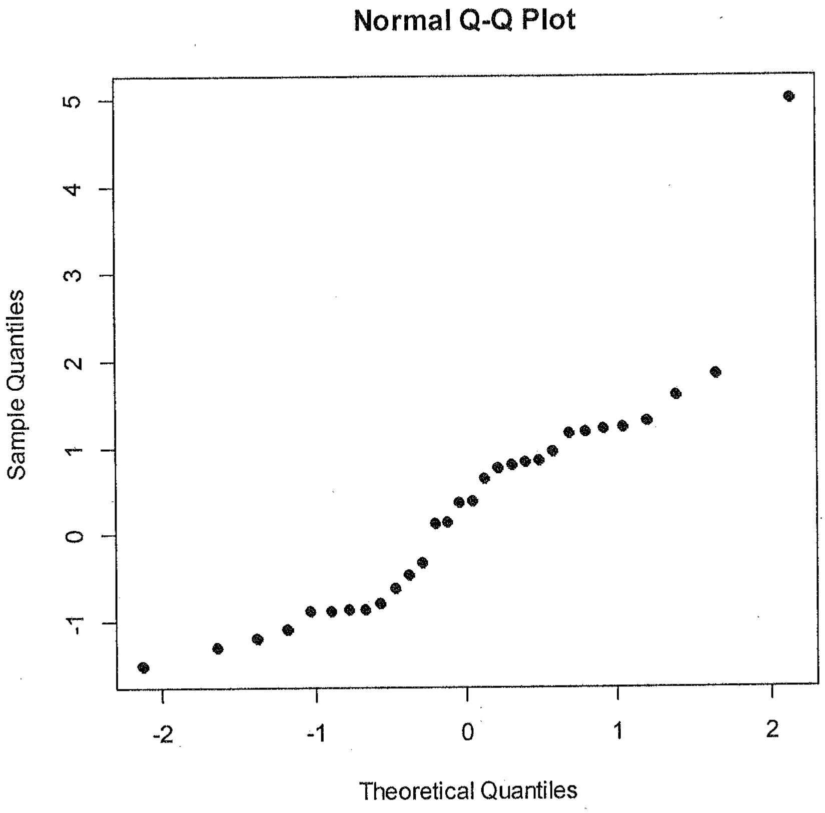
e. Identify and fix the error(s) in the following line of R code for calculating Hotelling's T^2 statistic for testing the null hypothesis that a mean vector is zero from a sample with mean vector, ***xbar***, and sample variance-covariance matrix, ***S***:

```
T.sq <- xbar%*%S%*%xbar
```

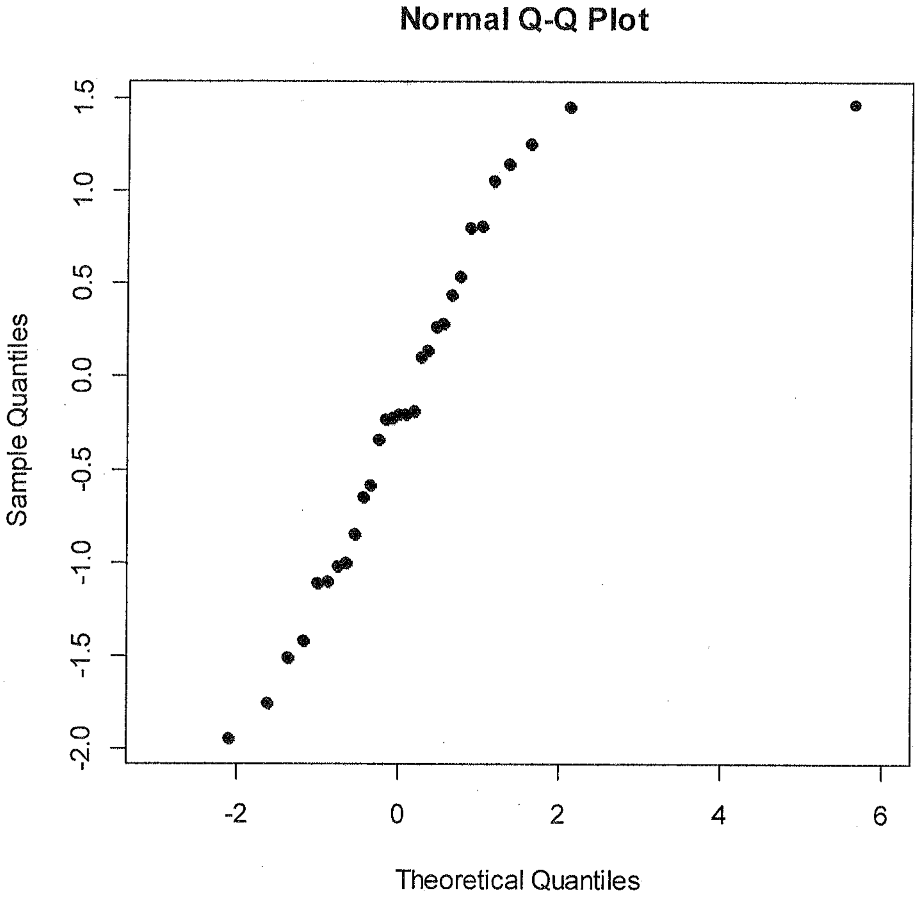
f. Compose R code that will convert a vector of univariate observations, ***x.obs***, to standard units.

7. Do the following Q-Q plots show any evidence of nonnormality? If so, describe the nature of the departure from the normal distribution (e.g., skewed to the right, skewed to the left, an outlier in the left-hand tail, etc.). Also, one or more of them may have an impossible shape for a Q-Q plot. If so, identify it/them.

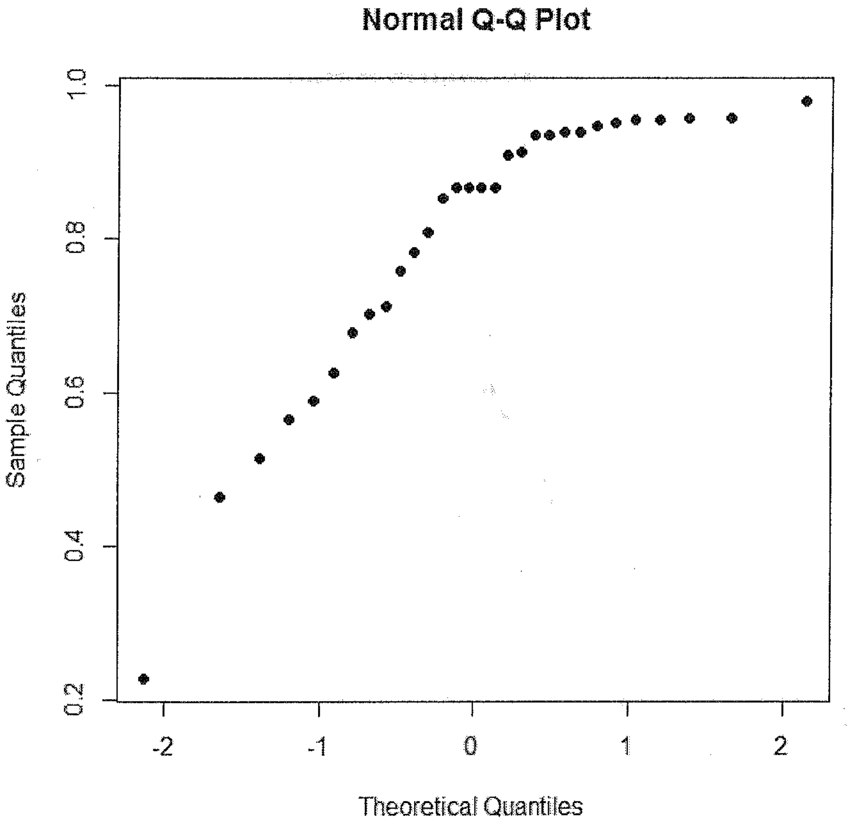
   a.



b.


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2
0
2
4
6

c.


Normal Q-Q Plot
Sample Quantiles
0.2
0.4
0.6
0.8
1.0
Theoretical Quantiles
-2
-1
0
1
2

Extra page for calculations, etc.

Extra page for calculations, etc.

Extra page for calculations, etc.