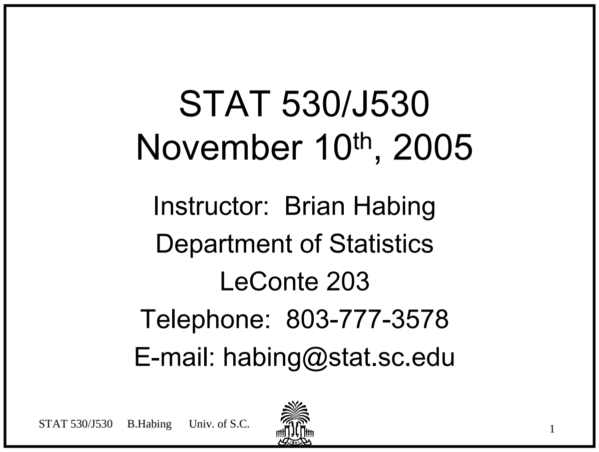
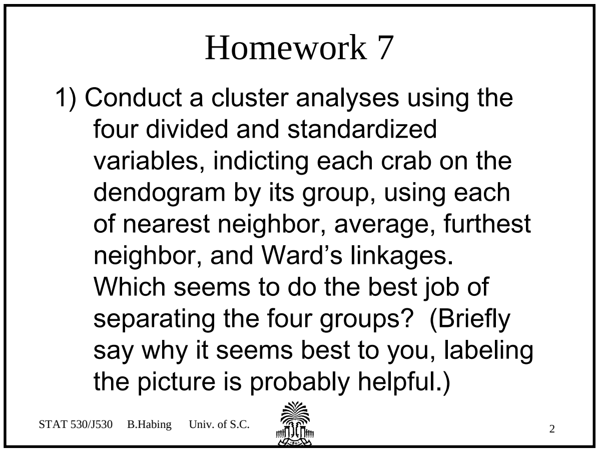


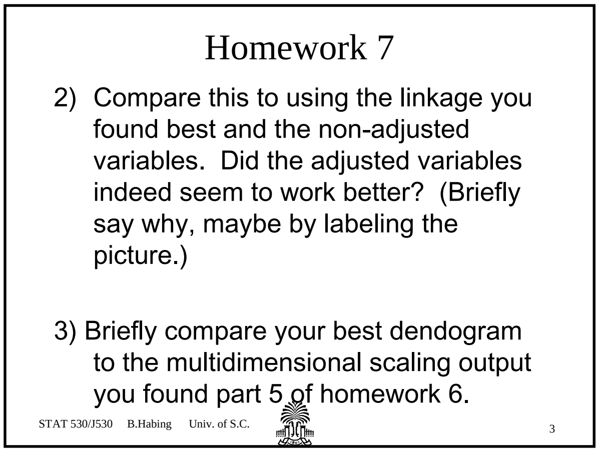
STAT 530/J530
November 10th, 2005

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



Homework 7

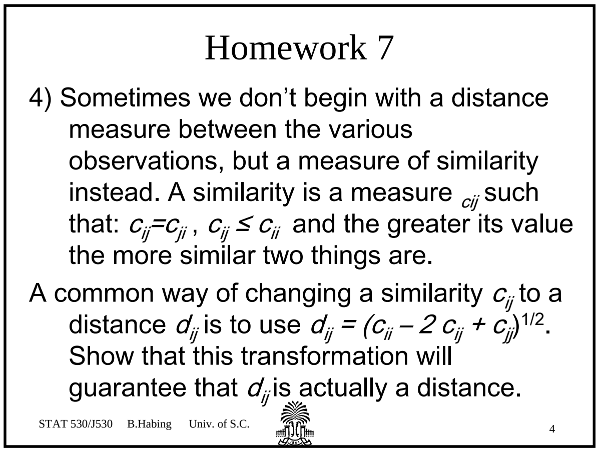
1) Conduct a cluster analyses using the four divided and standardized variables, indicting each crab on the dendogram by its group, using each of nearest neighbor, average, furthest neighbor, and Ward's linkages. Which seems to do the best job of separating the four groups? (Briefly say why it seems best to you, labeling the picture is probably helpful.)



Homework 7

2) Compare this to using the linkage you found best and the non-adjusted variables. Did the adjusted variables indeed seem to work better? (Briefly say why, maybe by labeling the picture.)

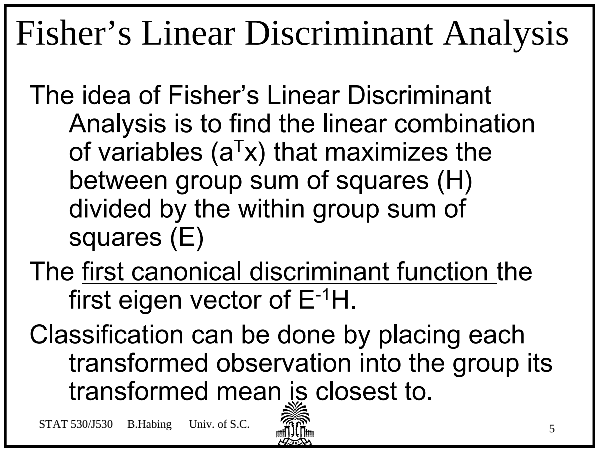
3) Briefly compare your best dendogram to the multidimensional scaling output you found part 5 of homework 6.



Homework 7

4) Sometimes we don't begin with a distance measure between the various observations, but a measure of similarity instead. A similarity is a measure c_{ij} such that: $c_{ij}=c_{ji}$, $c_{ij} \leq c_{ii}$ and the greater its value the more similar two things are.

A common way of changing a similarity c_{ij} to a distance d_{ij} is to use $d_{ij} = (c_{ii} – 2\, c_{ij} + c_{jj})^{1/2}$. Show that this transformation will guarantee that d_{ij} is actually a distance.

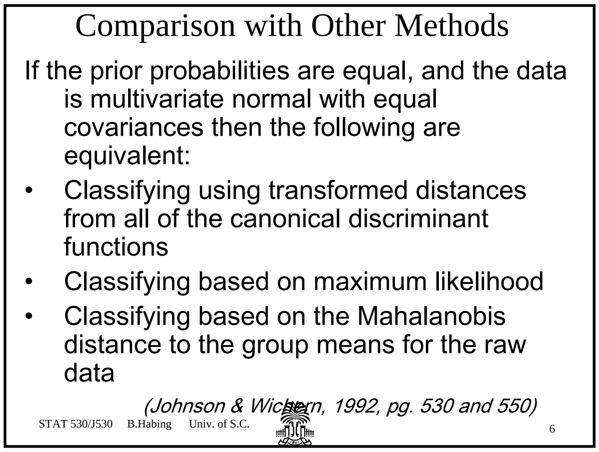


Fisher's Linear Discriminant Analysis

The idea of Fisher's Linear Discriminant Analysis is to find the linear combination of variables ($a^T x$) that maximizes the between group sum of squares (H) divided by the within group sum of squares (E)

The first canonical discriminant function the first eigen vector of $E^{-1}H$.

Classification can be done by placing each transformed observation into the group its transformed mean is closest to.



Comparison with Other Methods

If the prior probabilities are equal, and the data is multivariate normal with equal covariances then the following are equivalent:

- Classifying using transformed distances from all of the canonical discriminant functions
- Classifying based on maximum likelihood
- Classifying based on the Mahalanobis distance to the group means for the raw data

(Johnson & Wichern, 1992, pg. 530 and 550)

Maximum Likelihood

The estimated pdf for observation x given it is in group *i* is:

$$p(\mathrm{x} \mid group\ \ i) = |\, 2\pi S\, |^{-1/2}\ e^{-\frac{1}{2}\left((\mathrm{x}-\bar{\mathrm{x}}_i)^T S^{-1} (\mathrm{x}-\bar{\mathrm{x}}_i)\right)}$$

$$= |\, 2\pi S\, |^{-1/2}\ e^{-\frac{1}{2} D_i^2}$$

Bayes’s Theorem can let us find

$$P(group\, i \mid x) = \frac{p_i P(group\, i \mid x)}{p_1 P(group\, 1 \mid x) + \cdots + p_m P(group\, m \mid x)}$$

Logistic Regression

For two groups we have a binary response variable Y=0 or 1) and predictors x_1, … x_q

$$P\left[Y = 1 \mid \mathrm{x}\right] = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots \beta_q x_q)}}$$

Example

```
PROC LOGISTIC DATA=crabs ;
CLASS CrabType;
OUTPUT OUT=CrabOut PREDPROBS=I;
MODEL CrabType=sFL sRW sCL sCW/ LINK=GLOGIT;
RUN;

PROC PRINT DATA=CrabOut;
RUN;
```