# Chapter 21
# The CANDISC Procedure

## Chapter Table of Contents

# Chapter 21
# The CANDISC Procedure

## Overview

Canonical discriminant analysis is a dimension-reduction technique related to principal component analysis and canonical correlation. The methodology used in deriving the canonical coefficients parallels that of a one-way MANOVA. Whereas in MANOVA the goal is to test for equality of the mean vector across class levels, in a canonical discriminant analysis we find linear combinations of the quantitative variables that provide maximal separation between the classes or groups. Given a classification variable and several quantitative variables, the CANDISC procedure derives *canonical variables*, linear combinations of the quantitative variables that summarize between-class variation in much the same way that principal components summarize total variation.

The CANDISC procedure performs a canonical discriminant analysis, computes squared Mahalanobis distances between class means, and performs both univariate and multivariate one-way analyses of variance. Two output data sets can be produced: one containing the canonical coefficients and another containing, among other things, scored canonical variables. The canonical coefficients output data set can be rotated by the FACTOR procedure. It is customary to standardize the canonical coefficients so that the canonical variables have means that are equal to zero and pooled within-class variances that are equal to one. PROC CANDISC displays both standardized and unstandardized canonical coefficients. Correlations between the canonical variables and the original variables as well as the class means for the canonical variables are also displayed; these correlations, sometimes known as loadings, are called canonical structures. The scored canonical variables output data set can be used in conjunction with the PLOT procedure or the %PLOTIT macro to plot pairs of canonical variables to aid visual interpretation of group differences.

Given two or more groups of observations with measurements on several quantitative variables, canonical discriminant analysis derives a linear combination of the variables that has the highest possible multiple correlation with the groups. This maximal multiple correlation is called the *first canonical correlation*. The coefficients of the linear combination are the *canonical coefficients* or *canonical weights*. The variable defined by the linear combination is the *first canonical variable* or *canonical component*. The second canonical correlation is obtained by finding the linear combination uncorrelated with the first canonical variable that has the highest possible multiple correlation with the groups. The process of extracting canonical variables can be repeated until the number of canonical variables equals the number of original variables or the number of classes minus one, whichever is smaller.

The first canonical correlation is at least as large as the multiple correlation between the groups and any of the original variables. If the original variables have high within-group correlations, the first canonical correlation can be large even if all the multiple correlations are small. In other words, the first canonical variable can show substantial differences between the classes, even if none of the original variables do. Canonical variables are sometimes called *discriminant functions*, but this usage is ambiguous because the DISCRIM procedure produces very different functions for classification that are also called discriminant functions.

For each canonical correlation, PROC CANDISC tests the hypothesis that it and all smaller canonical correlations are zero in the population. An $F$ approximation (Rao 1973; Kshirsagar 1972) is used that gives better small-sample results than the usual chi-square approximation. The variables should have an approximate multivariate normal distribution within each class, with a common covariance matrix in order for the probability levels to be valid.

Canonical discriminant analysis is equivalent to canonical correlation analysis between the quantitative variables and a set of dummy variables coded from the class variable. Canonical discriminant analysis is also equivalent to performing the following steps:

1. Transform the variables so that the pooled within-class covariance matrix is an identity matrix.
2. Compute class means on the transformed variables.
3. Perform a principal component analysis on the means, weighting each mean by the number of observations in the class. The eigenvalues are equal to the ratio of between-class variation to within-class variation in the direction of each principal component.
4. Back-transform the principal components into the space of the original variables, obtaining the canonical variables.

An interesting property of the canonical variables is that they are uncorrelated whether the correlation is calculated from the total sample or from the pooled within-class correlations. The canonical coefficients are not orthogonal, however, so the canonical variables do not represent perpendicular directions through the space of the original variables.

# Getting Started

The data in this example are measurements on 159 fish caught off the coast of Finland. The species, weight, three different length measurements, height, and width of each fish is tallied. The complete data set is displayed in Chapter 60, “The STEPDISC Procedure”; the STEPDISC procedure identified all the variables as significant indicators of the differences among the seven fish species.

```
proc format;
   value specfmt
      1=’Bream’
      2=’Roach’
      3=’Whitefish’
      4=’Parkki’
      5=’Perch’
      6=’Pike’
      7=’Smelt’;
data fish (drop=HtPct WidthPct);
   title ’Fish Measurement Data’;
   input Species Weight Length1 Length2 Length3 HtPct
         WidthPct @@;
   Height=HtPct*Length3/100;
   Width=WidthPct*Length3/100;
   format Species specfmt.;
   symbol = put(Species, specfmt2.);
   datalines;
1  242.0 23.2 25.4 30.0 38.4 13.4
1  290.0 24.0 26.3 31.2 40.0 13.8
1  340.0 23.9 26.5 31.1 39.8 15.1
1  363.0 26.3 29.0 33.5 38.0 13.3
 ...[155 more records]
;
```

The following program uses PROC CANDISC to find the three canonical variables that best separate the species of fish in the fish data and creates the output data set outcan. The NCAN= option is used to request that only the first three canonical variables are displayed. The %PLOTIT macro is invoked to create a plot of the first two canonical variables. See Appendix B, “Using the %PLOTIT Macro,” for more information on the %PLOTIT macro.

```
proc candisc data=fish ncan=3 out=outcan;
   class Species;
   var Weight Length1 Length2 Length3 Height Width;
run;
%plotit(data=outcan, plotvars=Can2 Can1,
        labelvar=_blank_, symvar=symbol, typevar=symbol,
        symsize=1, symlen=4, tsize=1.5, exttypes=symbol, ls=100,
        plotopts=vaxis=-5 to 15 by 5, vtoh=, extend=close);
```

PROC CANDISC begins by displaying summary information about the variables in the analysis. This information includes the number of observations, the number of quantitative variables in the analysis (specified with the VAR statement), and the number of classes in the classification variable (specified with the CLASS statement). The frequency of each class is also displayed.

```
                             Fish Measurement Data

                             The CANDISC Procedure

            Observations     158          DF Total               157
            Variables          6          DF Within Classes      151
            Classes            7          DF Between Classes       6


                            Class Level Information

                         Variable
         Species         Name           Frequency       Weight    Proportion

         Bream           Bream                 34      34.0000      0.215190
         Parkki          Parkki                11      11.0000      0.069620
         Perch           Perch                 56      56.0000      0.354430
         Pike            Pike                  17      17.0000      0.107595
         Roach           Roach                 20      20.0000      0.126582
         Smelt           Smelt                 14      14.0000      0.088608
         Whitefish       Whitefish              6       6.0000      0.037975
```

**Figure 21.1.** Summary Information

PROC CANDISC performs a multivariate one-way analysis of variance (one-way MANOVA) and provides four multivariate tests of the hypothesis that the class mean vectors are equal. These tests, shown in Figure 21.2, indicate that not all of the mean vectors are equal $(p < .0001)$.

```
                              Fish Measurement Data

                              The CANDISC Procedure

                  Multivariate Statistics and F Approximations

                              S=6    M=-0.5    N=72

Statistic                        Value    F Value    Num DF    Den DF    Pr > F

Wilks' Lambda               0.00036325      90.71        36    643.89    <.0001
Pillai's Trace              3.10465132      26.99        36       906    <.0001
Hotelling-Lawley Trace     52.05799676     209.24        36    413.64    <.0001
Roy's Greatest Root        39.13499776     984.90         6       151    <.0001

          NOTE: F Statistic for Roy's Greatest Root is an upper bound.
```

**Figure 21.2.** MANOVA and Multivariate Tests

The first canonical correlation is the greatest possible multiple correlation with the classes that can be achieved using a linear combination of the quantitative variables. The first canonical correlation, displayed in Figure 21.3, is 0.987463.

```
                              Fish Measurement Data

                              The CANDISC Procedure

                            Adjusted    Approximate       Squared
                 Canonical   Canonical       Standard     Canonical
               Correlation  Correlation          Error   Correlation

          1       0.987463     0.986671       0.001989      0.975084
          2       0.952349     0.950095       0.007425      0.906969
          3       0.838637     0.832518       0.023678      0.703313
          4       0.633094     0.623649       0.047821      0.400809
          5       0.344157     0.334170       0.070356      0.118444
          6       0.005701      .             0.079806      0.000033
```

**Figure 21.3.** Canonical Correlations

A likelihood ratio test is displayed of the hypothesis that the current canonical correlation and all smaller ones are zero. The first line is equivalent to Wilks' Lambda multivariate test.

```
                  Test of H0: The canonical correlations in the
                    current row and all that follow are zero

                Likelihood    Approximate
                     Ratio        F Value    Num DF    Den DF    Pr > F

          1     0.00036325          90.71        36    643.89    <.0001
          2     0.01457896          46.46        25    547.58    <.0001
          3     0.15671134          23.61        16    452.79    <.0001
          4     0.52820347          12.09         9    362.78    <.0001
          5     0.88152702           4.88         4       300    0.0008
          6     0.99996749           0.00         1       151    0.9442
```

**Figure 21.4.** Likelihood Ratio Test

The first canonical variable, Can1, shows that the linear combination of the centered variables $\text{Can1} = -0.0006 \times \text{Weight} - 0.33 \times \text{Length1} - 2.49 \times \text{Length2} + 2.60 \times \text{Length3} + 1.12 \times \text{Height} - 1.45 \times \text{Width}$ separates the species most effectively (see Figure 21.5).

```
                           Fish Measurement Data

                           The CANDISC Procedure

                         Raw Canonical Coefficients

      Variable               Can1              Can2              Can3

      Weight         -0.000648508      -0.005231659      -0.005596192
      Length1        -0.329435762      -0.626598051      -2.934324102
      Length2        -2.486133674      -0.690253987       4.045038893
      Length3         2.595648437       1.803175454      -1.139264914
      Height          1.121983854      -0.714749340       0.283202557
      Width          -1.446386704      -0.907025481       0.741486686
```

**Figure 21.5.** Raw Canonical Coefficients

PROC CANDISC computes the means of the canonical variables for each class. The first canonical variable is the linear combination of the variables Weight, Length1, Length2, Length3, Height, and Width that provides the greatest difference (in terms of a univariate $F$-test) between the class means. The second canonical variable provides the greatest difference between class means while being uncorrelated with the first canonical variable.

```
                           Fish Measurement Data

                           The CANDISC Procedure

                    Class Means on Canonical Variables

     Species                 Can1              Can2              Can3

     Bream            10.94142464        0.52078394        0.23496708
     Parkki            2.58903743       -2.54722416       -0.49326158
     Perch            -4.47181389       -1.70822715        1.29281314
     Pike             -4.89689441        8.22140791       -0.16469132
     Roach            -0.35837149        0.08733611       -1.10056438
     Smelt            -4.09136653       -2.35805841       -4.03836098
     Whitefish        -0.39541755       -0.42071778        1.06459242
```

**Figure 21.6.** Class Means for Canonical Variables

A plot of the first two canonical variables (Figure 21.7) shows that Can1 discriminates between three groups: 1) bream; 2) whitefish, roach, and parkki; and 3) smelt, pike, and perch. Can2 best discriminates between pike and the other species.

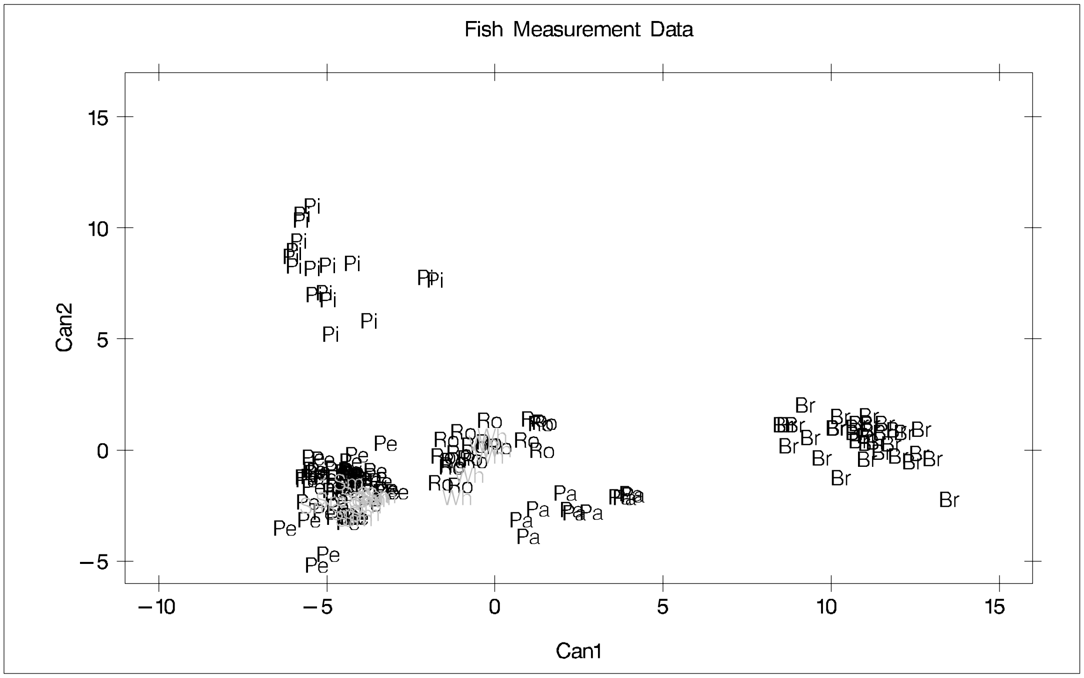


**Figure 21.7.** Plot of First Two Canonical Variables

# Syntax

The following statements are available in PROC CANDISC.

```
PROC CANDISC < options > ;
    CLASS variable ;
    BY variables ;
    FREQ variable ;
    VAR variables ;
    WEIGHT variable ;
```

The BY, CLASS, FREQ, VAR, and WEIGHT statements are described after the PROC CANDISC statement.

## PROC CANDISC Statement

**PROC CANDISC** < *options* > **;**

This statement invokes the CANDISC procedure. The options listed in the following table can appear in the PROC CANDISC statement.

**Table 21.1.** CANDISC Procedure Options

| **Task** | **Options** |
|---|---|
| Specify Data Sets | DATA=<br>OUT=<br>OUTSTAT= |
| Control Canonical Variables | NCAN=<br>PREFIX= |
| Determine Singularity | SINGULAR= |
| Control Displayed Correlations | BCORR<br>PCORR<br>TCORR<br>WCORR |
| Control Displayed Covariances | BCOV<br>PCOV<br>TCOV<br>WCOV |
| Control Displayed SSCP Matrices | BSSCP<br>PSSCP<br>TSSCP<br>WSSCP |
| Suppress Output | NOPRINT<br>SHORT |
| Miscellaneous | ALL<br>ANOVA<br>DISTANCE<br>SIMPLE<br>STDMEAN |

**ALL**

activates all of the display options.

**ANOVA**

displays univariate statistics for testing the hypothesis that the class means are equal in the population for each variable.

**BCORR**

displays between-class correlations.

**BCOV**

displays between-class covariances. The between-class covariance matrix equals the between-class SSCP matrix divided by $n(c-1)/c$, where $n$ is the number of observations and $c$ is the number of classes. The between-class covariances should be interpreted in comparison with the total-sample and within-class covariances, not as formal estimates of population parameters.

**BSSCP**

displays the between-class SSCP matrix.

**DATA=***SAS-data-set*

specifies the data set to be analyzed. The data set can be an ordinary SAS data set or one of several specially structured data sets created by SAS statistical procedures. These specially structured data sets include TYPE=CORR, COV, CSSCP, and SSCP. If you omit the DATA= option, the procedure uses the most recently created SAS data set.

**DISTANCE**

displays squared Mahalanobis distances between the group means, $F$ statistics, and the corresponding probabilities of greater squared Mahalanobis distances between the group means.

**NCAN=***n*

specifies the number of canonical variables to be computed. The value of $n$ must be less than or equal to the number of variables. If you specify NCAN=0, the procedure displays the canonical correlations, but not the canonical coefficients, structures, or means. A negative value suppresses the canonical analysis entirely. Let $v$ be the number of variables in the VAR statement and $c$ be the number of classes. If you omit the NCAN= option, only $\min(v, c-1)$ canonical variables are generated; if you also specify an OUT= output data set, $v$ canonical variables are generated, and the last $v-(c-1)$ canonical variables have missing values.

**NOPRINT**

suppresses the normal display of results. Note that this option temporarily disables the Output Delivery System (ODS); see Chapter 15, "Using the Output Delivery System," for more information.

**OUT=***SAS-data-set*

creates an output SAS data set containing the original data and the canonical variable scores. To create a permanent SAS data set, specify a two-level name (refer to *SAS Language Reference: Concepts*, for more information on permanent SAS data sets).

**OUTSTAT=***SAS-data-set*

creates a TYPE=CORR output SAS data set that contains various statistics including class means, standard deviations, correlations, canonical correlations, canonical structures, canonical coefficients, and means of canonical variables for each class. To create a permanent SAS data set, specify a two-level name (refer to *SAS Language Reference: Concepts*, for more information on permanent SAS data sets).

**PCORR**

displays pooled within-class correlations (partial correlations based on the pooled within-class covariances).

**PCOV**

displays pooled within-class covariances.

**PREFIX=*name***

specifies a prefix for naming the canonical variables. By default the names are Can1, Can2, Can3 and so forth. If you specify PREFIX=Abc, the components are named Abc1, Abc2, and so on. The number of characters in the prefix, plus the number of digits required to designate the canonical variables, should not exceed 32. The prefix is truncated if the combined length exceeds 32.

**PSSCP**

displays the pooled within-class corrected SSCP matrix.

**SHORT**

suppresses the display of canonical structures, canonical coefficients, and class means on canonical variables; only tables of canonical correlations and multivariate test statistics are displayed.

**SIMPLE**

displays simple descriptive statistics for the total sample and within each class.

**SINGULAR=*p***

specifies the criterion for determining the singularity of the total-sample correlation matrix and the pooled within-class covariance matrix, where $0 < p < 1$. The default is SINGULAR=1E$-8$.

Let **S** be the total-sample correlation matrix. If the $R^2$ for predicting a quantitative variable in the VAR statement from the variables preceding it exceeds $1 - p$, **S** is considered singular. If **S** is singular, the probability levels for the multivariate test statistics and canonical correlations are adjusted for the number of variables with $R^2$ exceeding $1 - p$.

If **S** is considered singular and the inverse of **S** (Squared Mahalanobis Distances) is required, a quasi-inverse is used instead. For details see the “Quasi-Inverse” section in Chapter 25, “The DISCRIM Procedure.”

**STDMEAN**

displays total-sample and pooled within-class standardized class means.

**TCORR**

displays total-sample correlations.

**TCOV**

displays total-sample covariances.

**TSSCP**

displays the total-sample corrected SSCP matrix.

**WCORR**
displays within-class correlations for each class level.

**WCOV**
displays within-class covariances for each class level.

**WSSCP**
displays the within-class corrected SSCP matrix for each class level.

## BY Statement

**BY** *variables* **;**

You can specify a BY statement with PROC CANDISC to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED or DESCENDING in the BY statement for the CANDISC procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables using the DATASETS procedure (in base SAS software).

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Concepts*. For more information on the DATASETS procedure, refer to the discussion in the *SAS Procedures Guide*.

## CLASS Statement

**CLASS** *variable* **;**

The values of the CLASS variable define the groups for analysis. Class levels are determined by the formatted values of the CLASS variable. The CLASS variable can be numeric or character. A CLASS statement is required.

## FREQ Statement

**FREQ** *variable* **;**

If a variable in the data set represents the frequency of occurrence for the other values in the observation, include the name of the variable in a FREQ statement. The procedure then treats the data set as if each observation appears $n$ times, where $n$ is the value of the FREQ variable for the observation. The total number of observations is considered to be equal to the sum of the FREQ variable when the procedure determines degrees of freedom for significance probabilities.

If the value of the FREQ variable is missing or is less than one, the observation is not used in the analysis. If the value is not an integer, the value is truncated to an integer.

## VAR Statement

**VAR** *variables* **;**

You specify the quantitative variables to include in the analysis using a VAR statement. If you do not use a VAR statement, the analysis includes all numeric variables not listed in other statements.

## WEIGHT Statement

**WEIGHT** *variable* **;**

To use relative weights for each observation in the input data set, place the weights in a variable in the data set and specify the name in a WEIGHT statement. This is often done when the variance associated with each observation is different and the values of the WEIGHT variable are proportional to the reciprocals of the variances. If the value of the WEIGHT variable is missing or is less than zero, then a value of zero for the weight is assumed.

The WEIGHT and FREQ statements have a similar effect except that the WEIGHT statement does not alter the degrees of freedom.

# Details

## Missing Values

If an observation has a missing value for any of the quantitative variables, it is omitted from the analysis. If an observation has a missing CLASS value but is otherwise complete, it is not used in computing the canonical correlations and coefficients; however, canonical variable scores are computed for that observation for the OUT= data set.

## Computational Details

### *General Formulas*

Canonical discriminant analysis is equivalent to canonical correlation analysis between the quantitative variables and a set of dummy variables coded from the class variable. In the following notation the dummy variables will be denoted by $\mathbf{y}$ and the quantitative variables by $\mathbf{x}$. The total sample covariance matrix for the $\mathbf{x}$ and $\mathbf{y}$ variables is

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{xx} & \mathbf{S}_{xy} \\ \mathbf{S}_{yx} & \mathbf{S}_{yy} \end{bmatrix}$$

When $c$ is the number of groups, $n_t$ is the number of observations in group $t$, and $\mathbf{S}_t$ is the sample covariance matrix for the $\mathbf{x}$ variables in group $t$, the within-class pooled covariance matrix for the $\mathbf{x}$ variables is

$$\mathbf{S}_p = \frac{1}{\sum n_t - c} \sum (n_t - 1)\mathbf{S}_t$$

The canonical correlations, $\rho_i$, are the square roots of the eigenvalues, $\lambda_i$, of the following matrix. The corresponding eigenvectors are $\mathbf{v}_i$.

$$\mathbf{S}_p{}^{-1/2}\mathbf{S}_{xy}\mathbf{S}_{yy}{}^{-1}\mathbf{S}_{yx}\mathbf{S}_p{}^{-1/2}$$

Let $\mathbf{V}$ be the matrix with the eigenvectors $\mathbf{v}_i$ that correspond to nonzero eigenvalues as columns. The raw canonical coefficients are calculated as follows

$$\mathbf{R} = \mathbf{S}_p{}^{-1/2}\mathbf{V}$$

The pooled within-class standardized canonical coefficients are

$$\mathbf{P} = \mathrm{diag}(\mathbf{S}_p)^{1/2}\mathbf{R}$$

And the total sample standardized canonical coefficients are

$$\mathbf{T} = \mathrm{diag}(\mathbf{S}_{xx})^{1/2}\mathbf{R}$$

Let $\mathbf{X}_c$ be the matrix with the centered $\mathbf{x}$ variables as columns. The canonical scores may be calculated by any of the following

$$\mathbf{X}_c\,\mathbf{R}$$

$$\mathbf{X}_c\,\mathrm{diag}(\mathbf{S}_p)^{-1/2}\mathbf{P}$$

$$\mathbf{X}_c\,\mathrm{diag}(\mathbf{S}_{xx})^{-1/2}\mathbf{T}$$

For the Multivariate tests based on $\mathbf{E}^{-1}\mathbf{H}$

$$\mathbf{E} = (n-1)(\mathbf{S}_{yy} - \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy})$$

$$\mathbf{H} = (n-1)\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$$

where $n$ is the total number of observations.

## Input Data Set

The input DATA= data set can be an ordinary SAS data set or one of several specially structured data sets created by statistical procedures available with SAS/STAT software. For more information on special types of data sets, see Appendix A, “Special SAS Data Sets.” The BY variable in these data sets becomes the CLASS variable in PROC CANDISC. These specially structured data sets include

- TYPE=CORR data sets created by PROC CORR using a BY statement
- TYPE=COV data sets created by PROC PRINCOMP using both the COV option and a BY statement
- TYPE=CSSCP data sets created by PROC CORR using the CSSCP option and a BY statement, where the OUT= data set is assigned TYPE=CSSCP with the TYPE= data set option
- TYPE=SSCP data sets created by PROC REG using both the OUTSSCP= option and a BY statement.

When the input data set is TYPE=CORR, TYPE=COV, or TYPE=CSSCP, PROC CANDISC reads the number of observations for each class from the observations with _TYPE_=’N’ and the variable means in each class from the observations with _TYPE_=’MEAN’. The CANDISC procedure then reads the within-class correlations from the observations with _TYPE_=’CORR’, the standard deviations from the observations with _TYPE_=’STD’ (data set TYPE=CORR), the within-class covariances from the observations with _TYPE_=’COV’ (data set TYPE=COV), or the within-class corrected sums of squares and crossproducts from the observations with _TYPE_=’CSSCP’ (data set TYPE=CSSCP).

When the data set does not include any observations with _TYPE_=’CORR’ (data set TYPE=CORR), _TYPE_=’COV’ (data set TYPE=COV), or _TYPE_=’CSSCP’ (data set TYPE=CSSCP) for each class, PROC CANDISC reads the pooled within-class information from the data set. In this case, PROC CANDISC reads the pooled within-class correlations from the observations with _TYPE_='PCORR', the pooled within-class standard deviations from the observations with _TYPE_='PSTD' (data set TYPE=CORR), the pooled within-class covariances from the observations with _TYPE_='PCOV' (data set TYPE=COV), or the pooled within-class corrected SSCP matrix from the observations with_TYPE_='PSSCP' (data set TYPE=CSSCP).

When the input data set is TYPE=SSCP, PROC CANDISC reads the number of observations for each class from the observations with _TYPE_='N', the sum of weights of observations from the variable INTERCEPT in observations with _TYPE_='SSCP' and _NAME_='INTERCEPT', the variable sums from the variable=*variablenames* in observations with _TYPE_='SSCP' and _NAME_='INTERCEPT', and the uncorrected sums of squares and crossproducts from the variable=*variablenames* in observations with _TYPE_='SSCP' and _NAME_=*variablenames*.

## Output Data Sets

### *OUT= Data Set*

The OUT= data set contains all the variables in the original data set plus new variables containing the canonical variable scores. You determine the number of new variables using the NCAN= option. The names of the new variables are formed as described in the PREFIX= option. The new variables have means equal to zero and pooled within-class variances equal to one. An OUT= data set cannot be created if the DATA= data set is not an ordinary SAS data set.

### *OUTSTAT= Data Set*

The OUTSTAT= data set is similar to the TYPE=CORR data set produced by the CORR procedure but contains many results in addition to those produced by the CORR procedure.

The OUTSTAT= data set is TYPE=CORR, and it contains the following variables:

- the BY variables, if any
- the CLASS variable
- _TYPE_, a character variable of length 8 that identifies the type of statistic
- _NAME_, a character variable of length 32 that identifies the row of the matrix or the name of the canonical variable
- the quantitative variables (those in the VAR statement, or if there is no VAR statement, all numeric variables not listed in any other statement)

The observations, as identified by the variable _TYPE_, have the following _TYPE_ values:

| **_TYPE_** | **Contents** |
|---|---|
| N | number of observations for both the total sample (CLASS variable missing) and within each class (CLASS variable present) |
| SUMWGT | sum of weights for both the total sample (CLASS variable missing) and within each class (CLASS variable present) if a WEIGHT statement is specified |
| MEAN | means for both the total sample (CLASS variable missing) and within each class (CLASS variable present) |
| STDMEAN | total-standardized class means |
| PSTDMEAN | pooled within-class standardized class means |
| STD | standard deviations for both the total sample (CLASS variable missing) and within each class (CLASS variable present) |
| PSTD | pooled within-class standard deviations |
| BSTD | between-class standard deviations |
| RSQUARED | univariate $R^2$s |

The following kinds of observations are identified by the combination of the variables _TYPE_ and _NAME_. When the _TYPE_ variable has one of the following values, the _NAME_ variable identifies the row of the matrix.

| **_TYPE_** | **Contents** |
|---|---|
| CSSCP | corrected SSCP matrix for the total sample (CLASS variable missing) and within each class (CLASS variable present) |
| PSSCP | pooled within-class corrected SSCP matrix |
| BSSCP | between-class SSCP matrix |
| COV | covariance matrix for the total sample (CLASS variable missing) and within each class (CLASS variable present) |
| PCOV | pooled within-class covariance matrix |
| BCOV | between-class covariance matrix |
| CORR | correlation matrix for the total sample (CLASS variable missing) and within each class (CLASS variable present) |
| PCORR | pooled within-class correlation matrix |
| BCORR | between-class correlation matrix |

When the _TYPE_ variable has one of the following values, the _NAME_ variable identifies the canonical variable:

| **_TYPE_** | **Contents** |
|---|---|
| CANCORR | canonical correlations |
| STRUCTUR | canonical structure |
| BSTRUCT | between canonical structure |
| PSTRUCT | pooled within-class canonical structure |
| SCORE | total sample standardized canonical coefficients |
| PSCORE | pooled within-class standardized canonical coefficients |
| RAWSCORE | raw canonical coefficients |
| CANMEAN | means of the canonical variables for each class |

## Computational Resources

In the following discussion, let

| | | |
|---|---|---|
| $n$ | $=$ | number of observations |
| $c$ | $=$ | number of class levels |
| $v$ | $=$ | number of variables in the VAR list |
| $l$ | $=$ | length of the CLASS variable |

### *Memory Requirements*

The amount of memory in bytes for temporary storage needed to process the data is

$$c(4v^2 + 28v + 4l + 68) + 16v^2 + 96v + 4l$$

With the ANOVA option, the temporary storage must be increased by 16v bytes. The DISTANCE option requires an additional temporary storage of $4v^2 + 4v$ bytes.

### *Time Requirements*

The following factors determine the time requirements of the CANDISC procedure.

- The time needed for reading the data and computing covariance matrices is proportional to $nv^2$. PROC CANDISC must also look up each class level in the list. This is faster if the data are sorted by the CLASS variable. The time for looking up class levels is proportional to a value ranging from $n$ to $n\log(c)$.
- The time for inverting a covariance matrix is proportional to $v^3$.
- The time required for the canonical discriminant analysis is proportional to $v^3$.

Each of the preceding factors has a different constant of proportionality.

## Displayed Output

The output produced by PROC CANDISC includes

- Class Level Information, including the values of the classification variable, the Frequency and Weight of each value, and its Proportion in the total sample.

Optional output includes

- Within-Class SSCP Matrices for each group
- Pooled Within-Class SSCP Matrix
- Between-Class SSCP Matrix
- Total-Sample SSCP Matrix
- Within-Class Covariance Matrices for each group
- Pooled Within-Class Covariance Matrix
- Between-Class Covariance Matrix, equal to the between-class SSCP matrix divided by $n(c-1)/c$, where $n$ is the number of observations and $c$ is the number of classes
- Total-Sample Covariance Matrix
- Within-Class Correlation Coefficients and $\mathrm{Pr} > |r|$ to test the hypothesis that the within-class population correlation coefficients are zero
- Pooled Within-Class Correlation Coefficients and $\mathrm{Pr} > |r|$ to test the hypothesis that the partial population correlation coefficients are zero
- Between-Class Correlation Coefficients and $\mathrm{Pr} > |r|$ to test the hypothesis that the between-class population correlation coefficients are zero
- Total-Sample Correlation Coefficients and $\mathrm{Pr} > |r|$ to test the hypothesis that the total population correlation coefficients are zero
- Simple Statistics including N (the number of observations), Sum, Mean, Variance, and Standard Deviation both for the total sample and within each class
- Total-Sample Standardized Class Means, obtained by subtracting the grand mean from each class mean and dividing by the total sample standard deviation
- Pooled Within-Class Standardized Class Means, obtained by subtracting the grand mean from each class mean and dividing by the pooled within-class standard deviation
- Pairwise Squared Distances Between Groups
- Univariate Test Statistics, including Total-Sample Standard Deviations, Pooled Within-Class Standard Deviations, Between-Class Standard Deviations, $R^2$, $R^2/(1-R^2)$, $F$, and $\mathrm{Pr} > F$ (univariate $F$ values and probability levels for one-way analyses of variance)

By default, PROC CANDISC displays these statistics:

- Multivariate Statistics and $F$ Approximations including Wilks' Lambda, Pillai's Trace, Hotelling-Lawley Trace, and Roy's Greatest Root with $F$ approximations, degrees of freedom (Num DF and Den DF), and probability values $(\mathrm{Pr} > F)$. Each of these four multivariate statistics tests the hypothesis that the class means are equal in the population. See the "Multivariate Tests" section in Chapter 3, "Introduction to Regression Procedures," for more information.
- Canonical Correlations
- Adjusted Canonical Correlations (Lawley 1959). These are asymptotically less biased than the raw correlations and can be negative. The adjusted canonical correlations may not be computable and are displayed as missing values if two canonical correlations are nearly equal or if some are close to zero. A missing value is also displayed if an adjusted canonical correlation is larger than a previous adjusted canonical correlation.
- Approx Standard Error, approximate standard error of the canonical correlations
- Squared Canonical Correlations
- Eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$. Each eigenvalue is equal to $\rho^2/(1-\rho^2)$, where $\rho^2$ is the corresponding squared canonical correlation and can be interpreted as the ratio of between-class variation to pooled within-class variation for the corresponding canonical variable. The table includes Eigenvalues, Differences between successive eigenvalues, the Proportion of the sum of the eigenvalues, and the Cumulative proportion.
- Likelihood Ratio for the hypothesis that the current canonical correlation and all smaller ones are zero in the population. The likelihood ratio for the hypothesis that all canonical correlations equal zero is Wilks' lambda.
- Approx $F$ statistic based on Rao's approximation to the distribution of the likelihood ratio (Rao 1973, p. 556; Kshirsagar 1972, p. 326)
- Num DF (numerator degrees of freedom), Den DF (denominator degrees of freedom), and $\mathrm{Pr} > F$, the probability level associated with the $F$ statistic

The following statistics can be suppressed with the SHORT option:

- Total Canonical Structure, giving total-sample correlations between the canonical variables and the original variables
- Between Canonical Structure, giving between-class correlations between the canonical variables and the original variables
- Pooled Within Canonical Structure, giving pooled within-class correlations between the canonical variables and the original variables
- Total-Sample Standardized Canonical Coefficients, standardized to give canonical variables with zero mean and unit pooled within-class variance when applied to the total-sample standardized variables

- Pooled Within-Class Standardized Canonical Coefficients, standardized to give canonical variables with zero mean and unit pooled within-class variance when applied to the pooled within-class standardized variables
- Raw Canonical Coefficients, standardized to give canonical variables with zero mean and unit pooled within-class variance when applied to the centered variables
- Class Means on Canonical Variables

## ODS Table Names

PROC CANDISC assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in the following table. For more information on ODS, see Chapter 15, "Using the Output Delivery System."

**Table 21.2.** ODS Tables Produced in PROC CANDISC

| ODS Table Name | Description | PROC CANDISC Option |
|---|---|---|
| ANOVA | Univariate statistics | ANOVA |
| AveRSquare | Average R-square | ANOVA |
| BCorr | Between-class correlations | BCORR |
| BCov | Between-class covariances | BCOV |
| BSSCP | Between-class SSCP matrix | BSSCP |
| BStruc | Between canonical structure | default |
| CanCorr | Canonical correlations | default |
| CanonicalMeans | Class means on canonical variables | default |
| Counts | Number of observations, variables, classes, df | default |
| CovDF | DF for covariance matrices, not printed | any *COV option |
| Dist | Squared distances | MAHALANOBIS |
| DistFValues | $F$ statistics based on squared distances | MAHALANOBIS |
| DistProb | Probabilities for $F$ statistics from squared distances | MAHALANOBIS |
| Levels | Class level information | default |
| MultStat | MANOVA | default |
| PCoef | Pooled standard canonical coefficients | default |
| PCorr | Pooled within-class correlations | PCORR |
| PCov | Pooled within-class covariances | PCOV |
| PSSCP | Pooled within-class SSCP matrix | PSSCP |
| PStdMeans | Pooled standardized class means | STDMEAN |
| PStruc | Pooled within canonical structure | default |
| RCoef | Raw canonical coefficients | default |
| SimpleStatistics | Simple statistics | SIMPLE |
| TCoef | Total-sample standard canonical coefficients | default |
| TCorr | Total-sample correlations | TCORR |
| TCov | Total-sample covariances | TCOV |
| TSSCP | Total-sample SSCP matrix | TSSCP |

**Table 21.2.** (continued)

| ODS Table Name | Description | PROC CANDISC Option |
|---|---|---|
| TStdMeans | Total standardized class means | STDMEAN |
| TStruc | Total canonical structure | default |
| WCorr | Within-class correlations | WCORR |
| WCov | Within-class covariances | WCOV |
| WSSCP | Within-class SSCP matrices | WSSCP |

# Example

## Example 21.1. Analysis of Iris Data Using PROC CANDISC

The iris data published by Fisher (1936) have been widely used for examples in discriminant analysis and cluster analysis. The sepal length, sepal width, petal length, and petal width are measured in millimeters on fifty iris specimens from each of three species: *Iris setosa, I. versicolor, and I. virginica*.

This example is a canonical discriminant analysis that creates an output data set containing scores on the canonical variables and plots the canonical variables. The following statements produce Output 21.1.1 through Output 21.1.7:

```
proc format;
   value specname
      1=’Setosa    ’
      2=’Versicolor’
      3=’Virginica ’;
run;

data iris;
   title ’Fisher (1936) Iris Data’;
   input SepalLength SepalWidth PetalLength PetalWidth
         Species @@;
   format Species specname.;
   label SepalLength=’Sepal Length in mm.’
         SepalWidth =’Sepal Width in mm.’
         PetalLength=’Petal Length in mm.’
         PetalWidth =’Petal Width in mm.’;
   symbol = put(Species, specname10.);
   datalines;
50 33 14 02 1 64 28 56 22 3 65 28 46 15 2 67 31 56 24 3
63 28 51 15 3 46 34 14 03 1 69 31 51 23 3 62 22 45 15 2
59 32 48 18 2 46 36 10 02 1 61 30 46 14 2 60 27 51 16 2
65 30 52 20 3 56 25 39 11 2 65 30 55 18 3 58 27 51 19 3
68 32 59 23 3 51 33 17 05 1 57 28 45 13 2 62 34 54 23 3
77 38 67 22 3 63 33 47 16 2 67 33 57 25 3 76 30 66 21 3
49 25 45 17 3 55 35 13 02 1 67 30 52 23 3 70 32 47 14 2
64 32 45 15 2 61 28 40 13 2 48 31 16 02 1 59 30 51 18 3
55 24 38 11 2 63 25 50 19 3 64 32 53 23 3 52 34 14 02 1
49 36 14 01 1 54 30 45 15 2 79 38 64 20 3 44 32 13 02 1
67 33 57 21 3 50 35 16 06 1 58 26 40 12 2 44 30 13 02 1
```

```
77 28 67 20 3 63 27 49 18 3 47 32 16 02 1 55 26 44 12 2
50 23 33 10 2 72 32 60 18 3 48 30 14 03 1 51 38 16 02 1
61 30 49 18 3 48 34 19 02 1 50 30 16 02 1 50 32 12 02 1
61 26 56 14 3 64 28 56 21 3 43 30 11 01 1 58 40 12 02 1
51 38 19 04 1 67 31 44 14 2 62 28 48 18 3 49 30 14 02 1
51 35 14 02 1 56 30 45 15 2 58 27 41 10 2 50 34 16 04 1
46 32 14 02 1 60 29 45 15 2 57 26 35 10 2 57 44 15 04 1
50 36 14 02 1 77 30 61 23 3 63 34 56 24 3 58 27 51 19 3
57 29 42 13 2 72 30 58 16 3 54 34 15 04 1 52 41 15 01 1
71 30 59 21 3 64 31 55 18 3 60 30 48 18 3 63 29 56 18 3
49 24 33 10 2 56 27 42 13 2 57 30 42 12 2 55 42 14 02 1
49 31 15 02 1 77 26 69 23 3 60 22 50 15 3 54 39 17 04 1
66 29 46 13 2 52 27 39 14 2 60 34 45 16 2 50 34 15 02 1
44 29 14 02 1 50 20 35 10 2 55 24 37 10 2 58 27 39 12 2
47 32 13 02 1 46 31 15 02 1 69 32 57 23 3 62 29 43 13 2
74 28 61 19 3 59 30 42 15 2 51 34 15 02 1 50 35 13 03 1
56 28 49 20 3 60 22 40 10 2 73 29 63 18 3 67 25 58 18 3
49 31 15 01 1 67 31 47 15 2 63 23 44 13 2 54 37 15 02 1
56 30 41 13 2 63 25 49 15 2 61 28 47 12 2 64 29 43 13 2
51 25 30 11 2 57 28 41 13 2 65 30 58 22 3 69 31 54 21 3
54 39 13 04 1 51 35 14 03 1 72 36 61 25 3 65 32 51 20 3
61 29 47 14 2 56 29 36 13 2 69 31 49 15 2 64 27 53 19 3
68 30 55 21 3 55 25 40 13 2 48 34 16 02 1 48 30 14 01 1
45 23 13 03 1 57 25 50 20 3 57 38 17 03 1 51 38 15 03 1
55 23 40 13 2 66 30 44 14 2 68 28 48 14 2 54 34 17 02 1
51 37 15 04 1 52 35 15 02 1 58 28 51 24 3 67 30 50 17 2
63 33 60 25 3 53 37 15 02 1
;
proc candisc data=iris out=outcan distance anova;
   class Species;
   var SepalLength SepalWidth PetalLength PetalWidth;
run;
```

PROC CANDISC first displays information about the observations and the classes in the data set in Output 21.1.1.

**Output 21.1.1.** Iris Data: Summary Information

```
                                  Fisher (1936) Iris Data

                                   The CANDISC Procedure

                  Observations     150          DF Total               149
                  Variables          4          DF Within Classes      147
                  Classes            3          DF Between Classes       2


                                  Class Level Information

                         Variable
              Species    Name          Frequency       Weight    Proportion

              Setosa     Setosa               50      50.0000      0.333333
              Versicolor Versicolor           50      50.0000      0.333333
              Virginica  Virginica            50      50.0000      0.333333
```

The DISTANCE option in the PROC CANDISC statement displays squared Mahalanobis distances between class means. Results from the DISTANCE option is shown in Output 21.1.2 and Output 21.1.3.

**Output 21.1.2.** Iris Data: Squared Mahalanobis Distances

```
                         Fisher (1936) Iris Data

                          The CANDISC Procedure

                 Pairwise Squared Distances Between Groups

                  2          _   _          -1  _   _
                 D (i|j) = (X - X )' COV   (X - X )
                             i   j           i   j


                       Squared Distance to Species

            From
            Species           Setosa    Versicolor     Virginica

            Setosa                 0      89.86419     179.38471
            Versicolor      89.86419             0      17.20107
            Virginica      179.38471      17.20107             0
```

**Output 21.1.3.** Iris Data: Squared Mahalanobis Distance Statistics

```
                         Fisher (1936) Iris Data

                          The CANDISC Procedure

                 Pairwise Squared Distances Between Groups

                  2          _   _          -1  _   _
                 D (i|j) = (X - X )' COV   (X - X )
                             i   j           i   j


        F Statistics, NDF=4, DDF=144 for Squared Distance to Species

            From
            Species           Setosa    Versicolor     Virginica

            Setosa                 0     550.18889          1098
            Versicolor     550.18889             0     105.31265
            Virginica           1098     105.31265             0


         Prob > Mahalanobis Distance for Squared Distance to Species

            From
            Species           Setosa    Versicolor     Virginica

            Setosa            1.0000        <.0001        <.0001
            Versicolor        <.0001        1.0000        <.0001
            Virginica         <.0001        <.0001        1.0000
```

The ANOVA option specifies testing of the hypothesis that the class means are equal using univariate statistics. The resulting $R^2$ values (see Output 21.1.4) range from 0.4008 for SepalWidth to 0.9414 for PetalLength, and each variable is significant at the 0.0001 level. The multivariate test for differences between the classes (which is displayed by default) is also significant at the 0.0001 level; you would expect this from the highly significant univariate test results.

**Output 21.1.4.** Iris Data: Univariate and Multivariate Statistics

```
                                    Fisher (1936) Iris Data

                                     The CANDISC Procedure

                                  Univariate Test Statistics

                            F Statistics,    Num DF=2,   Den DF=147

                                    Total    Pooled   Between
                                 Standard  Standard  Standard           R-Square
Variable    Label               Deviation Deviation Deviation R-Square / (1-RSq) F Value Pr > F

SepalLength Sepal Length in mm.    8.2807    5.1479    7.9506   0.6187    1.6226  119.26 <.0001
SepalWidth  Sepal Width in mm.     4.3587    3.3969    3.3682   0.4008    0.6688   49.16 <.0001
PetalLength Petal Length in mm.   17.6530    4.3033   20.9070   0.9414   16.0566 1180.16 <.0001
PetalWidth  Petal Width in mm.     7.6224    2.0465    8.9673   0.9289   13.0613  960.01 <.0001


                                       Average R-Square

                               Unweighted              0.7224358
                               Weighted by Variance    0.8689444


                         Multivariate Statistics and F Approximations

                                     S=2    M=0.5    N=71

        Statistic                        Value    F Value    Num DF    Den DF    Pr > F

        Wilks' Lambda               0.02343863     199.15         8       288    <.0001
        Pillai's Trace              1.19189883      53.47         8       290    <.0001
        Hotelling-Lawley Trace     32.47732024     582.20         8     203.4    <.0001
        Roy's Greatest Root        32.19192920    1166.96         4       145    <.0001

                  NOTE: F Statistic for Roy's Greatest Root is an upper bound.
                         NOTE: F Statistic for Wilks' Lambda is exact.
```

The $R^2$ between Can1 and the class variable, 0.969872, is much larger than the corresponding $R^2$ for Can2, 0.222027. This is displayed in Output 21.1.5.

**Output 21.1.5.** Iris Data: Canonical Correlations and Eigenvalues

```
                                    Fisher (1936) Iris Data

                                     The CANDISC Procedure

                                         Adjusted    Approximate        Squared
                           Canonical     Canonical       Standard      Canonical
                         Correlation   Correlation          Error    Correlation

                       1    0.984821      0.984508       0.002468       0.969872
                       2    0.471197      0.461445       0.063734       0.222027

                                                      Test of H0: The canonical correlations in
                                                               the current row and all
                  Eigenvalues of Inv(E)*H                        that follow are zero
                     = CanRsq/(1-CanRsq)
                                                     Likelihood Approximate
   Eigenvalue Difference Proportion Cumulative      Ratio     F Value Num DF Den DF Pr > F

 1    32.1919    31.9065     0.9912     0.9912 0.02343863      199.15      8    288 <.0001
 2     0.2854                0.0088     1.0000 0.77797337       13.79      3    145 <.0001
```

**Output 21.1.6.** Iris Data: Correlations Between Canonical and Original Variables

```
                              Fisher (1936) Iris Data

                               The CANDISC Procedure

                             Total Canonical Structure

Variable         Label                            Can1              Can2

SepalLength      Sepal Length in mm.          0.791888          0.217593
SepalWidth       Sepal Width in mm.          -0.530759          0.757989
PetalLength      Petal Length in mm.          0.984951          0.046037
PetalWidth       Petal Width in mm.           0.972812          0.222902


                            Between Canonical Structure

Variable         Label                            Can1              Can2

SepalLength      Sepal Length in mm.          0.991468          0.130348
SepalWidth       Sepal Width in mm.          -0.825658          0.564171
PetalLength      Petal Length in mm.          0.999750          0.022358
PetalWidth       Petal Width in mm.           0.994044          0.108977


                         Pooled Within Canonical Structure

Variable         Label                            Can1              Can2

SepalLength      Sepal Length in mm.          0.222596          0.310812
SepalWidth       Sepal Width in mm.          -0.119012          0.863681
PetalLength      Petal Length in mm.          0.706065          0.167701
PetalWidth       Petal Width in mm.           0.633178          0.737242
```

The raw canonical coefficients (shown in Output 21.1.7) for the first canonical variable, Can1, show that the classes differ most widely on the linear combination of the centered variables $-0.0829378 \times$ SepalLength $- 0.153447 \times$ SepalWidth $+ 0.220121 \times$ PetalLength $+ 0.281046 \times$ PetalWidth.

**Output 21.1.7.** Iris Data: Canonical Coefficients

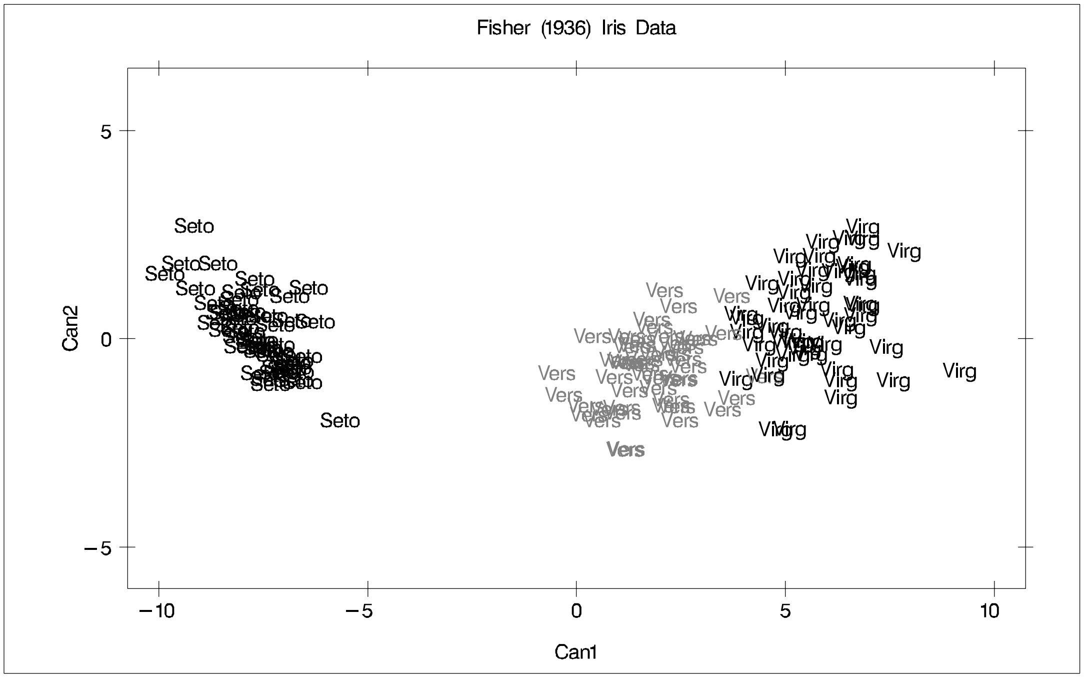
```
                              Fisher (1936) Iris Data

                               The CANDISC Procedure

                   Total-Sample Standardized Canonical Coefficients

Variable         Label                            Can1              Can2

SepalLength      Sepal Length in mm.      -0.686779533       0.019958173
SepalWidth       Sepal Width in mm.       -0.668825075       0.943441829
PetalLength      Petal Length in mm.       3.885795047      -1.645118866
PetalWidth       Petal Width in mm.        2.142238715       2.164135931


                 Pooled Within-Class Standardized Canonical Coefficients

Variable         Label                            Can1              Can2

SepalLength      Sepal Length in mm.      -.4269548486      0.0124075316
SepalWidth       Sepal Width in mm.       -.5212416758      0.7352613085
PetalLength      Petal Length in mm.      0.9472572487      -.4010378190
PetalWidth       Petal Width in mm.       0.5751607719      0.5810398645
```

```
                                 Fisher (1936) Iris Data

                                 The CANDISC Procedure

                               Raw Canonical Coefficients

      Variable          Label                           Can1              Can2

      SepalLength       Sepal Length in mm.      -.0829377642      0.0024102149
      SepalWidth        Sepal Width in mm.       -.1534473068      0.2164521235
      PetalLength       Petal Length in mm.      0.2201211656      -.0931921210
      PetalWidth        Petal Width in mm.       0.2810460309      0.2839187853


                            Class Means on Canonical Variables

                   Species                 Can1              Can2

                   Setosa          -7.607599927       0.215133017
                   Versicolor       1.825049490      -0.727899622
                   Virginica        5.782550437       0.512766605
```

The plot of canonical variables in Output 21.1.8 shows that of the two canonical variables Can1 has the most discriminatory power. The following invocation of the %PLOTIT macro creates this plot:

```
   %plotit(data=outcan, plotvars=Can2 Can1,
           labelvar=_blank_, symvar=symbol, typevar=symbol,
           symsize=1, symlen=4, exttypes=symbol, ls=100,
           tsize=1.5, extend=close);
```

**Output 21.1.8.** Iris Data: Plot of First Two Canonical Variables



# References

Fisher, R.A. (1936), “The Use of Multiple Measurements in Taxonomic Problems,” *Annals of Eugenics*, 7, 179–188.

Kshirsagar, A.M. (1972), *Multivariate Analysis*, New York: Marcel Dekker, Inc.

Lawley, D.N. (1959), “Tests of Significance in Canonical Analysis,” *Biometrika*, 46, 59–66.

Rao, C.R. (1973), *Linear Statistical Inference*, New York: John Wiley & Sons, Inc.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

**SAS/STAT® User's Guide, Version 8**

ISBN 1–58025–494–2

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999

The Institute is a private company devoted to the support and further development of its software and related services.