

Discriminant Analysis

Summary

The **Discriminant Analysis** procedure is designed to help distinguish between two or more groups of data based on a set of p observed quantitative variables. It does so by constructing *discriminant functions* that are linear combinations of the variables. The objective of such an analysis is usually one or both of the following:

1. to be able to *describe* observed cases mathematically in a manner that separates them into groups as well as possible.
2. to be able to *classify* new observations as belonging to one or another of the groups.

In constructing the discriminant functions, the procedure allows you to include all of the variables or to use a stepwise selection procedure that includes only those variables that are statistically significant discriminators amongst the groups. Statistical summaries and tests of significance for the number of discriminant functions needed are included.

The derived discriminant functions may be used to classify new cases into groups. Prior probabilities of belonging to each group may be entered or derived from the observed data.

Sample StatFolio: *discriminant.sgp*

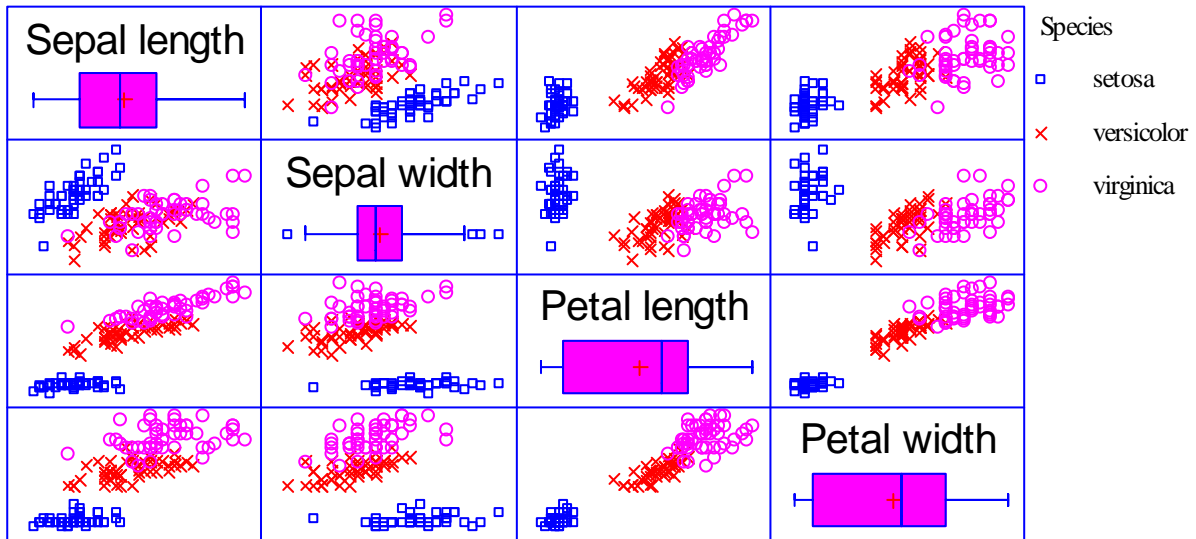
Sample Data:

The file *iris.sgd* contains a famous set of data from Fisher (1936). The data consist of a total of $n = 150$ irises, 50 from each of $g = 3$ different species: *setosa*, *versicolor*, and *virginica*.

Measurements were made on $p = 4$ variables, describing the length and width of the sepal and petal. The table below shows a partial list of the data in that file:

| <i>Sample</i> | <i>Sepal length</i> | <i>Sepal width</i> | <i>Petal length</i> | <i>Petal width</i> | <i>Species</i> |
|---------------|---------------------|--------------------|---------------------|--------------------|----------------|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 2 | 4.9 | 3 | 1.4 | 0.2 | setosa |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 5 | 5 | 3.6 | 1.4 | 0.2 | setosa |
| 6 | 5.4 | 3.9 | 1.7 | 0.4 | setosa |
| 7 | 4.6 | 3.4 | 1.4 | 0.3 | setosa |
| 8 | 5 | 3.4 | 1.5 | 0.2 | setosa |
| 9 | 4.4 | 2.9 | 1.4 | 0.2 | setosa |
| 10 | 4.9 | 3.1 | 1.5 | 0.1 | setosa |
| ... | ... | ... | ... | ... | ... |

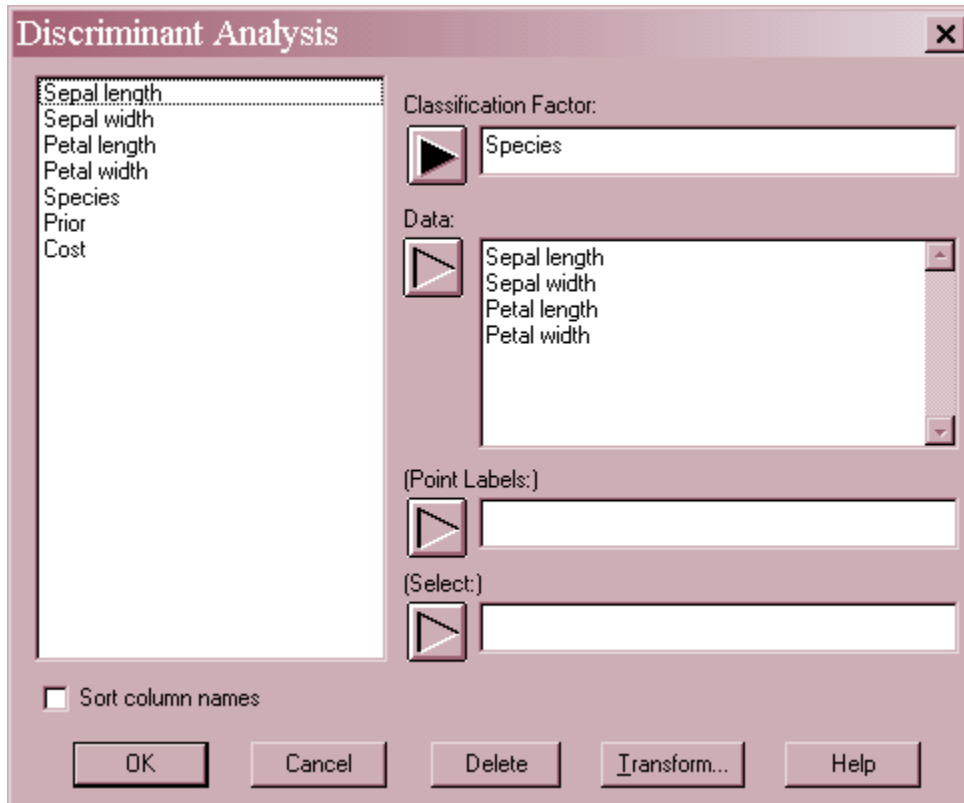
A matrix plot of the observed data is shown below:



Note how the species are naturally divided into groups. There is, however, some overlap between the groups, particularly *versicolor* and *virginica*.

Data Input

The data input dialog box requests the name of a column identifying the groups and the names of p variables to be used to discriminate amongst the groups:



- **Classification Factor:** numeric or nonnumeric column containing an identifier of which group each observation belongs to. There must be g unique values in this column.
- **Data:** the names of the p variables to be used to discriminate amongst the groups.
- **Point Labels:** optional labels for each observation.
- **Select:** subset selection.

Statistical Model

The goal of the *Discriminant Analysis* procedure is to construct s linear combinations of the p input variables that best discriminate amongst the g groups. The j -th discriminant function takes the form

$$D_j = d_{j1}Z_1 + d_{j2}Z_2 + \dots + d_{jp}Z_p \quad (1)$$

where the Z 's are the standardized input variables X , created by subtracting the sample means and dividing by the sample standard deviations.

The s discriminant functions are found by determining the eigenvalues of

$$W^{-1}B \quad (2)$$

where W is the sample *within groups* sum of squares and cross-products matrix and B is the sample *between groups* sum of squares and cross-products matrix. The coefficients of the discriminant functions are derived from the eigenvectors. Basically, the discriminant functions are derived so as to maximize the separation of the groups.

To classify new cases into groups, classification functions are also derived. To classify an observation, a score is derived for each group. The score for the j -th group is calculated from

$$C_j = c_{j1}X_1 + c_{j2}X_2 + \dots + c_{jp}X_p + c_{j0} \quad (3)$$

New cases are classified as belonging to whichever group has the largest value of $C_j * prior_j$ where $prior_j$ is the prior probability of belonging to the j -th group. The priors may be entered by the user, estimated from the data, or assumed to be equal.

Analysis Summary

The *Analysis Summary* table is shown below:

| <u>Discriminant Analysis</u> | | | | |
|---|------------|-------------|-------------|---------|
| Classification variable: Species (type of iris) | | | | |
| Independent variables: | | | | |
| Sepal length (centimeters) | | | | |
| Sepal width (centimeters) | | | | |
| Petal length (centimeters) | | | | |
| Petal width (centimeters) | | | | |
| Number of complete cases: 150 | | | | |
| Number of groups: 3 | | | | |
| Discriminant | Eigenvalue | Relative | Canonical | |
| Function | | Percentage | Correlation | |
| 1 | 32.1919 | 99.12 | 0.98482 | |
| 2 | 0.285391 | 0.88 | 0.47120 | |
| Functions | Wilks | | | |
| Derived | Lambda | Chi-Squared | DF | P-Value |
| 1 | 0.0234386 | 546.1153 | 8 | 0.0000 |
| 2 | 0.777973 | 36.5297 | 3 | 0.0000 |

Displayed in the top section of the table are:

- **Data variables:** the names of the p input variables.
- **Number of complete cases:** the number of cases n for which none of the data were missing.
- **Number of groups:** number of different groups g into which the cases are divided.
- **Discriminant Function:** the index of the discriminant function j .
- **Eigenvalue:** λ_j , the j -th eigenvalue of $W^{-1}B$.
- **Relative Percentage:** the percentage of the sum of the variances of the p independent variables accounted for by the j -th discriminant function, calculated by dividing the j -th eigenvalue by the sum of all the eigenvalues.
- **Canonical Correlation:** the canonical correlation ρ_j^{*2} associated with the j -th eigenvalue, calculated from

$$\rho_j^{*2} = \frac{\lambda_j}{1 + \lambda_j} \tag{4}$$

which represents its relative ability to discriminate amongst the groups.

- **Wilk's Lambda:** a statistic calculated from the canonical correlations according to

$$\Lambda_j = \prod_{i=j}^q (1 - \rho_i^{*2}) \tag{5}$$

- **Chi-Squared:** a test statistic used to test the hypothesis that all canonical correlations numbered j and higher are equal to 0. It is calculated from

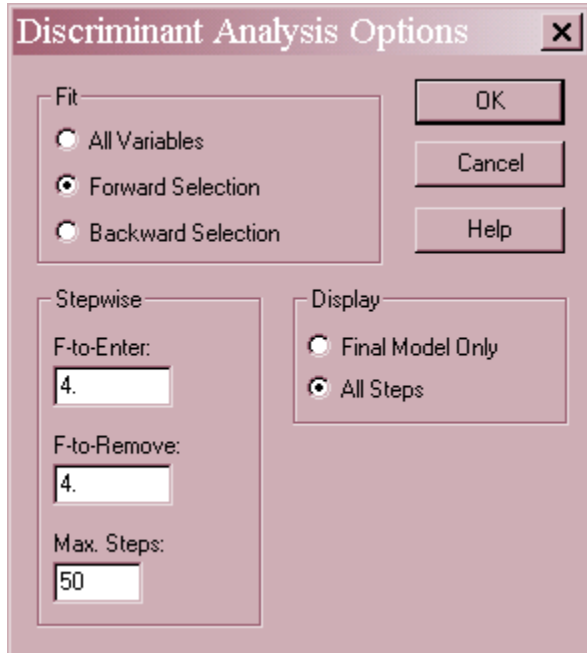
$$X^2 = -\left(n - 1 - \frac{1}{2}(p + g)\right) \ln \Lambda_j \quad (6)$$

- **D.F.:** the degrees of freedom $(p-j+1)(g-j)$ associated with the chi-squared statistic.
- **P-Value:** a one-sided P-Value for the observed chi-squared statistic. Small P-values (less than 0.05 if operating at the 5% significance level) correspond to discriminant functions that are significantly different from zero.

In the example, both of the discriminant functions are statistically significant, although the first accounts for the vast majority of the variance in the data.

Analysis Options

The *Analysis Options* dialog box determines whether all p variables will be included in the analysis or if a stepwise variable selection procedure will be used to potentially select only a subset of the variables:



- **Fit** – specifies whether all independent variables specified on the data input dialog box should be included in the final model, or whether a stepwise selection of variables should be applied. Stepwise selection attempts to find the best model that contains only statistically significant variables. An example of stepwise regression is included below.
- **F-to-Enter** - In a stepwise regression, variables will be entered into the model at a given step if their F values are greater than or equal to the *F-to-Enter* value specified.
- **F-to-Remove** - In a stepwise regression, variables will be removed from the model at a given step if their F values are less than the *F-to-Remove* value specified.
- **Max Steps** – maximum number of steps permitted when doing a stepwise regression.
- **Display** – whether to display the results at each step when doing a stepwise regression.

Example – Stepwise Regression

Analysis Options may be used to perform either a forward stepwise selection or a backward stepwise selection.

- **Forward selection** – Begins with a model involving only a constant term and enters one variable at a time based on its statistical significance if added to the current model. At each step, the algorithm brings into the model the variable that will be the most statistically significant if entered. Selection of variables is based on an *F-to-*

enter test. As long as the most significant variable has an F value greater or equal to that specified on the *Analysis Summary* dialog box, it will be brought into the model. When no variable has a large enough F value, variable selection stops. In addition, variables brought into the model early in the procedure may be removed later if their F value falls below the *F-to-remove* criterion.

- **Backward selection** – Begins with a model involving all the variables specified on the data input dialog box and removes one variable at a time based on its statistical significance in the current model. At each step, the algorithm removes from the model the variable that is the least statistically significant. Removal of variables is based on an *F-to-remove* test. If the least significant variable has an F value less than that specified on the *Analysis Summary* dialog box, it will be removed from the model. When all remaining variables have large F values, the procedure stops. In addition, variables removed from the model early in the procedure may be re-entered later if their F values reach the *F-to-enter* criterion.

The output below shows the results of a *Forward selection* for the example data:

Stepwise regression
Method: forward selection
F-to-enter: 4.0
F-to-remove: 4.0

Step 0:
0 variables in the model.

Step 1:
Adding variable Petal length with F-to-enter = 1180.16
1 variables in the model.
Wilk's lambda = 0.0586283 Approximate F = 1180.16 with P-value = 0.0000

Step 2:
Adding variable Sepal width with F-to-enter = 43.0355
2 variables in the model.
Wilk's lambda = 0.0368841 Approximate F = 307.105 with P-value = 0.0000

Step 3:
Adding variable Petal width with F-to-enter = 34.5687
3 variables in the model.
Wilk's lambda = 0.0249755 Approximate F = 257.503 with P-value = 0.0000

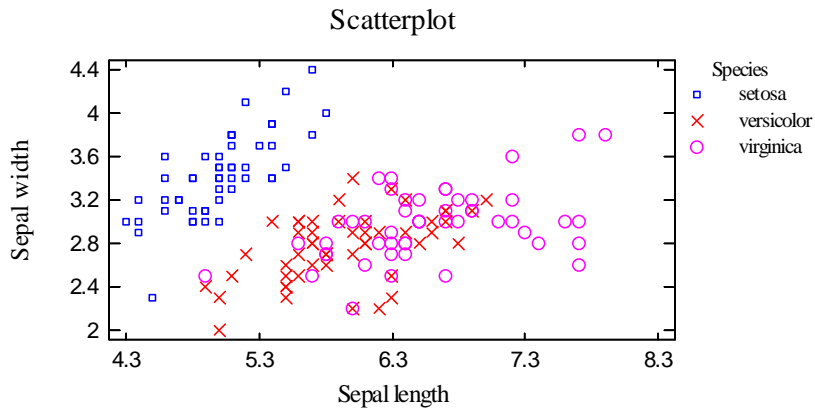
Step 4:
Adding variable Sepal length with F-to-enter = 4.72115
4 variables in the model.
Wilk's lambda = 0.0234386 Approximate F = 199.145 with P-value = 0.0000

Final model selected.

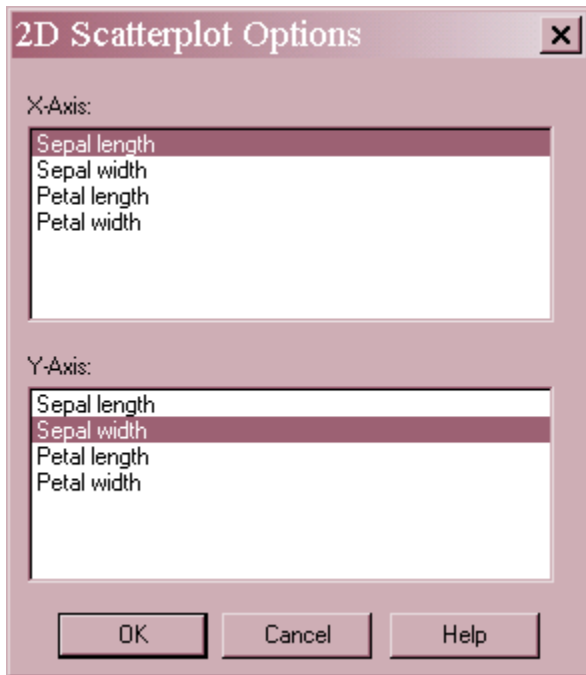
All four variables add significantly to the fit as they are entered.

2D Scatterplot

The *2D Scatterplot* plots the data for any two of the *X* variables.



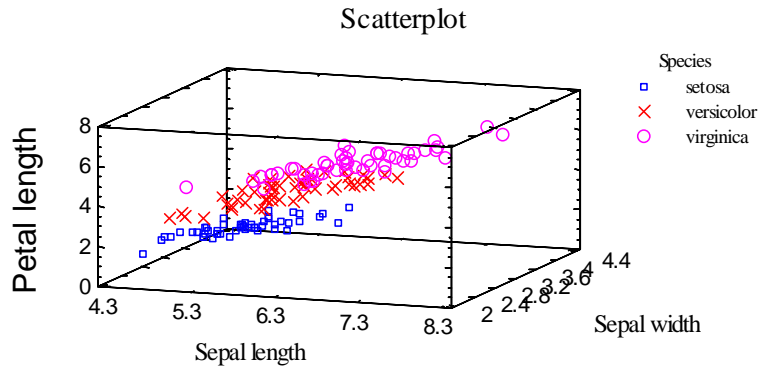
Pane Options



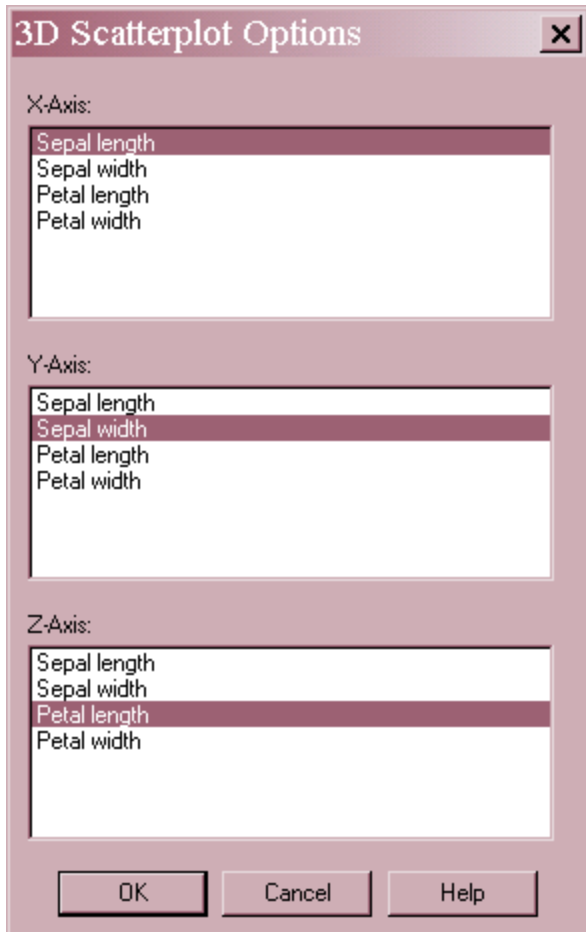
Select variables to define the horizontal and vertical axes.

3D Scatterplot

The 3D Scatterplot plots the data for any three of the X variables.



Pane Options



Select variables to define the three axes.

Group Statistics

This table displays the samples means and sample standard deviations for each of the p variables in each of the g groups.

| Summary Statistics by Group | | | | |
|-----------------------------|----------|------------|-----------|----------|
| Species | setosa | versicolor | virginica | TOTAL |
| COUNTS | 50 | 50 | 50 | 150 |
| MEANS | | | | |
| Sepal length | 5.006 | 5.936 | 6.588 | 5.84333 |
| Sepal width | 3.428 | 2.77 | 2.974 | 3.05733 |
| Petal length | 1.462 | 4.26 | 5.552 | 3.758 |
| Petal width | 0.246 | 1.326 | 2.026 | 1.19933 |
| STD. DEVIATIONS | | | | |
| Sepal length | 0.35249 | 0.516171 | 0.63588 | 0.828066 |
| Sepal width | 0.379064 | 0.313798 | 0.322497 | 0.435866 |
| Petal length | 0.173664 | 0.469911 | 0.551895 | 1.7653 |
| Petal width | 0.105386 | 0.197753 | 0.27465 | 0.762238 |

Groups Correlations

This table shows the pooled within-group estimates of the covariance and correlation matrices.

| Pooled Within-Group Statistics for Species | | | | |
|--|--------------|-------------|--------------|-------------|
| Within-Group Covariance Matrix | | | | |
| | Sepal length | Sepal width | Petal length | Petal width |
| Sepal length | 0.265008 | 0.0927211 | 0.167514 | 0.0384014 |
| Sepal width | 0.0927211 | 0.115388 | 0.0552435 | 0.0327102 |
| Petal length | 0.167514 | 0.0552435 | 0.185188 | 0.0426653 |
| Petal width | 0.0384014 | 0.0327102 | 0.0426653 | 0.0418816 |
| Within-Group Correlation Matrix | | | | |
| | Sepal length | Sepal width | Petal length | Petal width |
| Sepal length | 1.0 | 0.530236 | 0.756164 | 0.364506 |
| Sepal width | 0.530236 | 1.0 | 0.377916 | 0.470535 |
| Petal length | 0.756164 | 0.377916 | 1.0 | 0.484459 |
| Petal width | 0.364506 | 0.470535 | 0.484459 | 1.0 |

Discriminant Functions

Discriminant Functions are linear combinations of the input variables used to separate the data into the different groups. This pane shows both the standardized and unstandardized coefficients:

| Discriminant Function Coefficients for Species | | |
|--|-----------|-----------|
| | 1 | 2 |
| Sepal length | 0.426955 | 0.0124075 |
| Sepal width | 0.521242 | 0.735261 |
| Petal length | -0.947257 | -0.401038 |
| Petal width | -0.575161 | 0.58104 |

| Unstandardized Coefficients | | |
|-----------------------------|----------|-----------|
| | 1 | 2 |
| Sepal length | 0.829378 | 0.0241021 |
| Sepal width | 1.53447 | 2.16452 |
| Petal length | -2.20121 | -0.931921 |
| Petal width | -2.81046 | 2.83919 |
| CONSTANT | 2.10511 | -6.66147 |

The j -th standardized discriminant function takes the form

$$D_j = d_{j1}Z_1 + d_{j2}Z_2 + \dots + d_{jp}Z_p \quad (7)$$

where the Z 's are the standardized form of the input variables X , created by subtracting the sample means and dividing by the sample standard deviations. The j -th unstandardized discriminant function takes the form

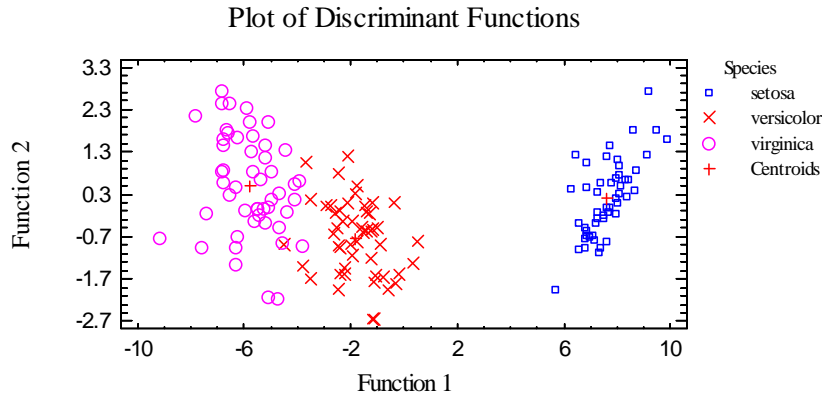
$$U_j = u_{j1}X_1 + u_{j2}X_2 + \dots + u_{jp}X_p + u_{j0} \quad (8)$$

When the variables are in different units or have different variances, more insight is usually gained from the standardized coefficients.

In the sample data, note that the first discriminant function is basically a contrast between the sepal size and the petal size. The second discriminant function is primarily a contrast between the combined width of the sepal and petal and the petal length.

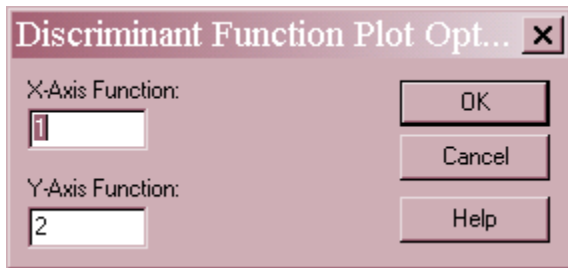
Discriminant Functions Plot

This pane displays the values of any two discriminant functions for each of the n cases.



It is very helpful in visualizing how well the functions separate the data. Clearly, the first discriminant function completely separates *setosa* from the other two species, leaving a small amount of overlap between *versicolor* and *virginica*. The second discriminant function may help a small amount in separating the latter two species. In addition to the observations, the location of the mean discriminant functions values for each group are shown as + signs.

Pane Options



Enter the numbers of the two discriminant functions to plot on the horizontal and vertical axes.

Group Centroids

The pane shows the centroid or mean values for each of the g groups on each of the s discriminant functions

| Group Centroids for Species | | |
|-----------------------------|----------|----------|
| Group | 1 | 2 |
| setosa | 7.6076 | 0.215133 |
| versicolor | -1.82505 | -0.7279 |
| virginica | -5.78255 | 0.512767 |

Classification Functions

The classification functions are used to determine which of the g groups any individual sample is most likely to belong to:

| Classification Function Coefficients for Species | | | |
|--|---------------|-------------------|------------------|
| | <i>setosa</i> | <i>Versicolor</i> | <i>virginica</i> |
| Sepal length | 23.5442 | 15.6982 | 12.4458 |
| Sepal width | 23.5879 | 7.07251 | 3.68528 |
| Petal length | -16.4306 | 5.21145 | 12.7665 |
| Petal width | -17.3984 | 6.43423 | 21.0791 |
| CONSTANT | -86.3085 | -72.8526 | -104.368 |

A score is calculated for each observation i and each group j according to

$$C_{ij} = c_{j1}X_{i1} + c_{j2}X_{i2} + \dots + c_{jp}X_{ip} \tag{9}$$

If the data are assumed to come from multivariate normal distributions, then the scores are related to the probabilities that an observation belongs to a particular group.

Classification Table

The *Classification Table* shows the result of using the derived classification rule to assign both observed cases and new cases to groups. For a given set of X values, a case is assigned to whichever group gives the largest value of $C_{ij} * prior_j$, where $prior_j$ is the prior probability that an individual comes from group j . Since the size of the populations from which each group's samples are taken may not be the same, the probability that an individual belongs to a particular group prior to examining the data may vary from group to group. For example, in screening for a disease, the proportion of individuals given a diagnostic test who actually have the disease may be very small, a fact that needs to be accounted for. Using *Pane Options*, the user specifies how to handle the prior probabilities. They can be assumed to be the same for all groups, to be proportional to the fraction of the data within each group, or input by the user.

The table below shows typical output:

| Classification Table | | | | |
|----------------------|-------|-----------|------------|-----------|
| Actual | Group | Predicted | Species | |
| Species | Size | setosa | versicolor | virginica |
| setosa | 50 | 50 | 0 | 0 |
| | | (100.00%) | (0.00%) | (0.00%) |
| versicolor | 50 | 0 | 48 | 2 |
| | | (0.00%) | (96.00%) | (4.00%) |
| virginica | 50 | 0 | 1 | 49 |
| | | (0.00%) | (2.00%) | (98.00%) |

Percent of cases correctly classified: **98.00%**

| Group | Prior Probability |
|-------|-------------------|
| 1 | 0.3333 |
| 2 | 0.3333 |
| 3 | 0.3333 |

| Row | Actual Group | Highest Group | Highest Value | Squared Distance | Prob. | 2nd Highest Group | 2nd Highest Value | Squared Distance | Prob. |
|-----|--------------|---------------|---------------|------------------|--------|-------------------|-------------------|------------------|--------|
| 71 | versicolor | *virginica | 80.0769 | 4.55382 | 0.7468 | versicolor | 78.9954 | 6.71675 | 0.2532 |
| 84 | versicolor | *virginica | 79.093 | 3.59634 | 0.8566 | versicolor | 77.3056 | 7.17114 | 0.1434 |
| 134 | virginica | *versicolor | 82.0789 | 4.0068 | 0.7294 | virginica | 81.0874 | 5.98984 | 0.2706 |
| 151 | | virginica | 99.945 | 0.73244 | 0.9996 | versicolor | 91.9996 | 16.6234 | 0.0004 |

* = incorrectly classified.

The top section of the table shows how well the classification rule performed in classifying the trial data. Each row tabulates the results for cases that actually belong to a particular group. The columns show how often they were classified as belonging to each group. Displayed along the bottom is the percentage of cases correctly classified.

The center part of the table displays the prior probabilities. For the example data, the prior probabilities were assumed to be the same for all groups.

The lower part of the table shows the two groups that received the highest scores for selected cases. The table shows:

1. *Highest and second highest group* – the two groups with the highest scores.
2. *Values* – the values of the scores calculated for the two groups.
3. *Squared Distance* – the squared Mahalanobis distance of the observations from the group centroids, in the space of the discriminant functions. The farther an observation is from a group centroid, the less likely it belongs to that group.
4. *Probability* – the estimated probability that the case belongs to a particular group. The probability is based on the ratio of the height of the normal density function at the distance of the observation from each group centroid and on the prior probabilities.

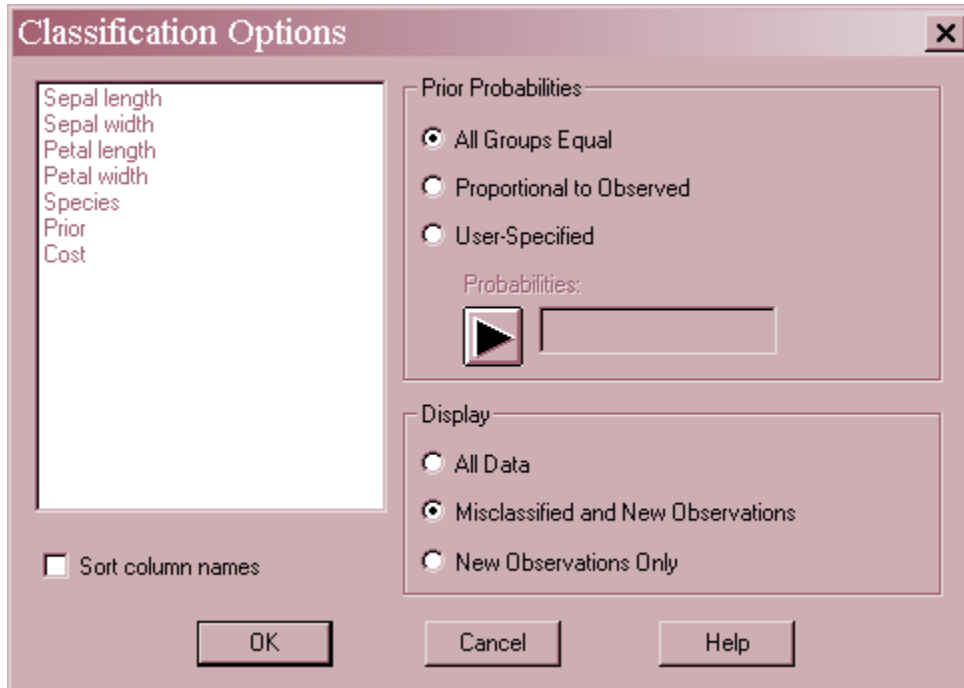
For example, suppose a new iris was observed with the following features:

sepal length = 6.6 inches
 sepal width = 2.9 inches
 petal length = 5.1 inches
 petal width = 2.2 inches

These values would be placed in row #151 of the datasheet. The table shows that the group with the highest score for those values is *virginica*, followed by *versicolor*. The large difference

between the distances and thus the posterior probabilities implies that the sample most likely belongs to the *virginica* group.

Pane Options



- **Prior Probabilities:** method for determining the probability of group membership before the data is examined. Select *All Groups Equal* to assume equal priors for all groups, *Proportional to Observed* to set the priors equal to the fraction of n represented by each group, or *User-Specified* to enter a column with g values that sum to 1.
- **Display:** *All Data* will display all of the observations in the datasheet, *Misclassified and New Observations* will display any cases that were either misclassified or had a missing value for the group indicator, while *New Observations Only* will display only data not used to determine the discriminant functions.

Save Results

The following results can be saved to a datasheet:

1. *Discriminant Function Values* – the values D of the discriminant functions for each of the n observations.
2. *Classification Function Values*– the values C of the classification functions for each of the n observations.
3. *Standardized Discriminant Function Coefficients*- s columns containing the values of the p coefficients d_{ij} of each standardized discriminant function.
4. *Unstandardized Discriminant Function Coefficients* - s columns containing the values of the $p+1$ coefficients u_{ij} of each unstandardized discriminant function.
5. *Classification Function Coefficients* – s columns containing the values of the $p+1$ coefficients c_{ij} of each classification function.
6. *Prior Probabilities* – the prior probabilities of belonging to each of the g groups.
7. *Sample Means* – the means of each of the p X variables.
8. *Sample Standard Deviations* – the standard deviations of each of the p X variables.