STAT 530/J530
November 3rd, 2005

Instructor: Brian Habing
Department of Statistics
LeConte 203
Telephone: 803-777-3578
E-mail: habing@stat.sc.edu

Homework 6

The data set contains four groups of fifteen crabs each: O = orange male, o = orange female, B= blue male, b = blue female.

Each crab has eight measurements. The first four are: FL = frontal lobe size (mm), RW = rear width (mm), CL = carapace length (mm), and CW = carapace width (mm). The next four variables were created by first dividing each of the original measurements by the total body depth and then standardizing.

1) If you want to use the measurements to tell apart the four different groups of crab, why might you want to divide the four individual measurements by the overall size?

2) If you want to use the measurements to tell apart the four different groups of crab apart, why might you want to standardize the measurements?

3) If you decided to use two-dimensional multi-dimensional scaling on the divided and standardized data set, which of Classical or Isometric do you think would be best at showing the distinct clusters in this data set? Why?

4) Perform the scaling method you chose in part 3 and construct a plot of the scaling where each crab is denoted solely by a dot. By looking at the separation of points in the scaling, divide the scaling into separate clusters.

5) Re-plot the scaling in 4, this time labeling each crab by its type. Summarize how well you think these measurements separate the four groups.

6) For the method you chose in 3, compare the estimated stress for the one, two, and three dimensional multi-dimensional scalings. How many dimensions should you have used?

Hotelling's T^2

Hotelling's T^2 tests H_0: $\mu_1 = \mu_2$ where μ_1 and μ_2 are the (qx1) mean vectors.

Assumes:

1. The variables in each population are multivariate normal.
2. The two populations have the same covariance matrix.
3. The observations are independent.

Hotelling's T^2

The formula is:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}\left(\bar{x}_1 - \bar{x}_2\right)^T S^{-1}\left(\bar{x}_1 - \bar{x}_2\right)$$

Where S^{-1} is the pooled covariance estimate.

With proper scaling can be compared to an F distribution.

More Than Two Groups?

In general Multivariate Analysis of Variance (MANOVA) is designed to test: H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$

where the μ_i are the (qx1) mean vectors.

The assumptions are the same as before: independence, multivariate normality, and equal variances.

The Basic Idea

The basic idea is similar to one-way ANOVA. A between-group sum of squares (H) and within-group sum of squares (E) are calculated, and the ratio is taken. If the ratio is large then the null hypothesis is rejected.

The difficulty is that H and E are matrices!

Four Statistics

Wilk's Lamda: $\Lambda = \frac{|E|}{|H+E|}$

Roy's Greatest Root:

Largest Eigen value of $E^{-1}H$

Lawley-Hotelling Trace:

Trace of $E^{-1}H$

Pillai Trace:

Trace of $H(H+E)^{-1}$

Which One???

All result in the same p-value when *k*=2.

Steven, J.P. (2002). *Applied Multivariate Statistics for the Social Sciences*. Mahwah, NJ: Lawrence Erlbaum.

has an extensive overview…

Which One???

Wilk's, Lawley-Hotelling, and Pillai are all fairly robust if the covariances are not equal assuming the sample sizes are fairly equal (largest/smallest < 1.5).

Roy's is most powerful if the differences can be measured using only a single combination of variables.

Pillai's is slightly more powerful than the others if the differences are found on several orthogonal combinations of variables.

About the Assumptions - Independence

The independence assumption is critical. Dependence between observations can result in the α level being several times larger than it should be.

In some cases group averages or hierarchical models can be used to remove the effects of individuals being measured in groups.

About the Assumptions - Normality

Violation of Normality seems to have only a small effect on the α-level.

The tails of the distribution do seem to have a sometimes substantial effect on power.

The skewness seems to have much less effect although there seem to be fewer studies.

About the Assumptions - Covariances

The α-level is maintained fairly well if the sample sizes are equal except for extremely different covariance matrices.

Very unequal sample sizes can amplify even slight differences in the covariances and greatly affect the α-level.

In general large variance in a small group makes the test liberal, large variability in the larger group makes it conservative.

In General

- Univariate and Multivariate tests can disagree
- Don't include too many variables
- Always form the most specific hypotheses you can

Discriminant Analysis

Discriminant Analysis tries to find the linear combinations of variables that do the best job at distinguishing between the groups.