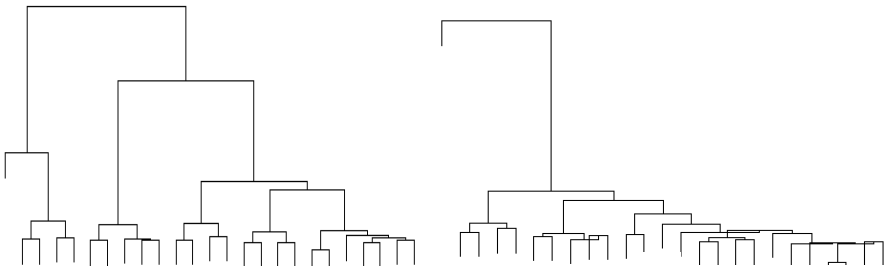
# STAT 445 Practice Questions for Final Examination

1. Sketch a QQ plot for a set of observations that show each of the following types of departures from the normal distribution:
    a. An outlier in the right-hand tail
    b. An outlier in the left-hand tail
    c. Positive skewness
    d. Negative skewness

2. Following are two dendrograms for the same set of items. One was produced by the centroid method; the other, by Ward's method.



    a. Which one do you think was likely produced by the centroid method?
    b. Provide two reasons for your choice.

3. Following are the mean vector and variance-covariance matrix for a set of 50 independent items sampled from the output from an industrial process manufacturing machine bolts. Three measurements were taken on each item: (i) length, (ii) diameter, and (iii) thread pitch.

$$\bar{\boldsymbol{x}} = \begin{pmatrix} 2.218 \\ 0.503 \\ 1.598 \end{pmatrix} \qquad \boldsymbol{S} = 0.001 \times \begin{pmatrix} 1.48 & 0.85 & 0.33 \\ 0.85 & 2.10 & 0.42 \\ 0.33 & 0.42 & 1.40 \end{pmatrix}$$

    a. Does the sample mean vector differ significantly at the 5% level from the target mean of $(2.20, 0.50, 1.60)'$? You can use the following values for the inverse of the variance-covariance matrix if you'd like.

$$\boldsymbol{S}^{-1} = \begin{pmatrix} 895.656 & -340.749 & -108.894 \\ -340.749 & 636.222 & -110.547 \\ -108.894 & -110.547 & 773.118 \end{pmatrix}$$

    *Note that you will not likely be asked to answer such a computationally intensive question on the actual final examination.*
    b. Construct 95% confidence limits for each of the three variables using the most appropriate method if these were the only such confidence intervals that you

were planning to construct from this dataset. (For the purpose of this review, you need only construct the first one.)

c. Someone who is familiar with the text has read that you can justify using ordinary *t*-based confidence intervals in such circumstances if the Hotelling's $T^2$-test rejects the null hypothesis.
    i. To which now-discredited multiple comparison technique in the context of univariate analysis of variance is this strategy related?
    ii. Describe a situation (in the context of the specific application in this question) in which it can provide misleading conclusions.

4. Following is a summary of the results of several applications of the *k*-means clustering method, one for each of several choices of *k*. There were 8 items with coordinates as follows:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $x_1$ | 68 | 66 | 72 | 48 | 35 | 20 | 34 | 42 |
| $x_2$ | 148 | 166 | 155 | 55 | 34 | 14 | 12 | 15 |

a. Fill in the values in the empty cells.
b. Apply one common criterion for selecting a reasonable value for *k*, and draw an appropriate conclusion. Draw a graph if appropriate.

| Number of Clusters | Within-Cluster Sum of Squares | Between-Cluster Sum of Squares |
|---|---|---|
| 1 | 35865.75 | |
| 2 | 1990.13 | 33875.62 |
| 3 | 1839.30 | 33926.45 |
| 4 | 1806.80 | |
| 5 | 557.67 | |
| 6 | 69.00 | |
| 7 | 36.50 | |
| 8 | | |

5. 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.

6. The following table was produced by JMP as part of a correspondence analysis. Several entries have been blanked out.
   a. Several entries have been blanked out. Fill them in.

**Details**

| Singular Value | Inertia | Portion | Cumulative |
|---|---|---|---|
| 0.53253 | | 0.5534 | 0.5534 |
| 0.41244 | | 0.3319 | |
| 0.24246 | | 0.1147 | |

| Type | c1 | c2 | c3 |
|---|---|---|---|
| A | 0.4204 | -0.4396 | -0.0057 |
| B | -0.0080 | 0.3584 | 0.6335 |
| C | -0.5641 | 0.0674 | -0.1075 |
| D | 0.9407 | 0.9369 | -0.2735 |

| Site | c1 | c2 | c3 |
|---|---|---|---|
| P0 | 0.9195 | 0.7522 | -0.2586 |
| P1 | 0.3782 | -0.6114 | -0.0020 |
| P2 | 0.1374 | -0.5859 | -0.1587 |
| P3 | 0.7002 | -0.2210 | 0.3306 |
| P4 | 0.1481 | 0.1351 | 0.4690 |
| P5 | -0.0784 | -0.0865 | -0.1902 |
| P6 | -0.7247 | 0.2122 | -0.0356 |

   b. Based on these values, do you feel that a correspondence analysis plot in two dimensions would provide a reasonable summary of the data? Explain your answer in at most two sentences.
   c. Use these values to draw a rough sketch of the correspondence analysis graph, but include only points for the first two types and first two sites.