Chapter 40
The DISTANCE Procedure

Overview: DISTANCE Procedure

The DISTANCE procedure computes various measures of distance, dissimilarity, or similarity between the observations (rows) of an input SAS data set, which can contain numeric or character variables, or both, depending on which proximity measure is used.

The proximity measures are stored as a lower triangular matrix or a square matrix (depending on the SHAPE= option) in an output data set that can then be used as input to the CLUSTER, MDS, and MODECLUS procedures.
The number of rows and columns in the output data set equals the number of observations in the input data set. If the input data set contains BY groups, an output matrix is computed for each BY group with the size determined by the maximum number of observations in any BY group.

The output data set is of type TYPE=DISTANCE or TYPE=SIMILAR, depending on the value of the METHOD= option. See the METHOD= option for more information about the association between the method and the output data set type.

Data set types do not persist when you copy or modify a data set. You must specify the TYPE= data set option for the new data set, as in the following example:

```
data dist2(type=distance);
   set dist;
run;
```

See the OUT= option for more information about data set type persistence.

PROC DISTANCE also provides various nonparametric and parametric methods for standardizing variables. Different variables can be standardized with different methods.

Distance matrices are used frequently in data mining, genomics, marketing, financial analysis, management science, education, chemistry, psychology, biology, and various other fields.

---

**Levels of Measurement**

*Measurement* of some attribute of a set of objects is the process of assigning numbers or other symbols to the objects in such a way that properties of the numbers or symbols reflect properties of the attribute being measured. There are different *levels* of measurement that involve different properties (relations and operations) of the numbers or symbols. Associated with each level of measurement is a set of transformations of the measurements that preserve the relevant properties; these transformations are called permissible transformations. A particular way of assigning numbers or symbols to measure something is called a *scale* of measurement.

The most commonly discussed levels of measurement are as follows:

- **Nominal**
  Two objects are assigned the same symbol if they have the same value of the attribute. Permissible transformations are any one-to-one or many-to-one transformation, although a many-to-one transformation loses information.

- **Ordinal**
  Objects are assigned numbers such that the order of the numbers reflects an order relation defined on the attribute. Two objects \(x\) and \(y\) with attribute values \(a(x)\) and \(a(y)\) are assigned numbers \(m(x)\) and \(m(y)\) such that if \(m(x) > m(y)\), then \(a(x) > a(y)\). Permissible transformations are any monotone increasing transformation, although a transformation that is not strictly increasing loses information.

- **Interval**
  Objects are assigned numbers such that differences between the numbers reflect differences of the attribute. If \(m(x) - m(y) > m(u) - m(v)\), then \(a(x) - a(y) > a(u) - a(v)\). Permissible transformations are any affine transformation \(t(m) = c \times m + d\), where \(c\) and \(d\) are constants; another way of saying this is that the origin and unit of measurement are arbitrary.
Log-interval

Objects are assigned numbers such that ratios between the numbers reflect ratios of the attribute. If \( m(x)/m(y) > m(u)/m(v) \), then \( a(x)/a(y) > a(u)/a(v) \). Permissible transformations are any power transformation \( t(m) = c \cdot m^d \), where \( c \) and \( d \) are constants.

Ratio

Objects are assigned numbers such that differences and ratios between the numbers reflect differences and ratios of the attribute. Permissible transformations are any linear (similarity) transformation \( t(m) = c \cdot m \), where \( c \) is a constant; another way of saying this is that the unit of measurement is arbitrary.

Absolute

Objects are assigned numbers such that all properties of the numbers reflect analogous properties of the attribute. The only permissible transformation is the identity transformation.

Proximity measures provided in the DISTANCE procedure accept four levels of measurement: nominal, ordinal, interval, and ratio. Ordinal variables are transformed to interval variables before processing. This is done by replacing the data with their rank scores, and by assuming that the classes of an ordinal variable are spaced equally along the interval scale. See the RANKSCORE= option in the section “PROC DISTANCE Statement” on page 2680 for choices on assigning scores to ordinal variables. There are also different approaches for how to transform an ordinal variable to an interval variable. See Anderberg (1973) for alternatives.

Symmetric versus Asymmetric Nominal Variables

A binary variable contains two possible outcomes: 1 (positive/present) or 0 (negative/absent). If there is no preference for which outcome should be coded as 0 and which as 1, the binary variable is called symmetric. For example, the binary variable “is evergreen?” for a plant has the possible states “loses leaves in winter” and “does not lose leaves in winter.” Both are equally valuable and carry the same weight when a proximity measure is computed. Commonly used measures that accept symmetric binary variables include the Simple Matching, Hamann, Roger and Tanimoto, Sokal and Sneath 1, and Sokal and Sneath 3 coefficients.

If the outcomes of a binary variable are not equally important, the binary variable is called asymmetric. An example of such a variable is the presence or absence of a relatively rare attribute, such as “is color-blind” for a human being. While you say that two people who are color-blind have something in common, you cannot say that people who are not color-blind have something in common. The most important outcome is usually coded as 1 (present) and the other is coded as 0 (absent). The agreement of two 1’s (a present-present match or a positive match) is more significant than the agreement of two 0’s (an absent-absent match or a negative match). Usually, the negative match is treated as irrelevant. Commonly used measures that accept asymmetric binary variables include Jaccard, Dice, Russell and Rao, Binary Lance and Williams nonmetric, and Kulcynski coefficients.

When nominal variables are employed, the comparison of one data unit with another can only be in terms of whether the data units score the same or different on the variables. If a variable is defined as an asymmetric nominal variable and two data units score the same but fall into the absent category, the absent-absent match is excluded from the computation of the proximity measure.
Standardization

Since variables with large variances tend to have more effect on the proximity measure than those with small variances, it is recommended that you standardize the variables before the computation of the proximity measure. The DISTANCE procedure provides a convenient way to standardize each variable with its own method before the proximity measures are computed. You can also perform the standardization by using the STDIZE procedure, with the limitation that all variables must be standardized with the same method.

Mandatory Standardization

Variable standardization is not required if any of the following conditions is true:

- if there is only one level of measurement
- if only asymmetric nominal and nominal levels are specified
- if the NOSTD option is specified in the PROC DISTANCE statement
- if METHOD=JACCARD or METHOD=DJACCARD is specified

Otherwise, standardization is mandatory.

When standardization is mandatory and no standardization method is specified, a default method of standardization will be used. This default method is determined by the measurement level. In general, the default method is STD for interval variables except when METHOD=GOWER or METHOD=DGOWER is specified. See the STD= option in the section “VAR Statement” on page 2686 for the default methods for GOWER and DGOWER as well as methods available for standardizing variables.

When standardization is mandatory, PROC DISTANCE ignores the REPONLY option, if it is specified.

Getting Started: DISTANCE Procedure

Creating a Distance Matrix as Input for a Subsequent Cluster Analysis

The following example demonstrates how you can use the DISTANCE procedure to obtain a distance matrix that will be used as input to a subsequent clustering procedure.
The following data, originated by A. Weber and cited in Hand et al. (1994, p. 297), measure the amount of protein consumed for nine food groups in 25 European countries. The nine food groups are red meat (RedMeat), white meat (WhiteMeat), eggs (Eggs), milk (Milk), fish (Fish), cereal (Cereal), starch (Starch), nuts (Nuts), and fruits and vegetables (FruitVeg). Suppose you want to determine whether national figures in protein consumption can be used to determine certain types or categories of countries; specifically, you want to perform a cluster analysis to determine whether these 25 countries can be formed into groups suggested by the data.

The following DATA step creates the SAS data set Protein:

```sas
data Protein;
  length Country $ 14;
  input Country &$ RedMeat WhiteMeat Eggs Milk Fish Cereal Starch Nuts FruitVeg;
  datalines;
  Albania 10.1 1.4 0.5 8.9 0.2 42.3 0.6 5.5 1.7
  Austria 8.9 14.0 4.3 19.9 2.1 28.0 3.6 1.3 4.3
  Belgium 13.5 9.3 4.1 17.5 4.5 26.6 5.7 2.1 4.0
  Bulgaria 7.8 6.0 1.6 8.3 1.2 56.7 1.1 3.7 4.2
  Czechoslovakia 9.7 11.4 2.8 12.5 2.0 34.3 5.0 1.1 4.0
  Denmark 10.6 10.8 3.7 25.0 9.9 21.9 4.8 0.7 2.4
  E Germany 8.4 11.6 3.7 11.1 5.4 24.6 6.5 0.8 3.6
  Finland 9.5 4.9 2.7 33.7 5.8 26.3 5.1 1.0 1.4
  France 18.0 9.9 3.3 19.5 5.7 28.1 4.8 2.4 6.5
  Greece 10.2 3.0 2.8 17.6 5.9 41.7 2.2 7.8 6.5
  Hungary 5.3 12.4 2.9 9.7 0.3 40.1 4.0 5.4 4.2
  Ireland 13.9 10.0 4.7 25.8 2.2 24.0 6.2 1.6 2.9
  Italy 9.0 5.1 2.9 13.7 3.4 36.8 2.1 4.3 6.7
  Netherlands 9.5 13.6 3.6 23.4 2.5 22.4 4.2 1.8 3.7
  Norway 9.4 4.7 2.7 23.3 9.7 23.0 4.6 1.6 2.7
  Poland 6.9 10.2 2.7 19.3 3.0 36.1 5.9 2.0 6.6
  Portugal 6.2 3.7 1.1 4.9 14.2 27.0 5.9 4.7 7.9
  Romania 6.2 6.3 1.5 11.1 1.0 49.6 3.1 5.3 2.8
  Spain 7.1 3.4 3.1 8.6 7.0 29.2 5.7 5.9 7.2
  Sweden 9.9 7.8 3.5 4.7 7.5 19.5 3.7 1.4 2.0
  Switzerland 13.1 10.1 3.1 23.8 2.3 25.6 2.8 2.4 4.9
  UK 17.4 5.7 4.7 20.6 4.3 24.3 4.7 3.4 3.3
  USSR 9.3 4.6 2.1 16.6 3.0 43.6 6.4 3.4 2.9
  W Germany 11.4 12.5 4.1 18.8 3.4 18.6 5.2 1.5 3.8
  Yugoslavia 4.4 5.0 1.2 9.5 0.6 55.9 3.0 5.7 3.2
;```

The data set Protein contains the character variable Country and the nine numeric variables representing the food groups. The $14. in the INPUT statement specifies that the variable Country has a length of 14.
Chapter 40: The DISTANCE Procedure

The following statements create the distance matrix and display part of it:

```sas
title 'Protein Consumption in Europe';
proc distance data=Protein out=Dist method=Euclid;
   var interval(RedMeat--FruitVeg / std=Std);
   id Country;
run;
proc print data=Dist(obs=10);
   title2 'First 10 Observations in Output Data Set from PROC DISTANCE';
run;
```

An output SAS data set called Dist, which contains the distance matrix, is created through the OUT= option. The METHOD=EUCLID option requests that Euclidean distances (which is the default) should be computed and produces an output data set of TYPE=DISTANCE.¹

The VAR statement lists the variables (RedMeat—FruitVeg) along with their measurement level to be used in the analysis. An interval level of measurement is assigned to those variables. Since variables with large variances tend to have more effect on the proximity measure than those with small variances, each variable is standardized by the STD method to have a mean of 0 and a standard deviation of 1. This is done by adding "/STD=STD" at the end of the variables list.

The ID statement specifies that the variable Country should be copied to the OUT= data set and used to generate names for the distance variables. The distance variables in the output data set are named by the values in the ID variable, and the maximum length for the names of these variables is 14.

There are 25 observations in the input data set; therefore, the output data set Dist contains a 25-by-25 lower triangular matrix.

The PROC PRINT statement displays the first 10 observations in the output data set Dist as shown in Figure 40.1.

¹Data set types do not persist when you copy or modify a data set. You must specify the TYPE= data set option for the new data set. See the METHOD= and OUT= options for more information about data set types.
Creating a Distance Matrix as Input for a Subsequent Cluster Analysis

Figure 40.1 First 10 Observations in the Output Data Set from PROC DISTANCE

Protein Consumption in Europe
First 10 Observations in Output Data Set from PROC DISTANCE

<table>
<thead>
<tr>
<th>Obs</th>
<th>Country</th>
<th>Albania</th>
<th>Austria</th>
<th>Belgium</th>
<th>Bulgaria</th>
<th>Czechoslovakia</th>
<th>Denmark</th>
<th>E_Germany</th>
<th>Finland</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Albania</td>
<td>0.00000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>Austria</td>
<td>6.12388</td>
<td>0.00000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>Belgium</td>
<td>5.94109</td>
<td>2.44987</td>
<td>0.00000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>4</td>
<td>Bulgaria</td>
<td>2.76446</td>
<td>4.88331</td>
<td>5.22711</td>
<td>0.00000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>5</td>
<td>Czechoslovakia</td>
<td>5.13959</td>
<td>2.11498</td>
<td>2.21330</td>
<td>3.94761</td>
<td>0.00000</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>6</td>
<td>Denmark</td>
<td>6.61002</td>
<td>3.01392</td>
<td>2.52541</td>
<td>6.00803</td>
<td>3.34049</td>
<td>0.00000</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>7</td>
<td>E_Germany</td>
<td>6.39178</td>
<td>2.56341</td>
<td>2.10211</td>
<td>5.40824</td>
<td>1.87962</td>
<td>2.72112</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>8</td>
<td>Finland</td>
<td>5.81458</td>
<td>4.04271</td>
<td>3.45779</td>
<td>5.74882</td>
<td>3.91378</td>
<td>2.61570</td>
<td>3.99426</td>
<td>0.00000</td>
</tr>
<tr>
<td>10</td>
<td>Greece</td>
<td>4.24495</td>
<td>5.16330</td>
<td>4.69515</td>
<td>3.74849</td>
<td>5.59084</td>
<td>5.61496</td>
<td>5.47453</td>
<td>.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs</th>
<th>France</th>
<th>Greece</th>
<th>Hungary</th>
<th>Ireland</th>
<th>Italy</th>
<th>Netherlands</th>
<th>Norway</th>
<th>Poland</th>
<th>Portugal</th>
<th>Romania</th>
<th>Spain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>4</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>5</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>6</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>7</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>8</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>9</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>10</td>
<td>0.00000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs</th>
<th>Sweden</th>
<th>Switzerland</th>
<th>UK</th>
<th>USSR</th>
<th>W_Germany</th>
<th>Yugoslavia</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>4</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>5</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>6</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>7</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>8</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>9</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>10</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

The following statements produce the dendrogram in Figure 40.2:

```plaintext
ods graphics on;
proc cluster data=Dist method=Ward plots=dendrogram(height=rsq);
    id Country;
run;
```
The CLUSTER procedure performs a Ward’s minimum-variance cluster analysis based on the distance matrix created by PROC DISTANCE. PROC CLUSTER, along with ODS Graphics, produces the dendrogram shown in Figure 40.2. The option PLOTS=DENDROGRAM(HEIGHT=RSQ) specifies the squared multiple correlation as the height variable in the dendrogram.

**Figure 40.2** Dendrogram of R Square
After inspecting the dendrogram in Figure 40.2, you will see that when the countries are grouped into six clusters, the proportion of variance accounted for by these clusters is slightly less than 70% (69.3%). The 25 countries are clustered as follows:

- Balkan countries: Albania, Bulgaria, Romania, and Yugoslavia
- Mediterranean countries: Greece and Italy
- Iberian countries: Portugal and Spain
- Western European countries: Austria, Netherlands, Switzerland, Belgium, former West Germany, Ireland, France, and U.K.
- Scandinavian countries: Denmark, Norway, Finland, and Sweden
- Eastern European countries: former Czechoslovakia, former East Germany, Poland, former U.S.S.R., and Hungary

Syntax: DISTANCE Procedure

The following statements are available in the DISTANCE procedure:

```plaintext
PROC DISTANCE <options> ;
    BY variables ;
    COPY variables ;
    FREQ variable ;
    ID variable ;
    VAR level (variables </ options>) ;
    WEIGHT variable ;
```

Both the PROC DISTANCE statement and the VAR statement are required.
PROC DISTANCE Statement

PROC DISTANCE <options> ;

The PROC DISTANCE statement invokes the DISTANCE procedure. Table 40.1 summarizes the options available in the PROC DISTANCE statement. These options are discussed in the following section.

Table 40.1 Summary of PROC DISTANCE Statement Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardize variables</td>
<td></td>
</tr>
<tr>
<td>ADD=</td>
<td>Specifies the constant to add to each value after standardizing and multiplying by the value specified in the MULT= option</td>
</tr>
<tr>
<td>FUZZ=</td>
<td>Specifies the relative fuzz factor for writing the output</td>
</tr>
<tr>
<td>INITIAL=</td>
<td>Specifies the method for computing initial estimates for the A-estimates</td>
</tr>
<tr>
<td>MULT=</td>
<td>Specifies the constant to multiply each value by after standardizing</td>
</tr>
<tr>
<td>NORM</td>
<td>Normalizes the scale estimator to be consistent for the standard deviation of a normal distribution</td>
</tr>
<tr>
<td>NOSTD</td>
<td>Suppresses standardization</td>
</tr>
<tr>
<td>SNORM</td>
<td>Normalizes the scale estimator to have an expectation of approximately 1 for a standard normal distribution</td>
</tr>
<tr>
<td>STDONLY</td>
<td>Standardizes variables only (suppresses computation of the distance matrix)</td>
</tr>
<tr>
<td>VARDEF=</td>
<td>Specifies the variances divisor</td>
</tr>
<tr>
<td>Generate distance matrix</td>
<td></td>
</tr>
<tr>
<td>ABSENT=</td>
<td>Specifies the value to be used as an absence value for all the asymmetric nominal variables</td>
</tr>
<tr>
<td>METHOD=</td>
<td>Specifies the method for computing proximity measures</td>
</tr>
<tr>
<td>PREFIX=</td>
<td>Specifies a prefix for naming the distance variables in the OUT= data set</td>
</tr>
<tr>
<td>RANKSCORE=</td>
<td>Specifies the method of assigning scores to ordinal variables</td>
</tr>
<tr>
<td>SHAPE=</td>
<td>Specifies the shape of the proximity matrix to be stored in the OUT= data set</td>
</tr>
<tr>
<td>UNDEF=</td>
<td>Specifies the numeric constant used to replace undefined distances</td>
</tr>
<tr>
<td>Replace missing values</td>
<td></td>
</tr>
<tr>
<td>NOMISS</td>
<td>Omits observations with missing values from computation of the location and scale measures, if standardization applies; outputs missing values to the distance matrix for observations with missing values</td>
</tr>
<tr>
<td>REPLACE</td>
<td>Replaces missing data with zero in the standardized data</td>
</tr>
<tr>
<td>REPONLY</td>
<td>Replaces missing data with the location measure (does not standardize the data)</td>
</tr>
<tr>
<td>Specify data set details</td>
<td></td>
</tr>
<tr>
<td>DATA=</td>
<td>Specifies the input data set</td>
</tr>
<tr>
<td>OUT=</td>
<td>Specifies the output data set</td>
</tr>
<tr>
<td>OUTSDZ=</td>
<td>Specifies the output data set for standardized scores</td>
</tr>
</tbody>
</table>
These options and their abbreviations are described (in alphabetical order) in the remainder of this section.

**ABSENT=**

ABSENT=number | qs

specifies the value to be used as an absence value in an irrelevant absent-absent match for all of the asymmetric nominal variables. If you want to specify a different absence value for a particular variable, use the ABSENT= option in the VAR statement. See the ABSENT= option in the section “VAR Statement” on page 2686 for details.

An absence value for a variable can be either a numeric value or a quoted string consisting of combinations of characters. For instance, ., -999, and “NA” are legal values for the ABSENT= option.

The default absence value for a character variable is “NONE” (notice that a blank value is considered a missing value), and the default absence value for a numeric variable is 0.

**ADD=**

ADD=c

specifies a constant, c, to add to each value after standardizing and multiplying by the value you specify in the MULT= option. The default value is 0.

**DATA=**

DATA=SAS-data-set

specifies the input data set containing observations from which the proximity is computed. If you omit the DATA= option, the most recently created SAS data set is used.

**FUZZ=**

FUZZ=c

specifies the relative fuzz factor for computing the standardized scores. The default value is 1E–14.

For the OUTSDZ= data set, the score is computed as follows:

\[
\text{if } |\text{standardized scores}| < m \times c, \text{ then standardized scores } = 0
\]

where m is the numeric constant specified in the MULT= option, or 1 if MULT= option is not specified.

**INITIAL=**

INITIAL=method

specifies the method of computing initial estimates for the A-estimates (ABW, AWAVE, and AHUBER). The following methods are not allowed for the INITIAL= option: ABW, AHUBER, AWAVE, and IN.

The default value is INITIAL=MAD.

**METHOD=**

METHOD=method

specifies the method of computing proximity measures.

For use in PROC CLUSTER, distance or dissimilarity measures such as METHOD=EUCLID or METHOD=DGOWER should be chosen.

The following six tables outline the proximity measures available for the METHOD= option. These tables are classified by levels of measurement accepted by each method. Each table contains four or five columns: the Method column shows the proximity measures, one or two Range columns show the upper and lower bounds, and the TYPE= column shows the type of proximity. The TYPE= column contains SIMILAR if a method generates similarity measures or DISTANCE if a method generates distance or dissimilarity measures. The output data set is of the type shown. For more information about the output data set, see the OUT= option.

For formulas and descriptions of these methods, see the section “Details: DISTANCE Procedure” on page 2693.
Table 40.2 shows the range and output matrix type of the GOWER and DGOWER methods. These two methods accept all measurement levels including ratio, interval, ordinal, nominal, and asymmetric nominal. METHOD=GOWER or METHOD=DGOWER always implies standardization. Assuming all the numeric (ordinal, interval, and ratio) variables are standardized by their corresponding default methods, the possible range values for both methods are from 0 and 1, inclusive. For more information about the default methods of standardization for METHOD=GOWER or METHOD=DGOWER, see the STD= option in the section “VAR Statement” on page 2686.

Table 40.2 Methods That Accept All Measurement Levels

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>Range</th>
<th>TYPE=</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOWER</td>
<td>Gower and Legendre (1986) similarity</td>
<td>0 to 1</td>
<td>SIMILAR</td>
</tr>
<tr>
<td>DGOWER</td>
<td>1 minus GOWER</td>
<td>0 to 1</td>
<td>DISTANCE</td>
</tr>
</tbody>
</table>

Table 40.3 shows methods that accept ratio, interval, and ordinal variables.

Table 40.3 Methods That Accept Ratio, Interval, and Ordinal Variables

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>Range</th>
<th>TYPE=</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUCLID</td>
<td>Euclidean distance</td>
<td>≥ 0</td>
<td>DISTANCE</td>
</tr>
<tr>
<td>SQUEUCLID</td>
<td>Squared Euclidean distance</td>
<td>≥ 0</td>
<td>DISTANCE</td>
</tr>
<tr>
<td>SIZE</td>
<td>Size distance</td>
<td>≥ 0</td>
<td>DISTANCE</td>
</tr>
<tr>
<td>SHAPE</td>
<td>Shape distance</td>
<td>≥ 0</td>
<td>DISTANCE</td>
</tr>
<tr>
<td>COV</td>
<td>Covariance</td>
<td>≥ 0</td>
<td>DISTANCE</td>
</tr>
<tr>
<td>CORR</td>
<td>Correlation</td>
<td>−1 to 1</td>
<td>SIMILAR</td>
</tr>
<tr>
<td>DCORR</td>
<td>Correlation transformed to Euclidean distance</td>
<td>0 to $\sqrt{2}$</td>
<td>DISTANCE</td>
</tr>
<tr>
<td>SQCORR</td>
<td>Squared correlation</td>
<td>0 to 1</td>
<td>SIMILAR</td>
</tr>
<tr>
<td>DSQCORR</td>
<td>One minus squared correlation</td>
<td>0 to 1</td>
<td>DISTANCE</td>
</tr>
<tr>
<td>$L(p)$</td>
<td>Minkowski ($L_p$) distance, where $p$ is a positive numeric value. Typical values of $p$ include 1 and 2. Very small or large values of $p$ might cause floating-point overflow.</td>
<td>&gt; 0</td>
<td>DISTANCE</td>
</tr>
<tr>
<td>CITYBLOCK</td>
<td>$L_1$, city-block, or Manhattan distance</td>
<td>≥ 0</td>
<td>DISTANCE</td>
</tr>
<tr>
<td>CHEBYCHEV</td>
<td>$L_\infty$</td>
<td>≥ 0</td>
<td>DISTANCE</td>
</tr>
<tr>
<td>POWER($p, r$)</td>
<td>Generalized Euclidean distance, where $p$ is a positive numeric value and $r$ is a nonnegative numeric value. The distance between two observations is the $r$th root of sum of the absolute differences to the $p$th power between the values for the observations.</td>
<td>≥ 0</td>
<td>DISTANCE</td>
</tr>
</tbody>
</table>

Table 40.4 shows methods that accept ratio variables. Notice that all possible range values are nonnegative, because ratio variables are assumed to be positive.
Table 40.4 Methods That Accept Ratio Variables

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>Range</th>
<th>TYPE=</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIMRATIO</td>
<td>Similarity ratio (if variables are binary, this is the Jaccard coefficient)</td>
<td>0 to 1</td>
<td>SIMILAR</td>
</tr>
<tr>
<td>DISRATIO</td>
<td>One minus similarity ratio</td>
<td>0 to 1</td>
<td>DISTANCE</td>
</tr>
<tr>
<td>NONMETRIC</td>
<td>Lance and Williams nonmetric coefficient</td>
<td>0 to 1</td>
<td>DISTANCE</td>
</tr>
<tr>
<td>CANBERRA</td>
<td>Canberra metric distance coefficient</td>
<td>0 to 1</td>
<td>DISTANCE</td>
</tr>
<tr>
<td>COSINE</td>
<td>Cosine coefficient</td>
<td>0 to 1</td>
<td>SIMILAR</td>
</tr>
<tr>
<td>DOT</td>
<td>Dot (inner) product coefficient</td>
<td>≥ 0</td>
<td>SIMILAR</td>
</tr>
<tr>
<td>OVERLAP</td>
<td>Overlap similarity</td>
<td>≥ 0</td>
<td>SIMILAR</td>
</tr>
<tr>
<td>DOVERLAP</td>
<td>Overlap dissimilarity</td>
<td>≥ 0</td>
<td>DISTANCE</td>
</tr>
<tr>
<td>CHISQ</td>
<td>Chi-square coefficient</td>
<td>0</td>
<td>DISTANCE</td>
</tr>
<tr>
<td>CHI</td>
<td>Square root of chi-square coefficient</td>
<td>≥ 0</td>
<td>DISTANCE</td>
</tr>
<tr>
<td>PHISQ</td>
<td>Phi-square coefficient</td>
<td>≥ 0</td>
<td>DISTANCE</td>
</tr>
<tr>
<td>PHI</td>
<td>Square root of phi-square coefficient</td>
<td>≥ 0</td>
<td>DISTANCE</td>
</tr>
</tbody>
</table>

Table 40.5 shows methods that accept nominal variables.

Table 40.5 Methods That Accept Nominal Variables

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>Range</th>
<th>TYPE=</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAMMING</td>
<td>Hamming distance</td>
<td>0 to 1</td>
<td>DISTANCE</td>
</tr>
<tr>
<td>MATCH</td>
<td>Simple matching coefficient</td>
<td>0 to 1</td>
<td>SIMILAR</td>
</tr>
<tr>
<td>DMATCH</td>
<td>Simple matching coefficient transformed to Euclidean distance</td>
<td>0 to 1</td>
<td>SIMILAR</td>
</tr>
<tr>
<td>DSQMATCH</td>
<td>Simple matching coefficient transformed to squared Euclidean distance</td>
<td>0 to 1</td>
<td>DISTANCE</td>
</tr>
<tr>
<td>HAMANN</td>
<td>Hamann coefficient</td>
<td>–1 to 1</td>
<td>SIMILAR</td>
</tr>
<tr>
<td>RT</td>
<td>Roger and Tanimoto</td>
<td>0 to 1</td>
<td>SIMILAR</td>
</tr>
<tr>
<td>SS1</td>
<td>Sokal and Sneath 1</td>
<td>0 to 1</td>
<td>SIMILAR</td>
</tr>
<tr>
<td>SS3</td>
<td>Sokal and Sneath 3</td>
<td>0 to 1</td>
<td>SIMILAR</td>
</tr>
</tbody>
</table>

Note that \( v \) denotes the number of variables (dimensionality).

Table 40.6 shows methods that accept asymmetric nominal variables. Use the ABSENT= option to create a value to be considered absent.

Table 40.6 Methods That Accept Asymmetric Nominal Variables

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>Range</th>
<th>TYPE=</th>
</tr>
</thead>
<tbody>
<tr>
<td>DICE</td>
<td>Dice coefficient or Czekanowski/Sorensen similarity coefficient</td>
<td>0 to 1</td>
<td>SIMILAR</td>
</tr>
<tr>
<td>RR</td>
<td>Russell and Rao</td>
<td>0 to 1</td>
<td>SIMILAR</td>
</tr>
<tr>
<td>BLWNM</td>
<td>Binary Lance and Williams nonmetric, or Bray-Curtis coefficient</td>
<td>0 to 1</td>
<td>DISTANCE</td>
</tr>
<tr>
<td>K1</td>
<td>Kulczynski 1</td>
<td>≥ 0</td>
<td>SIMILAR</td>
</tr>
</tbody>
</table>
Table 40.7 shows methods that accept asymmetric nominal and ratio variables. Use the ABSENT= option to create a value to be considered absent. The table contains five columns. The third column contains possible range values if only one level of measurement (either ratio or asymmetric nominal but not both) is specified; the fourth column contains possible range values if both levels are specified.

The JACCARD method is equivalent to the SIMRATIO method if there is no asymmetric nominal variable; if both ratio and asymmetric nominal variables are present, the coefficient is computed as the sum of the coefficient from the ratio variables and the coefficient from the asymmetric nominal variables. See “Proximity Measures” in the section “Details: DISTANCE Procedure” on page 2693 for the formula and descriptions of the JACCARD method.

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>Range (One Level)</th>
<th>Range (Two Levels)</th>
<th>TYPE=</th>
</tr>
</thead>
<tbody>
<tr>
<td>JACCARD</td>
<td>Jaccard similarity</td>
<td>0 to 1</td>
<td>0 to 2</td>
<td>SIMILAR</td>
</tr>
<tr>
<td></td>
<td>coefficient</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DJACCARD</td>
<td>Jaccard dissimilarity coefficient</td>
<td>0 to 1</td>
<td>0 to 2</td>
<td>DISTANCE</td>
</tr>
</tbody>
</table>

MULT=\(c\)

specifies a numeric constant, \(c\), by which to multiply each value after standardizing. The default value is 1.

NOMISS

omits observations with missing values from computation of the location and scale measures when standardizing; generates undefined (missing) distances for observations with missing values when computing distances. Use the UNDEF= option to specify the undefined values.

If a distance matrix is created to be used as an input to PROC CLUSTER, the NOMISS option should not be used because PROC CLUSTER does not accept distance matrices with missing values.

NORM

normalizes the scale estimator to be consistent for the standard deviation of a normal distribution when you specify the option STD=AGK, STD=IQR, STD=MAD, or STD=SPACING in the VAR statement.

NOSTD

suppresses standardization of the variables. The NOSTD option should not be specified with the STDONLY option or with the REPLACE option.

OUT=SAS-data-set

specifies the name of the SAS data set created by PROC DISTANCE. The output data set contains the BY variables, the ID variable, computed distance variables, the COPY variables, the FREQ variable, and the WEIGHT variables.

If you omit the OUT= option, PROC DISTANCE creates an output data set named according to the DATA\(n\) convention.

The output data set is of type TYPE=DISTANCE or TYPE=SIMILAR. See the METHOD= option for more information about the association between the method and the output data set type. Data set types do not persist when you copy or modify a data set. You must specify the TYPE= data set option for the new data set, as in the following example:
data dist2(type=distance);
  set dist;
run;

If you do not specify the TYPE=DISTANCE data set option, the new data set is the default TYPE=DATA. If you use the new data set in a procedure that accepts both TYPE=DATA or TYPE=DISTANCE data sets (such as PROC CLUSTER or PROC MODECLUS), the results will be incorrect.

OUTSDZ=SAS-data-set
specifies the name of the SAS data set containing the standardized scores. The output data set contains a copy of the DATA= data set, except that the analyzed variables have been standardized. Analyzed variables are those listed in the VAR statement.

PREFIX=name
specifies a prefix for naming the distance variables in the OUT= data set. By default, the names are Dist1, Dist2, \ldots, Distn. If you specify PREFIX=ABC, the variables are named ABC1, ABC2, \ldots, ABCn. If the ID statement is also specified, the variables are named by appending the value of the ID variable to the prefix.

RANKSCORE=MIDRANK | INDEX
specifies the method of assigning scores to ordinal variables. The available methods are listed as follows:

- MIDRANK assigns consecutive integers to each category with consideration of the frequency value. This is the default method.
- INDEX assigns consecutive integers to each category regardless of frequencies.

The following example explains how each method assigns the rank scores. Suppose the data contain an ordinal variable ABC with values A, B, C. There are two ways to assign numbers. One is to use midranks, which depend on the frequencies of each category. Another is to assign consecutive integers to each category, regardless of frequencies.

<table>
<thead>
<tr>
<th>ABC</th>
<th>MIDRANK</th>
<th>INDEX</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1.5</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>1.5</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
<td>3</td>
</tr>
</tbody>
</table>

REPLACE
replaces missing data with zero in the standardized data (to correspond to the location measure before standardizing). To replace missing data with something else, use the MISSING= option in the VAR statement. The REPLACE option implies standardization.

You cannot specify the following options together:
both the REPLACE and the REPONLY options
both the REPLACE and the NOSTD options

REPONLY
replaces missing data by the location measure that is specified by the MISSING= option or the STD= option (if the MISSING= option is not specified), but does not standardize the data. If the MISSING= option is not specified and METHOD=GOWER is specified, missing values are replaced by the location measure from the RANGE method (the minimum value), and the STD= option is.

You cannot specify both the REPLACE and the REPONLY options.

SHAPE=TRIANGLE | TRI | SQUARE | SQU | SQR
specifies the shape of the proximity matrix to be stored in the OUT= data set. SHAPE=TRIANGLE requests the matrix to be stored as a lower triangular matrix; SHAPE=SQUARE requests that the matrix be stored as a square matrix. Use SHAPE=SQUARE if the output data set is to be used as input to the MODECLUS procedures. The default is TRIANGLE.

SNORM
normalizes the scale estimator to have an expectation of approximately 1 for a standard normal distribution when the STD=SPACING option is specified.

STDONLY
standardizes variables only and computes no distance matrix. You must use the OUTSDZ= option to save the standardized scores. You cannot specify both the STDONLY option and the NOSTD option.

UNDEF=n
specifies the numeric constant used to replace undefined distances, such as when an observation has all missing values, or if a divisor is zero.

VARDEF=DF | N | WDF | WEIGHT | WGT
specifies the divisor to be used in the calculation of distance, dissimilarity, or similarity measures, and for standardizing variables whenever a variance or covariance is computed. By default, VARDEF=DF. The values and associated divisors are as follows:

<table>
<thead>
<tr>
<th>Value</th>
<th>Divisor</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>DF</td>
<td>degrees of freedom</td>
<td>$n - 1$</td>
</tr>
<tr>
<td>N</td>
<td>number of observations</td>
<td>$n$</td>
</tr>
<tr>
<td>WDF</td>
<td>sum of weights minus 1</td>
<td>$(\sum w_i) - 1$</td>
</tr>
<tr>
<td>WEIGHT</td>
<td>sum of weights</td>
<td>$\sum w_i$</td>
</tr>
</tbody>
</table>

VAR Statement

\[
\text{VAR level (variables } / \text{options})< \text{level (variables } / \text{options})> . . . ;
\]

where the syntax for the options is as follows:

ABSENT=value
MISSING=miss-method | value
ORDER=order-option
STD=std-method
WEIGHTS=weight-list
The VAR statement lists variables from which distances are to be computed. The VAR statement is required. The variables can be numeric or character depending on their measurement levels. A variable cannot appear more than once in either the same list or a different list.

`level` is required. It declares the levels of measurement for those variables specified within the parentheses. Available values for `level` are as follows:

- **ANOMINAL**: variables are asymmetric nominal and can be either numeric or character.
- **NOMINAL**: variables are symmetric nominal and can be either numeric or character.
- **ORDINAL**: variables are ordinal and can be either numeric or character. Values of ordinal variables are replaced by their corresponding rank scores. If standardization is required, the standardized rank scores are output to the data set specified in the OUTSDZ= option. See the RANKSCORE= option in the PROC DISTANCE statement for methods available for assigning rank scores to ordinal variables. After being replaced by scores, ordinal variables are considered interval.
- **INTERVAL**: variables are interval and numeric.
- **RATIO**: variables are ratio and numeric. Ratio variables should always contain positive measurements.

Each variable list can be followed by an `option` list. Use “/” after the list of variables to start the option list. An `option` list contains options that are applied to the variables. The following `options` are available in the option list:

- **ABSENT=** specifies the value to be used as an absence value in an irrelevant absent-absent match for asymmetric nominal variables.
- **MISSING=** specifies the method (or numeric value) with which to replace missing data.
- **ORDER=** selects the order for assigning scores to ordinal variables.
- **STD=** selects the standardization method.
- **WEIGHTS=** assigns weights to the variables in the list.

If an `option` is missing from the current attribute list, PROC DISTANCE provides default values for all the variables in the current list.

For example, in the VAR statement

```
var ratio(x1-x4/std= mad weights= .5 .5 .1 .5 missing= -99)
   interval(x5/std= range)
   ordinal(x6/order= desc);
```

the first option list defines x1–x4 as ratio variables to be standardized by the MAD method. Also, any missing values in x1–x4 should be replaced by -99. x1 is given a weight of 0.5, x2 is given a weight of 0.5, x3 is given a weight of 0.1, and x4 is given a weight of 0.5.

The second option list defines x5 as an interval variable to be standardized by the RANGE method. If the REPLACE option is specified in the PROC DISTANCE statement, missing values in x5 are replaced by the location estimate from the RANGE method. By default, x5 is given a weight of 1.

The last option list defines x6 as an ordinal variable. The scores are assigned from highest to lowest by its unformatted values. Although the STD= option is not specified, x6 is standardized by the default method.
(STD) because there is more than one level of measurements (ratio, interval, and ordinal) in the VAR statement. Again, if the REPLACE option is specified, missing values in x6 are replaced by the location estimate from the STD method. Finally, by default, x6 is given a weight of 1.

More details for the options are explained as follows.

**STD=std-method**

specifies the standardization method. Valid values for std-method are MEAN, MEDIAN, SUM, EUCLEN, USTD, STD, RANGE, MIDRANGE, MAXABS, IQR, MAD, ABW, AHUBER, AWA VE, AGK, SPACING, and L. Table 40.9 lists available methods of standardization as well as their corresponding location and scale measures.

<table>
<thead>
<tr>
<th>Method</th>
<th>Scale</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEAN</td>
<td>1</td>
<td>mean</td>
</tr>
<tr>
<td>MEDIAN</td>
<td>1</td>
<td>median</td>
</tr>
<tr>
<td>SUM</td>
<td>sum</td>
<td>0</td>
</tr>
<tr>
<td>EUCLEN</td>
<td>Euclidean length</td>
<td>0</td>
</tr>
<tr>
<td>USTD</td>
<td>standard deviation about origin</td>
<td>0</td>
</tr>
<tr>
<td>STD</td>
<td>standard deviation</td>
<td>mean</td>
</tr>
<tr>
<td>RANGE</td>
<td>range</td>
<td>minimum</td>
</tr>
<tr>
<td>MIDRANGE</td>
<td>range/2</td>
<td>midrange</td>
</tr>
<tr>
<td>MAXABS</td>
<td>maximum absolute value</td>
<td>0</td>
</tr>
<tr>
<td>IQR</td>
<td>interval quartile range</td>
<td>median</td>
</tr>
<tr>
<td>MAD</td>
<td>median absolute deviation from median</td>
<td>median</td>
</tr>
<tr>
<td>ABW(c)</td>
<td>biweight A-estimate</td>
<td>biweight 1-step M-estimate</td>
</tr>
<tr>
<td>AHUBER(c)</td>
<td>Huber A-estimate</td>
<td>Huber 1-step M-estimate</td>
</tr>
<tr>
<td>AWA VE(c)</td>
<td>Wave 1-step M-estimate</td>
<td>Wave A-estimate</td>
</tr>
<tr>
<td>AGK(p)</td>
<td>AGK estimate (ACECLUS)</td>
<td>mean</td>
</tr>
<tr>
<td>SPACING(p)</td>
<td>minimum spacing</td>
<td>mid minimum-spacing</td>
</tr>
<tr>
<td>L(p)</td>
<td>L_p</td>
<td>L_p</td>
</tr>
</tbody>
</table>

These standardization methods are further documented in the section on the METHOD= option in the PROC STDIZE statement of the STDIZE procedure (see the section “Standardization Methods” on page 9347 in Chapter 112, “The STDIZE Procedure”).

Standardization is not required if there is only one level of measurement, or if only asymmetric nominal and nominal levels are specified; otherwise, standardization is mandatory. When standardization is mandatory, a default method is provided when the STD= option is not specified. You can suppress the mandatory standardization by using the NOSTD option in the PROC DISTANCE statement. See the NOSTD option in the section “PROC DISTANCE Statement” on page 2680 and the section “Mandatory Standardization” on page 2674 for details.

The default method is STD for standardizing interval variables unless METHOD=GOWER or METHOD=DGOWER is specified. If METHOD=GOWER is specified, interval variables are standardized by the RANGE method, and whatever is specified in the STD= option is ignored; if METHOD=DGOWER is specified, the RANGE method is the default standardization method for
interval variables. The MAXABS method is the default standardization method for ratio variables for both the GOWER and DGOWER methods.

Notice that a ratio variable should always be positive.

Table 40.10 lists standardization methods and the levels of measurement that can be accepted by each method. For example, the SUM method can be used to standardize ratio variables but not interval or ordinal variables. Also, the AGK and SPACING methods should not be used to standardize ordinal variables. If you apply AGK and SPACING to ranks, the results are degenerate because all the spacings of a given order are equal.

Table 40.10 Legitimate Levels of Measurements for Each Method

<table>
<thead>
<tr>
<th>Standardization Method</th>
<th>Legitimate Levels of Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEAN</td>
<td>ratio, interval, ordinal</td>
</tr>
<tr>
<td>MEDIAN</td>
<td>ratio, interval, ordinal</td>
</tr>
<tr>
<td>SUM</td>
<td>ratio</td>
</tr>
<tr>
<td>EUCLEN</td>
<td>ratio</td>
</tr>
<tr>
<td>USTD</td>
<td>ratio</td>
</tr>
<tr>
<td>STD</td>
<td>ratio, interval, ordinal</td>
</tr>
<tr>
<td>RANGE</td>
<td>ratio, interval, ordinal</td>
</tr>
<tr>
<td>MIDRANGE</td>
<td>ratio, interval, ordinal</td>
</tr>
<tr>
<td>MAXABS</td>
<td>ratio</td>
</tr>
<tr>
<td>IQR</td>
<td>ratio, interval, ordinal</td>
</tr>
<tr>
<td>MAD</td>
<td>ratio, interval, ordinal</td>
</tr>
<tr>
<td>ABW(c)</td>
<td>ratio, interval, ordinal</td>
</tr>
<tr>
<td>AHUBER(c)</td>
<td>ratio, interval, ordinal</td>
</tr>
<tr>
<td>AWAIVE(c)</td>
<td>ratio, interval, ordinal</td>
</tr>
<tr>
<td>AGK(p)</td>
<td>ratio, interval</td>
</tr>
<tr>
<td>SPACING(p)</td>
<td>ratio, interval</td>
</tr>
<tr>
<td>L(p)</td>
<td>ratio, interval, ordinal</td>
</tr>
</tbody>
</table>

ABSENT=number | qs
specifies the value to be used as an absence value in an irrelevant absent-absent match for asymmetric nominal variables. The absence value specified here overwrites the absence value specified through the ABSENT= option in the PROC DISTANCE statement for those variables in the current variable list.

An absence value for a variable can be either a numeric value or a quoted string consisting of combinations of characters. For instance, ., –999, “NA” are legal values for the ABSENT= option.

The default for an absence value for a character variable is “NONE” (notice that a blank value is considered a missing value), and the default for an absence value for a numeric variable is 0.

MISSING=miss-method | value
specifies the method or a numeric value for replacing missing values. If you omit the MISSING= option, the REPLACE option replaces missing values with the location measure given by the STD= option. Specify the MISSING= option when you want to replace missing values with a different value. You can specify any method that is valid in the STD= option. The corresponding location measure is used to replace missing values.
If a numeric value is given, the value replaces missing values after standardizing the data. However, when standardization is not mandatory, you can specify the REPONLY option with the MISSING= option to suppress standardization for cases in which you want only to replace missing values.

If the NOSTD option is specified, there is no standardization, but missing values are replaced by the corresponding location measures or by the numeric value of the MISSING= option. See the section “Missing Values” on page 2699 for details about missing values replacement with and without standardization.

ORDER=ASCENDING | DESCENDING | ASCFORMATTED | DESFORMATTED | DSORDER
ORDER=ASC | DESC | ASCFMT | DESFMT | DATA

specifies the order for assigning score to ordinal variables. The value for the ORDER= option can be one of the following:

ASCENDING  scores are assigned in lowest-to-highest order of unformatted values.
DESCENDING  scores are assigned in highest-to-lowest order of unformatted values.
ASCFORMATTED scores are assigned in ascending order by their formatted values. This option can be applied to character variables only, since unformatted values are always used for numeric variables.
DESCFORMATTED scores are assigned in descending order by their formatted values. This option can be applied to character variables only, since unformatted values are always used for numeric variables.
DSORDER     scores are assigned according to the order of their appearance in the input data set.

The default value is ASCENDING.

WEIGHTS=weight-list

specifies a list of values for weighting individual variables while computing the proximity. Values in this list can be separated by blanks or commas. You can include one or more items of the form start TO stop BY increment. This list should contain at least one weight. The maximum number of weights you can list is equal to the number of variables. If the number of weights is less than the number of variables, the last value in the weight-list is used for the rest of the variables; conversely, if the number of weights is greater than the number of variables, the trailing weights are discarded.

The default value is 1.

ID Statement

ID variable;

The ID statement specifies a single variable to be copied to the OUT= data set and used to generate names for the distance variables. The ID variable must be character.

Typically, each ID value occurs only once in the input data set or, if you use a BY statement, only once within a BY group.

If you specify both the ID and BY statements, the ID variable must have the same values in the same order in each BY group.
**COPY Statement**

```
COPY variables;
```

The COPY statement specifies a list of additional variables to be copied to the OUT= data set.

---

**BY Statement**

```
BY variables;
```

You can specify a BY statement in PROC DISTANCE to obtain separate analyses of observations in groups that are 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 you specify more than one BY statement, only the last one specified is used.

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

- Sort the data by using the SORT procedure with a similar BY statement.
- Specify the NOTSORTED or DESCENDING option in the BY statement in the DISTANCE 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 by using the DATASETS procedure (in Base SAS software).

For more information about BY-group processing, see the discussion in *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the discussion in the *Base SAS Procedures Guide*.

---

**FREQ Statement**

```
FREQ | FREQUENCY variable;
```

The frequency variable is used both for standardizing variables and for assigning rank scores to ordinal variables. It has no direct effect on computing the distances. It has no direct effect on computing the distances.

For standardizing variables and assigning rank scores, PROC DISTANCE treats the data set as if each observation appeared $n$ times, where $n$ is the value of the FREQ variable for the observation. Nonintegral values of the FREQ variable are truncated to the largest integer less than the FREQ value. If the FREQ variable has a value that is less than 1 or is missing, the observation is not used in the analysis, and distances between that observation and other observations are set to missing.

---

**WEIGHT Statement**

```
WGT | WEIGHT variable;
```
The WEIGHT statement specifies a numeric variable in the input data set with values that are used to weight each observation. This weight variable is used for standardizing variables rather than computing the distances. Only one variable can be specified.

The WEIGHT variable values can be nonintegers. An observation is used in the analysis only if the value of the WEIGHT variable is greater than zero. The WEIGHT variable applies to variables that are standardized by the following options: STD=MEAN, STD=SUM, STD=EUCLEN, STD=USTD, STD=STD, STD=AGK, or STD=L.

PROC DISTANCE uses the value of the WEIGHT variable $w_i$ to compute the sample mean, uncorrected sample variances, and sample variances as follows:

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$$

$$u^2_w = \frac{\sum w_i x_i^2}{d}$$

$$s^2_w = \frac{\sum w_i (x_i - \bar{x}_w)^2}{d}$$

$w_i$ is the weight value of the $i$th observation, $x_i$ is the value of the $i$th observation, and $d$ is the divisor controlled by the VARDEF= option (see the VARDEF= option in the PROC DISTANCE statement for details).

PROC DISTANCE uses the value of the WEIGHT variable to calculate the following statistics for standardization:

**MEAN** the weighted mean, $\bar{x}_w$

**SUM** the weighted sum, $\sum w_i x_i$

**USTD** the weighted uncorrected standard deviation, $\sqrt{u^2_w}$

**STD** the weighted standard deviation, $\sqrt{s^2_w}$

**EUCLEN** the weighted Euclidean length, computed as the square root of the weighted uncorrected sum of squares:

$$\sqrt{\sum_i w_i x_i^2}$$

**AGK** the AGK estimate. This estimate is documented further in the ACECLUS procedure as the METHOD=COUNT option. See the discussion of the WEIGHT statement in Chapter 24, “The ACECLUS Procedure,” for information about how the WEIGHT variable is applied to the AGK estimate.

**L** the $L_p$ estimate. This estimate is documented further in the FASTCLUS procedure as the LEAST= option. See the discussion of the WEIGHT statement in Chapter 42, “The FASTCLUS Procedure,” for information about how the WEIGHT variable is used to compute weighted cluster means. Note that the number of clusters is always 1.
Details: DISTANCE Procedure

Proximity Measures

The following notation is used in this section:

\( v \) the number of variables or the dimensionality
\( x_j \) data for observation \( x \) and the \( j \)th variable, where \( j = 1 \) to \( v \)
\( y_j \) data for observation \( y \) and the \( j \)th variable, where \( j = 1 \) to \( v \)
\( w_j \) weight for the \( j \)th variable from the WEIGHTS= option in the VAR statement. \( w_j = 0 \) when either \( x_j \) or \( y_j \) is missing.
\( W \) the sum of total weights. No matter if the observation is missing or not, its weight is added to this metric.
\( \bar{x} \) mean for observation \( x \)
\( \bar{x} = \sum_{j=1}^{v} w_j x_j / \sum_{j=1}^{v} w_j \)
\( \bar{y} \) mean for observation \( y \)
\( \bar{y} = \sum_{j=1}^{v} w_j y_j / \sum_{j=1}^{v} w_j \)
\( d(x, y) \) the distance or dissimilarity between observations \( x \) and \( y \)
\( s(x, y) \) the similarity between observations \( x \) and \( y \)

The factor \( W/ \sum_{j=1}^{v} w_j \) is used to adjust some of the proximity measures for missing values.

Methods That Accept All Measurement Levels

GOWER Gower’s similarity
\[ s_1(x, y) = \sum_{j=1}^{v} w_j \delta^i_{x,y} d^i_{x,y} / \sum_{j=1}^{v} w_j \delta^i_{x,y} \]
\( \delta^i_{x,y} \) is computed as follows:
For nominal, ordinal, interval, or ratio variable,
\[ \delta^i_{x,y} = 1 \]
For asymmetric nominal variable,
\[ \delta^i_{x,y} = 1, \text{ if either } x_j \text{ or } y_j \text{ is present} \]
\[ \delta^i_{x,y} = 0, \text{ if both } x_j \text{ and } y_j \text{ are absent} \]
For nominal or asymmetric nominal variable,
\[ d^i_{x,y} = 1, \text{ if } x_j = y_j \]
\[ d^i_{x,y} = 0, \text{ if } x_j \neq y_j \]
For ordinal, interval, or ratio variable,
\[ d^i_{x,y} = 1 - |x_j - y_j| \]
DGOWER 1 minus Gower
d_2(x, y) = 1 - s_1(x, y)

## Methods That Accept Ratio, Interval, and Ordinal Variables

**EUCLID** Euclidean distance
\[ d_3(x, y) = \sqrt{\left(\sum_{j=1}^{v} w_j (x_j - y_j)^2\right) / \left(\sum_{j=1}^{v} w_j\right)} \]

**SQEUCLID** squared Euclidean distance
\[ d_4(x, y) = \left(\sum_{j=1}^{v} w_j (x_j - y_j)^2\right) / \left(\sum_{j=1}^{v} w_j\right) \]

**SIZE** size distance
\[ d_5(x, y) = |\sum_{j=1}^{v} w_j (x_j - y_j)| \sqrt{\left(\sum_{j=1}^{v} w_j\right)} \]

**SHAPE** shape distance
\[ d_6(x, y) = \sqrt{\left(\sum_{j=1}^{v} w_j [(x_j - \bar{x}) - (y_j - \bar{y})]^2\right) / \left(\sum_{j=1}^{v} w_j\right)} \]

**COV** covariance similarity coefficient
\[ s_7(x, y) = \sum_{j=1}^{v} w_j (x_j - \bar{x})(y_j - \bar{y}) / \text{vardiv}, \text{ where} \]
\[ \text{vardiv} = v \text{ if VARDEF = N} \]
\[ = v - 1 \text{ if VARDEF = DF} \]
\[ = \sum_{j=1}^{v} w_j \text{ if VARDEF = WEIGHT} \]
\[ = \sum_{j=1}^{v} w_j - 1 \text{ if VARDEF = WDF} \]

**CORR** correlation similarity coefficient
\[ s_8(x, y) = \frac{\sum_{j=1}^{v} w_j (x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_{j=1}^{v} w_j (x_j - \bar{x})^2 \sum_{j=1}^{v} w_j (y_j - \bar{y})^2}} \]

**DCORR** correlation transformed to Euclidean distance as \(\sqrt{1-CORR}\)
\[ d_9(x, y) = \sqrt{1 - s_8(x, y)} \]

**SQCORR** squared correlation
\[ s_{10}(x, y) = \frac{\left[\sum_{j=1}^{v} w_j (x_j - \bar{x})(y_j - \bar{y})\right]^2}{\sum_{j=1}^{v} w_j (x_j - \bar{x})^2 \sum_{j=1}^{v} w_j (y_j - \bar{y})^2} \]

**DSQCORR** squared correlation transformed to squared Euclidean distance as \((1-SQCORR)\)
\[ d_{11}(x, y) = 1 - s_{10}(x, y) \]

**L(p)** Minkowski \((L_p)\) distance, where \(p\) is a positive numeric value
Proximity Measures

\[ d_{12}(x, y) = \left[ \left( \sum_{j=1}^{v} w_j |x_j - y_j|^p \right) W / \left( \sum_{j=1}^{v} w_j \right) \right]^{1/p} \]

CITYBLOCK \( L_1 \)
\[ d_{13}(x, y) = \left( \sum_{j=1}^{v} w_j |x_j - y_j| \right) W / \left( \sum_{j=1}^{v} w_j \right) \]

CHEBYCHEV \( L_\infty \)
\[ d_{14}(x, y) = \max_{j=1}^{v} w_j |x_j - y_j| \]

POWER\((p,r)\) generalized Euclidean distance, where \( p \) is a nonnegative numeric value and \( r \) is a positive numeric value. The distance between two observations is the \( r \)th root of sum of the absolute differences to the \( p \)th power between the values for the observations:
\[ d_{15}(x, y) = \left[ \left( \sum_{j=1}^{v} w_j |x_j - y_j|^p \right) W / \left( \sum_{j=1}^{v} w_j \right) \right]^{1/r} \]

Methods That Accept Ratio Variables

SIMRATIO similarity ratio
\[ s_{16}(x, y) = \frac{\sum_{j=1}^{v} w_j (x_j y_j)}{\sum_{j=1}^{v} w_j (x_j y_j) + \sum_{j} w_j (x_j-y_j)^2} \]

DISRATIO one minus similarity ratio
\[ d_{17}(x, y) = 1 - s_{16}(x, y) \]

NONMETRIC Lance-Williams nonmetric coefficient
\[ d_{18}(x, y) = \frac{\sum_{j=1}^{v} w_j |x_j - y_j|}{\sum_{j=1}^{v} w_j (x_j + y_j)} \]

CANBERRA Canberra metric coefficient. See Sneath and Sokal (1973, pp. 125–126)
\[ d_{19}(x, y) = \sum_{j=1}^{v} \frac{w_j |x_j - y_j|}{w_j (x_j + y_j)} \]

COSINE cosine coefficient
\[ s_{20}(x, y) = \frac{\sum_{j=1}^{v} w_j (x_j y_j)}{\sqrt{\sum_{j=1}^{v} w_j x_j^2} \sqrt{\sum_{j=1}^{v} w_j y_j^2}} \]

DOT dot (inner) product coefficient
\[ s_{21}(x, y) = \left[ \sum_{j=1}^{v} w_j (x_j y_j) \right] / \sum_{j=1}^{v} w_j \]

OVERLAP sum of the minimum values
\[ s_{22}(x, y) = \sum_{j=1}^{v} w_j \min(x_j, y_j) \]

DOVERLAP maximum of the sum of the \( x \) and the sum of \( y \) minus overlap
\[ d_{23}(x, y) = \max(\sum_{j=1}^{v} w_j x_j, \sum_{j=1}^{v} w_j y_j) - s_{22}(x, y) \]

CHISQ chi-square
If the data represent the frequency counts, chi-square dissimilarity between two sets of
frequencies can be computed. A 2-by-$v$ contingency table is illustrated to explain how the chi-square dissimilarity is computed as follows:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Row</th>
<th>Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td></td>
<td>x_1 x_2 . . . x_v</td>
</tr>
<tr>
<td>Y</td>
<td></td>
<td>y_1 y_2 . . . y_v</td>
</tr>
<tr>
<td>Column Sum</td>
<td>c_1 c_2 . . . c_v</td>
<td>T</td>
</tr>
</tbody>
</table>

where

\[ r_x = \sum_{j=1}^{v} w_j x_j \]
\[ r_y = \sum_{j=1}^{v} w_j y_j \]
\[ c_j = w_j (x_j + y_j) \]
\[ T = r_x + r_y = \sum_{j=1}^{v} c_j \]

The chi-square measure is computed as follows:

\[ d_{24}(x, y) = \left( \sum_{j=1}^{v} \frac{(w_j x_j - E(x_j))^2}{E(x_j)} \right) + \sum_{j=1}^{v} \frac{(w_j y_j - E(y_j))^2}{E(y_j)} \]
\[ W / \left( \sum_{j=1}^{v} w_j \right) \]

where for $j = 1, 2, \ldots, v$

\[ E(x_j) = r_x c_j / T \]
\[ E(y_j) = r_y c_j / T \]

**CHI**

square root of chi-square

\[ d_{25}(x, y) = \sqrt{d_{24}(x, y)} \]

**PHISQ**

phi-square

This is the CHISQ dissimilarity normalized by the sum of weights

\[ d_{26}(x, y) = d_{24}(x, y) / \left( \sum_{j=1}^{v} w_j \right) \]

**PHI**

square root of phi-square

\[ d_{27}(x, y) = \sqrt{d_{25}(x, y)} \]

**Methods That Accept Symmetric Nominal Variables**

The following notation is used for computing $d_{28}(x, y)$ to $s_{35}(x, y)$. Notice that only the nonmissing pairs are discussed below; all the pairs with at least one missing value will be excluded from any of the computations in the following section because $w_j = 0$, if either $x_j$ or $y_j$ is missing.

\[ M \]

nonmissing matches

\[ M = \sum_{j=1}^{v} w_j \delta_{x,y}^j, \text{ where} \]

\[ \delta_{x,y}^j = 1, \text{ if } x_j = y_j \]
\[ \delta_{x,y}^j = 0, \text{ otherwise} \]
Proximity Measures

$X$ nonmissing mismatches

$X = \sum_{j=1}^{v} w_j \delta_{x,y}^j$, where

$\delta_{x,y}^j = 1, \text{ if } x_j \neq y_j$

$\delta_{x,y}^j = 0, \text{ otherwise}$

$N$ total nonmissing pairs

$N = \sum_{j=1}^{v} w_j$

HAMMING Hamming distance

$d_{28}(x, y) = X$

MATCH simple matching coefficient

$s_{29}(x, y) = M/N$

DMATCH simple matching coefficient transformed to Euclidean distance

$d_{30}(x, y) = \sqrt{1 - M/N} = \sqrt{(X/N)}$

DSQMATCH simple matching coefficient transformed to squared Euclidean distance

$d_{31}(x, y) = 1 - M/N = X/N$

HAMANN Hamann coefficient

$s_{32}(x, y) = (M - X)/N$

RT Roger and Tanimoto

$s_{33}(x, y) = M/(M + 2X)$

SS1 Sokal and Sneath 1

$s_{34}(x, y) = 2M/(2M + X)$

SS3 Sokal and Sneath 3. The coefficient between an observation and itself is always indeterminate (missing) since there is no mismatch.

$s_{35}(x, y) = M/X$

The following notation is used for computing $s_{36}(x, y)$ to $d_{41}(x, y)$. Notice that only the nonmissing pairs are discussed in the following section; all the pairs with at least one missing value are excluded from any of the computations in the following section because $w_j = 0$, if either $x_j$ or $y_j$ is missing.

Also, the observed nonmissing data of an asymmetric binary variable can have only two possible outcomes: presence or absence. Therefore, the notation, $PX$ (present mismatches), always has a value of zero for an asymmetric binary variable.
The following methods distinguish between the presence and absence of attributes.

$X$ mismatches with at least one present

$$X = \sum_{j=1}^{v} w_j \delta_{x,y}^j,$$

where

$$\delta_{x,y}^j = \begin{cases} 1, & \text{if } x_j \neq y_j \text{ and not both } x_j \text{ and } y_j \text{ are absent} \\ 0, & \text{otherwise} \end{cases}$$

$PM$ present matches

$$PM = \sum_{j=1}^{v} w_j \delta_{x,y}^j,$$

where

$$\delta_{x,y}^j = \begin{cases} 1, & \text{if } x_j = y_j \text{ and both } x_j \text{ and } y_j \text{ are present} \\ 0, & \text{otherwise} \end{cases}$$

$PX$ present mismatches

$$PX = \sum_{j=1}^{v} w_j \delta_{x,y}^j,$$

where

$$\delta_{x,y}^j = \begin{cases} 1, & \text{if } x_j \neq y_j \text{ and both } x_j \text{ and } y_j \text{ are present} \\ 0, & \text{otherwise} \end{cases}$$

$PP$ both present $= PM + PX$

$P$ at least one present $= PM + X$

$PAX$ present-absent mismatches

$$PAX = \sum_{j=1}^{v} w_j \delta_{x,y}^j,$$

where

$$\delta_{x,y}^j = \begin{cases} 1, & \text{if } x_j \neq y_j \text{ and either } x_j \text{ is present and } y_j \text{ is absent or } \\
& x_j \text{ is absent and } y_j \text{ is present} \\ 0, & \text{otherwise} \end{cases}$$

$N$ total nonmissing pairs

$$N = \sum_{j=1}^{v} w_j$$

**Methods That Accept Asymmetric Nominal and Ratio Variables**

JACCARD Jaccard similarity coefficient

The JACCARD method is equivalent to the SIMRATIO method if there are only ratio variables; if there are both ratio and asymmetric nominal variables, the coefficient is computed as sum of the coefficient from the ratio variables (SIMRATIO) and the coefficient from the asymmetric nominal variables.

$$s_{36}(x, y) = s_{16}(x, y) + PM/P$$
DJACCARD  Jaccard dissimilarity coefficient

The DJACCARD method is equivalent to the DISRATIO method if there are only ratio variables; if there are both ratio and asymmetric nominal variables, the coefficient is computed as sum of the coefficient from the ratio variables (DISRATIO) and the coefficient from the asymmetric nominal variables.

\[ d_{37}(x, y) = d_{17}(x, y) + \frac{X}{P} \]

Methods That Accept Asymmetric Nominal Variables

DICE  Dice coefficient or Czekanowski/Sorensen similarity coefficient

\[ s_{38}(x, y) = \frac{2PM}{(P + PM)} \]

RR  Russell and Rao. This is the binary equivalent of the dot product coefficient.

\[ s_{39}(x, y) = \frac{PM}{N} \]

BLWNM | BRAYCURTIS  Binary Lance and Williams, also known as Bray and Curtis coefficient

\[ d_{40}(x, y) = \frac{X}{(PAX + 2PP)} \]

K1  Kulczynski 1. The coefficient between an observation and itself is always indeterminate (missing) since there is no mismatch.

\[ d_{41}(x, y) = \frac{PM}{X} \]

Missing Values

Standardization versus No Standardization

You can replace the missing values with or without standardization. Missing values are replaced after standardization by specifying either the REPLACE option in the PROC DISTANCE statement or the MISSING= option in the VAR statement.

To replace missing values without standardization, use the following two options:

- the NOSTD option in the PROC DISTANCE statement. The NOSTD option suppresses standardization but still replaces the missing values with the location of the method or the numeric value specified in the MISSING= option in the VAR statement.

- the REPONLY option in the PROC DISTANCE statement. PROC DISTANCE replaces missing values with the location of the standardization method or with the numeric value specified in the MISSING= option in the VAR statement. This approach assumes that standardization is not mandatory (see the section “Standardization” on page 2674).
Eliminating Observations with Missing Values

If you specify the NOMISS option, PROC DISTANCE omits observations with any missing values in the analyzed variables from computation of the location and scale measures.

Distance Measures

If you specify the NOMISS option, PROC DISTANCE generates missing distance for observations with missing values. If the NOMISS option is not specified, the sum of total weights, no matter if an observation is missing or not, is incorporated into the computation of some of the proximity measures. See the section “Details: DISTANCE Procedure” on page 2693 for the formulas and descriptions.

Formatted versus Unformatted Values

PROC DISTANCE uses the formatted values from a character variable, if the variable has a format—for example, one assigned by a format statement. PROC DISTANCE uses the unformatted values from a numeric variable, even if it has a format.

Output Data Sets

OUT= Data Set

The DISTANCE procedure always produces an output data set, regardless of whether you specify the OUT= option in the PROC DISTANCE statement. PROC DISTANCE displays no output. Use PROC PRINT to display the output data set.

The output data set contains the following variables:

- the ID variable, if any
- the BY variables, if any
- the COPY variables, if any
- the FREQ variable, if any
- the WEIGHT variable, if any
- the new distance variables, named from PREFIX= options along with the ID values, or from the default values

The output data set is of type TYPE=DISTANCE or TYPE=SIMILAR. See the METHOD= option for more information about the output data set types. Data set types do not persist when you copy or modify a data set. You must specify the TYPE= data set option for the new data set, as in the following example:

```sas
   data dist2(type=distance);
   set dist;
   run;
```

See the OUT= option for more information about data set type persistence.
OUTSDZ= Data Set

The output data set is a copy of the DATA= data set except that the analyzed variables have been standardized. Analyzed variables are those listed in the VAR statement.

---

Examples: DISTANCE Procedure

---

Example 40.1: Divorce Grounds – the Jaccard Coefficient

A wide variety of distance and similarity measures are used in cluster analysis (Anderberg 1973; Sneath and Sokal 1973). If your data are in coordinate form and you want to use a non-Euclidean distance for clustering, you can compute a distance matrix by using the DISTANCE procedure.

Similarity measures must be converted to dissimilarities before being used in PROC CLUSTER. Such conversion can be done in a variety of ways, such as taking reciprocals or subtracting from a large value. The choice of conversion method depends on the application and the similarity measure. If applicable, PROC DISTANCE provides a corresponding dissimilarity measure for each similarity measure.

In the following example, the observations are states. Binary-valued variables correspond to various grounds for divorce and indicate whether the grounds for divorce apply in each of the U.S. states. A value of “1” indicates that the ground for divorce applies, and a value of “0” indicates the opposite. The 0-0 matches are treated as totally irrelevant; therefore, each variable has an asymmetric nominal level of measurement. The absence value is 0.

The DISTANCE procedure is used to compute the Jaccard coefficient (Anderberg 1973, pp. 89, 115, and 117) between each pair of states. The Jaccard coefficient is defined as the number of variables that are coded as 1 for both states divided by the number of variables that are coded as 1 for either or both states. Since dissimilarity measures are required by PROC CLUSTER, the DJACCARD coefficient is selected. Output 40.1.1 displays the distance matrix between the first 10 states.

The CENTROID method is used to perform the cluster analysis, and the resulting tree diagram from PROC CLUSTER is saved into the tree output data set. Output 40.1.2 displays the cluster history.

The TREE procedure generates nine clusters in the output data set out. After being sorted by the state, the out data set is then merged with the input data set divorce. After being sorted by the state, the merged data set is printed to display the cluster membership as shown in Output 40.1.3.

The following statements produce Output 40.1.1 through Output 40.1.3:

```sas
data divorce;
  length State $ 15;
  input State &$ Incompatibility Cruelty Desertion Non_Support Alcohol Felony Impotence Insanity Separation @@;
  datalines;
Alabama        1 1 1 1 1 1 1 1 1 Alaska        1 1 1 0 1 1 1 1 0
Arizona        1 0 0 0 0 0 0 0 0 Arkansas      0 1 1 1 1 1 1 1 1
California     1 0 0 0 0 0 1 0 Colorado      1 0 0 0 0 0 0 0 0
Connecticut    1 1 1 1 1 1 0 1 1 Delaware    1 0 0 0 0 0 0 0 1
```

---

Examples: DISTANCE Procedure
... more lines ...

Wisconsin 1 0 0 0 0 0 0 1
Wyoming 1 0 0 0 0 0 1 1

; title 'Grounds for Divorce';
proc distance data=divorce method=djaccard absent=0 out=distjacc;
    var anominal(Incompatibility--Separation);
    id state;
run;

proc print data=distjacc(obs=10);
    id state; var alabama--georgia;
    title2 'First 10 States';
run;

proc cluster data=distjacc method=centroid
    pseudo outtree=tree;
    id state;
    var alabama--wyoming;
run;

proc tree data=tree noprint n=9 out=out;
    id state;
run;

proc sort;
    by state;
run;

data clus;
    merge divorce out;
    by state;
run;

proc sort;
    by cluster;
run;

proc print;
    id state;
    var Incompatibility--Separation;
    by cluster;
run;
### Output 40.1.1 Distance Matrix Based on the Jaccard Coefficient

#### Grounds for Divorce
#### First 10 States

<table>
<thead>
<tr>
<th>State</th>
<th>Alabama</th>
<th>Alaska</th>
<th>Arizona</th>
<th>Arkansas</th>
<th>California</th>
<th>Colorado</th>
<th>Connecticut</th>
<th>Delaware</th>
<th>Florida</th>
<th>Georgia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>0.00000</td>
<td>0.22222</td>
<td>0.88889</td>
<td>0.11111</td>
<td>0.77778</td>
<td>0.77778</td>
<td>0.11111</td>
<td>0.77778</td>
<td>0.77778</td>
<td>0.22222</td>
</tr>
<tr>
<td>Alaska</td>
<td>0.22222</td>
<td>0.00000</td>
<td>0.85714</td>
<td>0.33333</td>
<td>0.87500</td>
<td>0.87500</td>
<td>0.33333</td>
<td>0.87500</td>
<td>0.87500</td>
<td>0.33333</td>
</tr>
<tr>
<td>Arizona</td>
<td>0.88889</td>
<td>0.85714</td>
<td>0.00000</td>
<td>0.87500</td>
<td>0.87500</td>
<td>0.87500</td>
<td>0.87500</td>
<td>0.87500</td>
<td>0.87500</td>
<td>0.87500</td>
</tr>
<tr>
<td>Arkansas</td>
<td>0.11111</td>
<td>0.33333</td>
<td>1.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>California</td>
<td>0.77778</td>
<td>0.71429</td>
<td>0.50000</td>
<td>0.88889</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>Colorado</td>
<td>0.88889</td>
<td>0.85714</td>
<td>0.00000</td>
<td>1.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>Connecticut</td>
<td>0.11111</td>
<td>0.33333</td>
<td>0.87500</td>
<td>0.22222</td>
<td>0.75000</td>
<td>0.87500</td>
<td>0.00000</td>
<td>0.75000</td>
<td>0.75000</td>
<td>0.00000</td>
</tr>
<tr>
<td>Delaware</td>
<td>0.77778</td>
<td>0.87500</td>
<td>0.50000</td>
<td>0.88889</td>
<td>0.66667</td>
<td>0.50000</td>
<td>0.00000</td>
<td>0.75000</td>
<td>0.66667</td>
<td>0.00000</td>
</tr>
<tr>
<td>Florida</td>
<td>0.77778</td>
<td>0.71429</td>
<td>0.50000</td>
<td>0.88889</td>
<td>0.00000</td>
<td>0.50000</td>
<td>0.75000</td>
<td>0.66667</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>Georgia</td>
<td>0.22222</td>
<td>0.00000</td>
<td>0.85714</td>
<td>0.33333</td>
<td>0.71429</td>
<td>0.85714</td>
<td>0.33333</td>
<td>0.87500</td>
<td>0.71429</td>
<td>0.00000</td>
</tr>
</tbody>
</table>

### Output 40.1.2 Clustering History

#### Grounds for Divorce
#### The CLUSTER Procedure
#### Centroid Hierarchical Cluster Analysis

**Root-Mean-Square Distance Between Observations** 0.694873
### Output 40.1.2 continued

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>Clusters Joined</th>
<th>Freq</th>
<th>Pseudo F Statistic</th>
<th>Pseudo t-Squared</th>
<th>Norm Centroid Distance</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>49 Arizona</td>
<td>Colorado</td>
<td>2</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>48 California</td>
<td>Florida</td>
<td>2</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>47 Alaska</td>
<td>Georgia</td>
<td>2</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>46 Delaware</td>
<td>Hawaii</td>
<td>2</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>45 Connecticut</td>
<td>Idaho</td>
<td>2</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>44 CL49</td>
<td>Iowa</td>
<td>3</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>43 CL47</td>
<td>Kansas</td>
<td>3</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>42 CL44</td>
<td>Kentucky</td>
<td>4</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>41 CL42</td>
<td>Michigan</td>
<td>5</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>40 CL41</td>
<td>Minnesota</td>
<td>6</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>39 CL43</td>
<td>Mississippi</td>
<td>4</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>38 CL40</td>
<td>Missouri</td>
<td>7</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>37 CL38</td>
<td>Montana</td>
<td>8</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>36 CL37</td>
<td>Nebraska</td>
<td>9</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>35 North Dakota</td>
<td>Oklahoma</td>
<td>2</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>34 CL36</td>
<td>Oregon</td>
<td>10</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>33 Massachusetts</td>
<td>Rhode Island</td>
<td>2</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>32 New Hampshire</td>
<td>Tennessee</td>
<td>2</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>31 CL46</td>
<td>Washington</td>
<td>3</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>30 CL31</td>
<td>Wisconsin</td>
<td>4</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>29 Nevada</td>
<td>Wyoming</td>
<td>2</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>28 Alabama</td>
<td>Arkansas</td>
<td>2</td>
<td>1561</td>
<td>0.1599</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>27 CL33</td>
<td>CL32</td>
<td>4</td>
<td>479</td>
<td>0.1799</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>26 CL39</td>
<td>CL35</td>
<td>6</td>
<td>265</td>
<td>0.1799</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>25 CL45</td>
<td>West Virginia</td>
<td>3</td>
<td>231</td>
<td>0.1799</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24 Maryland</td>
<td>Pennsylvania</td>
<td>2</td>
<td>199</td>
<td>0.2399</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23 CL28</td>
<td>Utah</td>
<td>3</td>
<td>167</td>
<td>3.2</td>
<td>0.2468</td>
<td></td>
</tr>
<tr>
<td>22 CL27</td>
<td>Ohio</td>
<td>5</td>
<td>136</td>
<td>5.4</td>
<td>0.2698</td>
<td></td>
</tr>
<tr>
<td>21 CL26</td>
<td>Maine</td>
<td>7</td>
<td>111</td>
<td>8.9</td>
<td>0.2998</td>
<td></td>
</tr>
<tr>
<td>20 CL23</td>
<td>CL21</td>
<td>10</td>
<td>75.2</td>
<td>8.7</td>
<td>0.3004</td>
<td></td>
</tr>
<tr>
<td>19 CL25</td>
<td>New Jersey</td>
<td>4</td>
<td>71.8</td>
<td>6.5</td>
<td>0.3053</td>
<td></td>
</tr>
<tr>
<td>18 CL19</td>
<td>Texas</td>
<td>5</td>
<td>69.1</td>
<td>2.5</td>
<td>0.3077</td>
<td></td>
</tr>
<tr>
<td>17 CL20</td>
<td>CL22</td>
<td>15</td>
<td>48.7</td>
<td>9.9</td>
<td>0.3219</td>
<td></td>
</tr>
<tr>
<td>16 New York</td>
<td>Virginia</td>
<td>2</td>
<td>50.1</td>
<td></td>
<td>0.3598</td>
<td></td>
</tr>
<tr>
<td>15 CL18</td>
<td>Vermont</td>
<td>6</td>
<td>49.4</td>
<td>2.9</td>
<td>0.3797</td>
<td></td>
</tr>
<tr>
<td>14 CL17</td>
<td>Illinois</td>
<td>16</td>
<td>47.0</td>
<td>3.2</td>
<td>0.4425</td>
<td></td>
</tr>
<tr>
<td>13 CL14</td>
<td>CL15</td>
<td>22</td>
<td>29.2</td>
<td>15.3</td>
<td>0.4722</td>
<td></td>
</tr>
<tr>
<td>12 CL48</td>
<td>CL29</td>
<td>4</td>
<td>29.5</td>
<td></td>
<td>0.4797</td>
<td></td>
</tr>
<tr>
<td>11 CL13</td>
<td>CL24</td>
<td>24</td>
<td>27.6</td>
<td>4.5</td>
<td>0.5042</td>
<td></td>
</tr>
<tr>
<td>10 CL11</td>
<td>South Dakota</td>
<td>25</td>
<td>28.4</td>
<td>2.4</td>
<td>0.5449</td>
<td></td>
</tr>
<tr>
<td>9 Louisiana</td>
<td>CL16</td>
<td>3</td>
<td>30.3</td>
<td>3.5</td>
<td>0.5844</td>
<td></td>
</tr>
<tr>
<td>8 CL34</td>
<td>CL30</td>
<td>14</td>
<td>23.3</td>
<td></td>
<td>0.7196</td>
<td></td>
</tr>
<tr>
<td>7 CL8</td>
<td>CL12</td>
<td>18</td>
<td>19.3</td>
<td>15.0</td>
<td>0.7175</td>
<td></td>
</tr>
<tr>
<td>6 CL10</td>
<td>South Carolina</td>
<td>26</td>
<td>21.4</td>
<td>4.2</td>
<td>0.7384</td>
<td></td>
</tr>
<tr>
<td>5 CL6</td>
<td>New Mexico</td>
<td>27</td>
<td>24.0</td>
<td>4.7</td>
<td>0.8303</td>
<td></td>
</tr>
<tr>
<td>4 CL5</td>
<td>Indiana</td>
<td>28</td>
<td>28.9</td>
<td>4.1</td>
<td>0.8343</td>
<td></td>
</tr>
</tbody>
</table>
### Output 40.1.2 continued

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>Clusters Joined</th>
<th>Freq</th>
<th>Pseudo F Statistic</th>
<th>Pseudo t-Squared</th>
<th>Norm Centroid Distance</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 CL4</td>
<td>CL9</td>
<td>31</td>
<td>31.7</td>
<td>10.9</td>
<td>0.8472</td>
<td></td>
</tr>
<tr>
<td>2 CL3</td>
<td>North Carolina</td>
<td>32</td>
<td>55.1</td>
<td>4.1</td>
<td>1.0017</td>
<td></td>
</tr>
<tr>
<td>1 CL2</td>
<td>CL7</td>
<td>50</td>
<td></td>
<td></td>
<td>55.1</td>
<td>1.0663</td>
</tr>
</tbody>
</table>

### Output 40.1.3 Cluster Membership

#### Grounds for Divorce

**CLUSTER=1**

<table>
<thead>
<tr>
<th>State</th>
<th>Incompatibility</th>
<th>Cruelty</th>
<th>Desertion</th>
<th>Non_Support</th>
<th>Alcohol</th>
<th>Felony</th>
<th>Impotence</th>
<th>Insanity</th>
<th>Separation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arizona</td>
<td>1    0 0 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Colorado</td>
<td>1    0 0 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Iowa</td>
<td>1    0 0 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Kentucky</td>
<td>1    0 0 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Michigan</td>
<td>1    0 0 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Minnesota</td>
<td>1    0 0 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Missouri</td>
<td>1    0 0 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Montana</td>
<td>1    0 0 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Nebraska</td>
<td>1    0 0 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Oregon</td>
<td>1    0 0 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**CLUSTER=2**

<table>
<thead>
<tr>
<th>State</th>
<th>Incompatibility</th>
<th>Cruelty</th>
<th>Desertion</th>
<th>Non_Support</th>
<th>Alcohol</th>
<th>Felony</th>
<th>Impotence</th>
<th>Insanity</th>
<th>Separation</th>
</tr>
</thead>
<tbody>
<tr>
<td>California</td>
<td>1    0 0 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Florida</td>
<td>1    0 0 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Nevada</td>
<td>1    0 0 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Wyoming</td>
<td>1    0 0 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Output 40.1.3  continued

**CLUSTER=3**

<table>
<thead>
<tr>
<th>State</th>
<th>Incompatibility</th>
<th>Cruelty</th>
<th>Desertion</th>
<th>Non_Support</th>
<th>Alcohol</th>
<th>Felony</th>
<th>Impotence</th>
<th>Insanity</th>
<th>Separation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Alaska</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Arkansas</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Connecticut</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Georgia</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Idaho</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Illinois</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Kansas</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Maine</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Maryland</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Mississippi</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>New Hampshire</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>New Jersey</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>North Dakota</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Ohio</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Oklahoma</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Rhode Island</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>South Dakota</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tennessee</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Texas</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Utah</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Vermont</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>West Virginia</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**CLUSTER=4**

<table>
<thead>
<tr>
<th>State</th>
<th>Incompatibility</th>
<th>Cruelty</th>
<th>Desertion</th>
<th>Non_Support</th>
<th>Alcohol</th>
<th>Felony</th>
<th>Impotence</th>
<th>Insanity</th>
<th>Separation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delaware</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Hawaii</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Washington</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**CLUSTER=5**

<table>
<thead>
<tr>
<th>State</th>
<th>Incompatibility</th>
<th>Cruelty</th>
<th>Desertion</th>
<th>Non_Support</th>
<th>Alcohol</th>
<th>Felony</th>
<th>Impotence</th>
<th>Insanity</th>
<th>Separation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Louisiana</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>New York</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Virginia</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**CLUSTER=6**

<table>
<thead>
<tr>
<th>State</th>
<th>Incompatibility</th>
<th>Cruelty</th>
<th>Desertion</th>
<th>Non_Support</th>
<th>Alcohol</th>
<th>Felony</th>
<th>Impotence</th>
<th>Insanity</th>
<th>Separation</th>
</tr>
</thead>
<tbody>
<tr>
<td>South Carolina</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Example 40.2: Financial Data – Stock Dividends

The following data set contains the average dividend yields for 15 utility stocks in the United States. The observations are names of the companies, and the variables correspond to the annual dividend yields for the period 1986–1990. The objective is to group similar stocks into clusters.

Before the cluster analysis is performed, the correlation similarity is chosen for measuring the closeness between each observation. Since distance type of measures are required by PROC CLUSTER, METHOD=DCORR is used in the PROC DISTANCE statement to transform the correlation measures to the distance measures. Notice that in Output 40.2.1, all the values in the distance matrix are between 0 and $\sqrt{2}$.

PROC CLUSTER performs hierarchical clustering by using agglomerative methods based on the distance data created from the previous PROC DISTANCE statement. Since the cubic clustering criterion is not suitable for distance data, only the pseudo $F$ statistic is requested to identify the number of clusters.

The two clustering methods are Ward’s and the average linkage methods. Since the results of the pseudo $t^2$ statistic from both Ward’s and the average linkage methods contain many missing values, only the plot of the pseudo $F$ statistic versus the number of clusters is requested along with the dendrogram by specifying PLOTS(ONLY)=(PSF DENDROGRAM) in the PROC CLUSTER statement.

Both Output 40.2.2 and Output 40.2.3 suggest four clusters. Both methods produce the same clustering result, as shown in Output 40.2.4 and Output 40.2.5. The four clusters are as follows:

- Cincinnati G&E and Detroit Edison
- Texas Utilities and Pennsylvania Power & Light
- Union Electric, Iowa-Ill Gas & Electric, Oklahoma Gas & Electric, and Wisconsin Energy
- Orange & Rockland Utilities, Kentucky Utilities, Kansas Power & Light, Allegheny Power, Green Mountain Power, Dominion Resources, and Minnesota Power & Light
title 'Stock Dividends';

data stock;
   length Company $ 27;
   datalines;
   Cincinnati G&E 8.4 8.2 8.4 8.1 8.0
   Texas Utilities 7.9 8.9 10.4 8.9 8.3
   Detroit Edison 9.7 10.7 11.4 7.8 6.5
   Orange & Rockland Utilities 6.5 7.2 7.3 7.7 7.9
   Kentucky Utilities 6.5 6.9 7.0 7.2 7.5
   Kansas Power & Light 5.9 6.4 6.9 7.4 8.0
   Union Electric 7.1 7.5 8.4 7.8 7.7
   Dominion Resources 6.7 6.9 7.0 7.0 7.4
   Allegheny Power 6.7 7.3 7.8 7.9 8.3
   Minnesota Power & Light 5.6 6.1 7.2 7.0 7.5
   Iowa-Ill Gas & Electric 7.1 7.5 8.5 7.8 8.0
   Pennsylvania Power & Light 7.2 7.6 7.7 7.4 7.1
   Oklahoma Gas & Electric 6.1 6.7 7.4 6.7 6.8
   Wisconsin Energy 5.1 5.7 6.0 5.7 5.9
   Green Mountain Power 7.1 7.4 7.8 7.8 8.3
;

proc distance data=stock method=dcorr out=distdcorr;
   var interval(div_1986 div_1987 div_1988 div_1989 div_1990);
   id company;
run;

proc print data=distdcorr;
   id company;
title2 'Distance Matrix for 15 Utility Stocks';
run;
title2;
ods graphics on;
/* compute pseudo statistic versus number of clusters and create plot */
proc cluster data=distdcorr method=ward pseudo plots(only)=(psf dendrogram);
   id company;
run;
/* compute pseudo statistic versus number of clusters and create plot */
proc cluster data=distdcorr method=average pseudo plots(only)=(psf dendrogram);
   id company;
run;
### Output 40.2.1 Distance Matrix Based on the DCORR Coefficient

**Stock Dividends**

**Distance Matrix for 15 Utility Stocks**

<table>
<thead>
<tr>
<th>Company</th>
<th>Cincinnati_G_E</th>
<th>TexasUtilities</th>
<th>Detroit_Edison</th>
<th>Orange_RocklandUtilities</th>
<th>KentuckyUtilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cincinnati G&amp;E</td>
<td>0.00000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Texas Utilities</td>
<td>0.82056</td>
<td>0.00000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detroit Edison</td>
<td>0.40511</td>
<td>0.65453</td>
<td>0.00000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orange &amp; Rockland Utilities</td>
<td>1.35380</td>
<td>0.88583</td>
<td>1.27306</td>
<td>0.00000</td>
<td></td>
</tr>
<tr>
<td>Kentucky Utilities</td>
<td>1.35581</td>
<td>0.92539</td>
<td>1.29382</td>
<td>0.12268</td>
<td>0.00000</td>
</tr>
<tr>
<td>Kansas Power &amp; Light</td>
<td>1.34227</td>
<td>0.94371</td>
<td>1.31696</td>
<td>0.19905</td>
<td>0.12874</td>
</tr>
<tr>
<td>Union Electric</td>
<td>0.98516</td>
<td>0.29043</td>
<td>0.89048</td>
<td>0.68798</td>
<td>0.71824</td>
</tr>
<tr>
<td>Dominion Resources</td>
<td>1.32945</td>
<td>0.96853</td>
<td>1.29016</td>
<td>0.33290</td>
<td>0.21510</td>
</tr>
<tr>
<td>Allegheny Power</td>
<td>1.30492</td>
<td>0.81666</td>
<td>1.24565</td>
<td>0.17844</td>
<td>0.15759</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Company</th>
<th>Kansas_Power___Light</th>
<th>Union_Electric</th>
<th>Dominion_Resources</th>
<th>Allegheny_Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cincinnati G&amp;E</td>
<td>0.00000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Texas Utilities</td>
<td>0.72082</td>
<td>0.00000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detroit Edison</td>
<td>0.24189</td>
<td>0.76587</td>
<td>0.00000</td>
<td></td>
</tr>
<tr>
<td>Orange &amp; Rockland Utilities</td>
<td>0.00000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kentucky Utilities</td>
<td>0.00000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kansas Power &amp; Light</td>
<td>0.17029</td>
<td>0.58452</td>
<td>0.27819</td>
<td>0.00000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Company</th>
<th>Minnesota_Power___Light</th>
<th>Iowa_II_Gas___Electric</th>
<th>Pennsylvania_Power___Light</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cincinnati G&amp;E</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Texas Utilities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detroit Edison</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orange &amp; Rockland Utilities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kentucky Utilities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kansas Power &amp; Light</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Union Electric</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dominion Resources</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allegheny Power</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Company</th>
<th>Oklahoma_Gas___Electric</th>
<th>Wisconsin_Energy</th>
<th>Green_Mountain_Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cincinnati G&amp;E</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Texas Utilities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detroit Edison</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orange &amp; Rockland Utilities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kentucky Utilities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kansas Power &amp; Light</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Union Electric</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dominion Resources</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allegheny Power</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Output 40.2.1 continued

#### Stock Dividends
Distance Matrix for 15 Utility Stocks

<table>
<thead>
<tr>
<th>Company</th>
<th>Cincinnati_G_E</th>
<th>TexasUtilities</th>
<th>Detroit_Edison</th>
<th>Orange_RocklandUtilities</th>
<th>KentuckyUtilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minnesota Power &amp; Light</td>
<td>1.24069</td>
<td>0.74082</td>
<td>1.20432</td>
<td>0.32581</td>
<td>0.30462</td>
</tr>
<tr>
<td>Iowa-Ill Gas &amp; Electric</td>
<td>1.04924</td>
<td>0.43100</td>
<td>0.97616</td>
<td>0.61166</td>
<td>0.61760</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>0.74931</td>
<td>0.37821</td>
<td>0.44256</td>
<td>1.03566</td>
<td>1.08878</td>
</tr>
<tr>
<td>Oklahoma Gas &amp; Electric</td>
<td>1.00604</td>
<td>0.30141</td>
<td>0.86200</td>
<td>0.68021</td>
<td>0.70259</td>
</tr>
<tr>
<td>Wisconsin Energy</td>
<td>1.17988</td>
<td>0.54830</td>
<td>1.03081</td>
<td>0.45013</td>
<td>0.47184</td>
</tr>
<tr>
<td>Green Mountain Power</td>
<td>1.30397</td>
<td>0.88063</td>
<td>1.27176</td>
<td>0.26948</td>
<td>0.17909</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Company</th>
<th>Kansas_Power___Light</th>
<th>Union_Electric</th>
<th>Dominion_Resources</th>
<th>Allegheny_Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minnesota Power &amp; Light</td>
<td>0.27231</td>
<td>0.48372</td>
<td>0.35733</td>
<td>0.15615</td>
</tr>
<tr>
<td>Iowa-Ill Gas &amp; Electric</td>
<td>0.61736</td>
<td>0.16923</td>
<td>0.63545</td>
<td>0.47900</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>1.12876</td>
<td>0.63285</td>
<td>1.14354</td>
<td>1.02358</td>
</tr>
<tr>
<td>Oklahoma Gas &amp; Electric</td>
<td>0.73158</td>
<td>0.17122</td>
<td>0.72977</td>
<td>0.58391</td>
</tr>
<tr>
<td>Wisconsin Energy</td>
<td>0.53381</td>
<td>0.37405</td>
<td>0.51969</td>
<td>0.37522</td>
</tr>
<tr>
<td>Green Mountain Power</td>
<td>0.15377</td>
<td>0.64869</td>
<td>0.17360</td>
<td>0.13958</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Company</th>
<th>Minnesota_Power___Light</th>
<th>Iowa_Ill_Gas___Electric</th>
<th>Pennsylvania_Power___Light</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minnesota Power &amp; Light</td>
<td>0.00000</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Iowa-Ill Gas &amp; Electric</td>
<td>0.36368</td>
<td>0.00000</td>
<td>.</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>0.99384</td>
<td>0.75596</td>
<td>0.00000</td>
</tr>
<tr>
<td>Oklahoma Gas &amp; Electric</td>
<td>0.50744</td>
<td>0.19673</td>
<td>0.60216</td>
</tr>
<tr>
<td>Wisconsin Energy</td>
<td>0.36319</td>
<td>0.30259</td>
<td>0.76085</td>
</tr>
<tr>
<td>Green Mountain Power</td>
<td>0.19370</td>
<td>0.52083</td>
<td>1.09269</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Company</th>
<th>Oklahoma_Gas___Electric</th>
<th>Wisconsin_Energy</th>
<th>Green_Mountain_Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minnesota Power &amp; Light</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Iowa-Ill Gas &amp; Electric</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Oklahoma Gas &amp; Electric</td>
<td>0.00000</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Wisconsin Energy</td>
<td>0.28070</td>
<td>0.00000</td>
<td>.</td>
</tr>
<tr>
<td>Green Mountain Power</td>
<td>0.64175</td>
<td>0.44814</td>
<td>0</td>
</tr>
</tbody>
</table>
Output 40.2.2  Pseudo F versus Number of Clusters When METHOD=WARD

Output 40.2.3  Pseudo F versus Number of Clusters When METHOD=AVERAGE
Output 40.2.4 Dendrogram of Semipartial R-Square Values When METHOD=WARD
Output 40.2.5  Dendrogram of Average Distance between Clusters When METHOD=AVG

References


Subject Index

absolute level of measurement, definition
  DISTANCE procedure, 2673
asymmetric binary variable
  DISTANCE procedure, 2673

Binary Lance and Williams nonmetric coefficient
  DISTANCE procedure, 2699
Bray and Curtis coefficient
  DISTANCE procedure, 2699
Canberra metric coefficient
  DISTANCE procedure, 2699
Chebychev distance coefficient
  DISTANCE procedure, 2695
chi-square coefficient
  DISTANCE procedure, 2695, 2696
City-block distance coefficient
  DISTANCE procedure, 2695
correlation dissimilarity coefficient
  DISTANCE procedure, 2694
correlation similarity coefficient
  DISTANCE procedure, 2694
cosine coefficient
  DISTANCE procedure, 2695
covariance similarity coefficient
  DISTANCE procedure, 2694
Czekanowski/Sorensen similarity coefficient
  DISTANCE procedure, 2699
Dice coefficient
  DISTANCE procedure, 2699
dot Product coefficient
  DISTANCE procedure, 2695
Euclidean distance coefficient
  DISTANCE procedure, 2694
examples, 2674, 2701, 2707
extension of binary variable, 2673
formulated values, 2700
formulas for proximity measures, 2693
frequencies, 2691
functional summary, 2680
fuzz factor, 2681
generalized Euclidean distance coefficient
  DISTANCE procedure, 2695
Gower's dissimilarity coefficient
  DISTANCE procedure, 2694
Gower's similarity coefficient
  DISTANCE procedure, 2693
Hamann coefficient
  DISTANCE procedure, 2697
Hamming distance coefficient
  DISTANCE procedure, 2697
identity transformation, 2673
initial estimates for A-estimates, 2681
interval level of measurement, 2672
Jaccard dissimilarity coefficient
  DISTANCE procedure, 2699
Jaccard similarity coefficient
  DISTANCE procedure, 2698
Kulcynski 1 coefficient
  DISTANCE procedure, 2699
Lance-Williams nonmetric coefficient
  DISTANCE procedure, 2695
levels of measurement, 2672
linear transformation, 2673
log-interval level of measurement, 2673
many-to-one transformation, 2672
Minkowski L_p distance coefficient
  DISTANCE procedure, 2694
missing values, 2684–2686, 2689, 2699
monotone increasing transformation, 2672
normal level of measurement, 2672
nominal variable, 2673
normalization, 2684, 2686
one-to-one transformation, 2672
ordinal level of measurement, 2672
output data sets, 2684, 2685, 2700
Overlap dissimilarity coefficient
  DISTANCE procedure, 2695
Overlap similarity coefficient, 2695
phi-square coefficient, 2696
power distance coefficient, 2695
power transformation, 2673
ratio level of measurement, 2673
Roger and Tanimoto coefficient, 2697
Russell and Rao similarity coefficient, 2699
scaling variables, 2674
shape distance coefficient, 2694
similarity ratio coefficient, 2695
simple Matching coefficient, 2697
simple matching dissimilarity coefficient, 2697
size distance coefficient, 2694
Sokal and Sneath 1 coefficient, 2697
Sokal and Sneath 3 coefficient, 2697
squared correlation dissimilarity coefficient, 2694
squared correlation similarity coefficient, 2694
squared Euclidean distance coefficient, 2694
squared simple matching dissimilarity coefficient, 2697
standardization methods, 2688
standardization suppression, 2684
standardization with frequencies, 2691
standardization with weights, 2691
standardization, default methods, 2674, 2688
standardization, example, 2676
standardization, mandatory, 2674
strictly increasing transformation, 2672
summary of options, 2680
symmetric binary variable, 2673
transforming ordinal variables to interval, 2673, 2685
weights, 2690, 2691
dot Product coefficient
DISTANCE procedure, 2695
Euclidean distance coefficient
DISTANCE procedure, 2694
formatted values
DISTANCE procedure, 2700
generalized Euclidean distance coefficient
DISTANCE procedure, 2695
Gower’s dissimilarity coefficient
DISTANCE procedure, 2694
Gower’s similarity coefficient
DISTANCE procedure, 2693
Hamann coefficient
DISTANCE procedure, 2697
Hamming distance coefficient
DISTANCE procedure, 2697
interval level of measurement
DISTANCE procedure, 2672
Jaccard dissimilarity coefficient
DISTANCE procedure, 2699
Jaccard similarity coefficient
DISTANCE procedure, 2698
Kulczynski 1 coefficient
DISTANCE procedure, 2699
Lance-Williams nonmetric coefficient
DISTANCE procedure, 2695
levels of measurement
DISTANCE procedure, 2672
log-interval level of measurement
DISTANCE procedure, 2673
Minkowski L_p distance coefficient
DISTANCE procedure, 2694
missing values
DISTANCE procedure, 2684–2686, 2689, 2699
nominal level of measurement
DISTANCE procedure, 2672
nominal variable
DISTANCE procedure, 2673
ordinal level of measurement
DISTANCE procedure, 2672
ordinal variables
transformed to interval (RANKSCORE=), 2685
Overlap dissimilarity coefficient
DISTANCE procedure, 2695
Overlap similarity coefficient
DISTANCE procedure, 2695
phi-square coefficient
DISTANCE procedure, 2696
power distance coefficient
DISTANCE procedure, 2695
proximity measures
available methods for computing (DISTANCE), 2681
formulas(DISTANCE), 2693
ratio level of measurement
DISTANCE procedure, 2673
Roger and Tanimoto coefficient
DISTANCE procedure, 2697
Russell and Rao similarity coefficient
DISTANCE procedure, 2699
scaling variables
DISTANCE procedure, 2674
shape distance coefficient
similarity ratio coefficient
DISTANCE procedure, 2695
simple Matching coefficient
DISTANCE procedure, 2697
simple matching dissimilarity coefficient
DISTANCE procedure, 2697
size distance coefficient
DISTANCE procedure, 2694
Sokal and Sneath 1 coefficient
DISTANCE procedure, 2697
Sokal and Sneath 3 coefficient
DISTANCE procedure, 2697
squared correlation dissimilarity coefficient
DISTANCE procedure, 2694
squared correlation similarity coefficient
DISTANCE procedure, 2694
squared Euclidean distance coefficient
DISTANCE procedure, 2694
squared simple matching dissimilarity coefficient
DISTANCE procedure, 2697
standardization
comparisons between DISTANCE and STDIZE procedures, 2674
standardization suppression
DISTANCE procedure, 2684
STD= option (DISTANCE), 2688
symmetric binary variable
DISTANCE procedure, 2673
transformations
affine(DISTANCE), 2672
identity(DISTANCE), 2673
linear(DISTANCE), 2673
many-to-one(DISTANCE), 2672
monotone increasing(DISTANCE), 2672
one-to-one(DISTANCE), 2672
power(DISTANCE), 2673
strictly increasing(DISTANCE), 2672
transforming ordinal variables to interval
DISTANCE procedure, 2673
Syntax Index

ABSENT= option
   PROC DISTANCE statement, 2681
   VAR statement, 2689
ADD= option
   PROC DISTANCE statement, 2681
BY statement
   DISTANCE procedure, 2691
COPY statement
   DISTANCE procedure, 2691
DATA= option
   PROC DISTANCE statement, 2681
   DISTANCE procedure, BY statement, 2691
   DISTANCE procedure, COPY statement, 2691
   DISTANCE procedure, FREQ statement, 2691
   DISTANCE procedure, ID statement, 2690
   DISTANCE procedure, PROC DISTANCE statement, 2680
      ABSENT= option, 2681
      ADD= option, 2681
      DATA= option, 2681
      FUZZ= option, 2681
      INITIAL= option, 2681
      METHOD= option, 2681
      NOMISS, 2684
      NORM option, 2684
      NOSTD, 2684
      OUT= option, 2684
      OUTSDZ= option, 2685
      PREFIX= option, 2685
      RANKSCORE= option, 2685
      REPLACE, 2685
      REPONLY, 2686
      SHAPE= option, 2686
      SNORM option, 2686
      STDONLY option, 2686
      UNDEF= option, 2686
      VARDEF= option, 2686
DISTANCE procedure, VAR statement
   ABSENT= option, 2689
   MISSING= option, 2689
   ORDER= option, 2690
   WEIGHTS= option, 2690
DISTANCE procedure, WGT statement, 2691
FREQ statement
   DISTANCE procedure, 2691
FUZZ= option
   PROC DISTANCE statement, 2681
ID statement
   DISTANCE procedure, 2690
INITIAL= option
   PROC DISTANCE statement, 2681
METHOD= option
   PROC DISTANCE statement, 2681
MISSING= option
   VAR statement, 2689
MULT= option
   PROC DISTANCE statement, 2684
NOMISS option
   PROC DISTANCE statement, 2684
NORM option
   PROC DISTANCE statement, 2684
NOSTD option
   PROC DISTANCE statement, 2684
ORDER= option
   VAR statement, 2690
OUT= option
   PROC DISTANCE statement, 2684
OUTSDZ= option
   PROC DISTANCE statement, 2685
PREFIX= option
   PROC DISTANCE statement, 2685
PROC DISTANCE statement, see DISTANCE procedure
   RANKSCORE= option
   PROC DISTANCE statement, 2685
   REPLACE option
   PROC DISTANCE statement, 2685
   REPONLY option
   PROC DISTANCE statement, 2686
SHAPE= option
   PROC DISTANCE statement, 2686
SNORM option
   PROC DISTANCE statement, 2686
STDONLY option
   PROC DISTANCE statement, 2686
UNDEF= option
PROC DISTANCE statement, 2686

VARDEF= option
    PROC DISTANCE statement, 2686

WEIGHTS= option
    VAR statement, 2690

WGT statement
    DISTANCE procedure, 2691