Chapter 59
The STDIZE Procedure

Chapter Table of Contents

Chapter 59
The STDIZE Procedure

Overview

The STDIZE procedure standardizes one or more numeric variables in a SAS data set by subtracting a location measure and dividing by a scale measure. A variety of location and scale measures are provided, including estimates that are resistant to outliers and clustering. Some of the well-known standardization methods such as mean, median, std, range, Huber's estimate, Tukey's biweight estimate, and Andrew's wave estimate are available in the STDIZE procedure.

In addition, you can multiply each standardized value by a constant and add a constant. Thus, the final output value is

$$result = add + multiply \times \frac{(original - location)}{scale}$$

where

result	= final output value
add	= constant to add (ADD= option)
multiply	= constant to multiply by (MULT= option)
original	= original input value
location	= location measure
scale	= scale measure

PROC STDIZE can also find quantiles in one pass of the data, a capability that is especially useful for very large data sets. With such data sets, the UNIVARIATE procedure may have high or excessive memory or time requirements.

Getting Started

The following example demonstrates how you can use the STDIZE procedure to obtain location and scale measures of your data.

In the following hypothetical data set, a random sample of grade 12 students is selected from a number of co-educational schools. Each school is classified as one of two types: Urban or Rural. There are 40 observations.

The variables are id (student identification), Type (type of school attended: 'urban'=urban area and 'rural'=rural area), and total (total assessment scores in History, Geometry, and Chemistry).

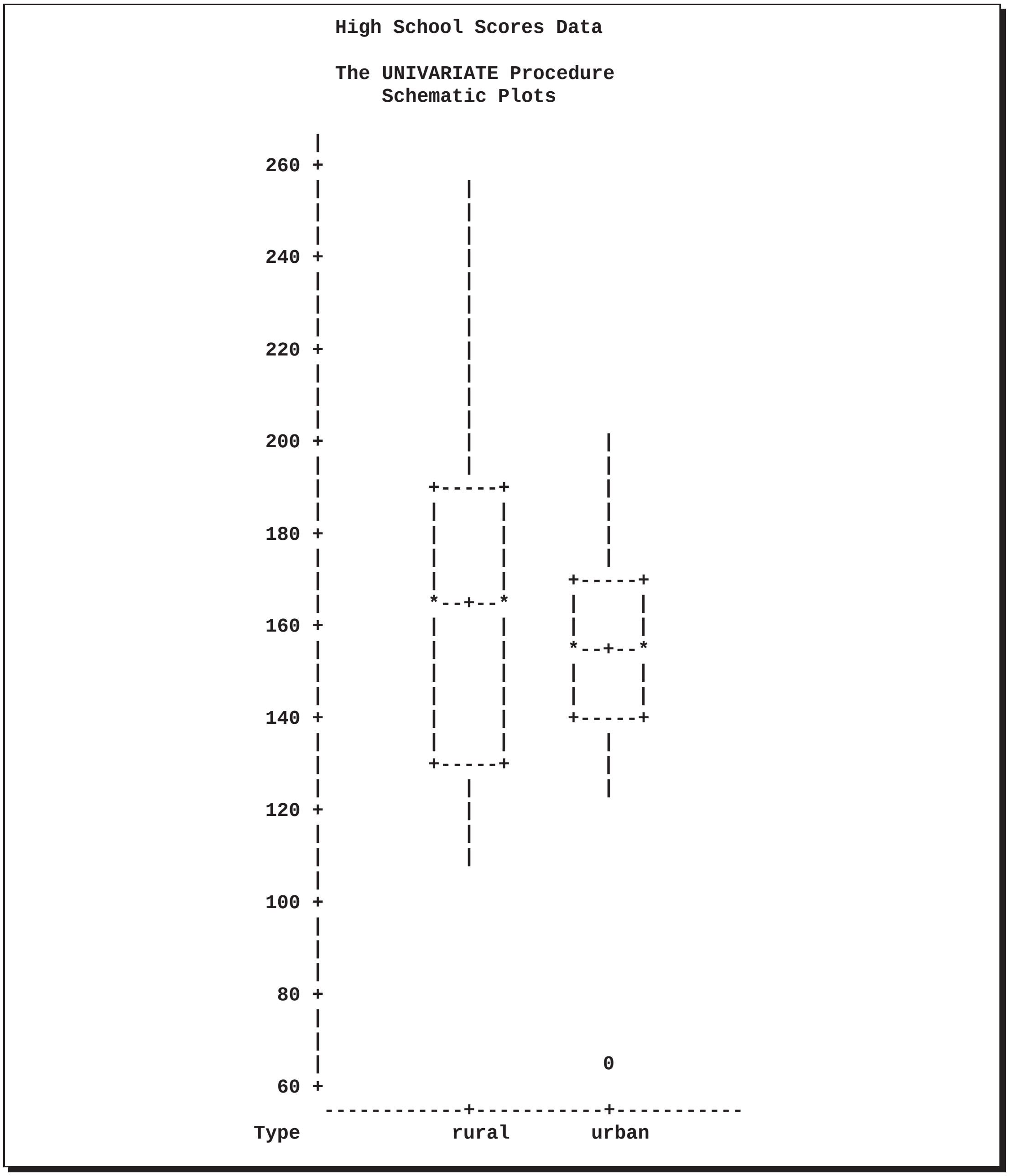
The following DATA step creates the SAS data set TotalScores.

```
data TotalScores;
   title 'High School Scores Data';
   input id Type $  total;
   datalines;
 1      rural        135
 2      rural        125
 3      rural        223
 4      rural        224
 5      rural        133
 6      rural        253
 7      rural        144
 8      rural        193
 9      rural        152
10      rural        178
11      rural        120
12      rural        180
13      rural        154
14      rural        184
15      rural        187
16      rural        111
17      rural        190
18      rural        128
19      rural        110
20      rural        217
21      urban        192
22      urban        186
23      urban         64
24      urban        159
25      urban        133
26      urban        163
27      urban        130
28      urban        163
29      urban        189
30      urban        144
31      urban        154
32      urban        198
33      urban        150
34      urban        151
35      urban        152
36      urban        151
37      urban        127
38      urban        167
39      urban        170
40      urban        123
;
run;
```

Suppose you would now like to standardize the total scores in different types of schools prior to any further analysis. Before standardizing the total scores, you can use the Schematic Plots from PROC UNIVARIATE to summarize the total scores for both types of schools.

```
proc univariate data=TotalScores plot;
   var total;
   by Type;
run;
```

The PLOT option in the PROC UNIVARIATE statement creates the Schematic Plots and several other types of plots. The Schematic Plots display side-by-side box plots for each BY group (Figure 59.1). The vertical axis represents the total scores, and the horizontal axis displays two box plots: the one on the left is for the rural scores and the one on the right is for the urban scores.



```
       High School Scores Data

       The UNIVARIATE Procedure
           Schematic Plots

    |
260 +
    |            |
    |            |
    |            |
240 +            |
    |            |
    |            |
    |            |
220 +            |
    |            |
    |            |
    |            |
200 +            |           |
    |            |           |
    |         +-----+        |
    |         |     |        |
180 +         |     |        |
    |         |     |        |
    |         |     |     +-----+
    |         *--+--*     |     |
160 +         |     |     |     |
    |         |     |     *--+--*
    |         |     |     |     |
    |         |     |     |     |
140 +         |     |     +-----+
    |         |     |        |
    |         +-----+        |
    |            |           |
120 +            |
    |            |
    |            |
    |
100 +
    |
    |
    |
 80 +
    |
    |
    |                        0
 60 +
      ------------+-----------+-----------
Type             rural       urban
```

Figure 59.1. Schematic Plots from PROC UNIVARIATE

Inspection reveals that one urban score is a low outlier. Also, if you compare the lengths of two boxplots, there seems to be twice as much dispersion for the rural scores as for the urban scores.

```
                            High School Scores Data

---------------------------------- Type=urban ----------------------------------

                            The UNIVARIATE Procedure
                                Variable:  total

                              Extreme Observations

                      ----Lowest----        ----Highest---

                      Value      Obs        Value      Obs

                         64        3          170       19
                        123       20          186        2
                        127       17          189        9
                        130        7          192        1
                        133        5          198       12
```

Figure 59.2. Table for Extreme Observations when Type=urban

Figure 59.2 displays the table from PROC UNIVARIATE for the lowest and highest five total scores for urban schools. The outlier (Obs = 3), marked in Figure 59.1 by the symbol ‘0’, has a score of 64.

The following statements use the traditional standardization method (METHOD=STD) to compute the location and scale measures:

```
proc stdize data=totalscores method=std pstat;
   title2 ’METHOD=STD’;
   var total;
   by Type;
run;
```

```
                            High School Scores Data
                                   METHOD=STD

---------------------------------- Type=rural ----------------------------------

                              The STDIZE Procedure

               Location and Scale Measures

               Location = mean     Scale = standard deviation

               Name         Location           Scale           N

               total      167.050000       41.956713          20


                            High School Scores Data
                                   METHOD=STD

---------------------------------- Type=urban ----------------------------------

                              The STDIZE Procedure

               Location and Scale Measures

               Location = mean     Scale = standard deviation

               Name         Location           Scale           N

               total      153.300000       30.066768          20
```

Figure 59.3. Location and Scale Measures Table when METHOD=STD

Figure 59.3 displays the table of location and scale measures from the PROC STDIZE statement. PROC STDIZE uses the mean as the location measure and the standard deviation as the scale measure for standardizing. The PSTAT option displays this table; otherwise, no display is created.

The ratio of the scale of rural scores to the scale of urban scores is approximately 1.4 (41.96/30.07). This ratio is smaller than the dispersion ratio observed in the previous Schematic Plots.

The STDIZE procedure provides several location and scale measures that are resistant to outliers. The following statements invoke three different standardization methods and display the Location and Scale Measures tables:

```
proc stdize data=totalscores method=mad pstat;
   title2 ’METHOD=MAD’;
   var total;
   by Type;
run;

proc stdize data=totalscores method=iqr pstat;
   title2 ’METHOD=IQR’;
   var total;
   by Type;
run;
```

```
proc stdize data=totalscores method=abw(4) pstat;
   title2 'METHOD=ABW(4)';
   var total;
   by Type;
run;
```

The results from this analysis are displayed in the following figures.

```
                            High School Scores Data
                                   METHOD=MAD

---------------------------------- Type=rural ----------------------------------

                              The STDIZE Procedure

               Location and Scale Measures

            Location = median     Scale = median abs dev from median

               Name         Location           Scale           N

               total      166.000000       32.000000          20


                            High School Scores Data
                                   METHOD=MAD

---------------------------------- Type=urban ----------------------------------

                              The STDIZE Procedure

               Location and Scale Measures

            Location = median     Scale = median abs dev from median

               Name         Location           Scale           N

               total      153.000000       15.500000          20
```

Figure 59.4. Location and Scale Measures Table when METHOD=MAD

Figure 59.4 displays the table of location and scale measures when the standardization method is MAD. The location measure is the median, and the scale measure is the median absolute deviation from median. The ratio of the scale of rural scores to the scale of urban scores is approximately 2.06 (32.0/15.5) and is close to the dispersion ratio observed in the previous Schematic Plots.

```
                            High School Scores Data
                                   METHOD=IQR

---------------------------------- Type=rural ----------------------------------

                              The STDIZE Procedure

               Location and Scale Measures

               Location = median     Scale = interquartile range

               Name          Location           Scale           N

               total       166.000000       61.000000          20


                            High School Scores Data
                                   METHOD=IQR

---------------------------------- Type=urban ----------------------------------

                              The STDIZE Procedure

               Location and Scale Measures

               Location = median     Scale = interquartile range

               Name          Location           Scale           N

               total       153.000000       30.000000          20
```

Figure 59.5. Location and Scale Measures Table when METHOD=IQR

Figure 59.5 displays the table of location and scale measures when the standardization method is IQR. The location measure is the median, and the scale measure is the interquartile range. The ratio of the scale of rural scores to the scale of urban scores is approximately 2.03 (61/30) and is, in fact, the dispersion ratio observed in the previous Schematic Plots.

```
                             High School Scores Data
                                  METHOD=ABW(4)

---------------------------------- Type=rural ----------------------------------

                               The STDIZE Procedure

               Location and Scale Measures

     Location = biweight 1-step M-estimate     Scale = biweight A-estimate

               Name         Location           Scale           N

               total      162.889603       56.662855          20


                             High School Scores Data
                                  METHOD=ABW(4)

---------------------------------- Type=urban ----------------------------------

                               The STDIZE Procedure

               Location and Scale Measures

     Location = biweight 1-step M-estimate     Scale = biweight A-estimate

               Name         Location           Scale           N

               total      156.014608       28.615980          20
```

Figure 59.6. Location and Scale Measures Table when METHOD=ABW

Figure 59.6 displays the table of location and scale measures when the standardization method is ABW. The location measure is the biweight 1-step M-estimate, and the scale measure is the biweight A-estimate. Note that the initial estimate for ABW is MAD. The tuning constant (4) of ABW is obtained by the following steps:

1. For rural scores, the location estimate for MAD is 166.0 and the scale estimate for MAD is 32.0. The maximum of the rural scores is 253 (not shown), and the minimum is 110 (not shown). Thus, the tuning constant needs to be 3 so that it does not reject any observation that has a score between 110 to 253.
2. For urban scores, the location estimate for MAD is 153.0 and the scale estimate for MAD is 15.5. The maximum of the rural scores is 198, and the minimum (also an outlier) is 64. Thus, the tuning constant needs to be 4 so that it rejects the outlier (64) but includes the maximum (198) as an normal observation.
3. The maximum of the tuning constants, obtained in steps 1 and 2, is 4.

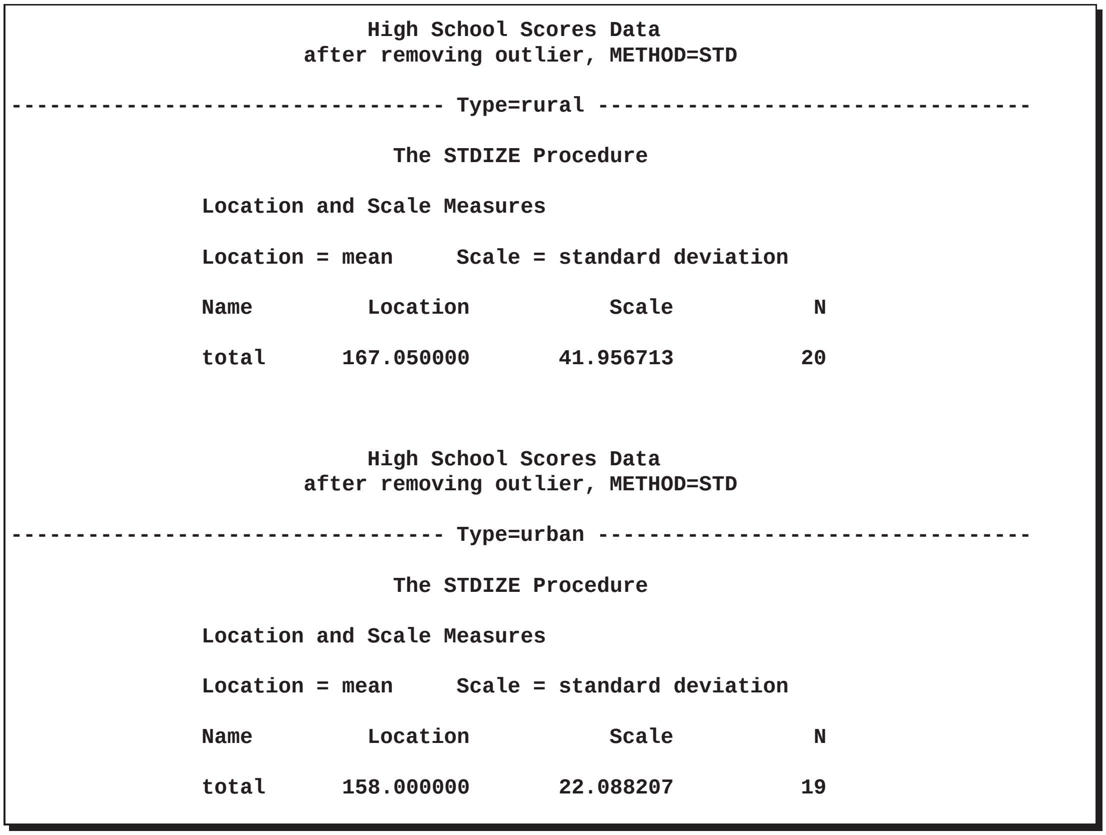
Refer to Goodall (1983, Chapter 11) for details on the tuning constant. The ratio of the scale of rural scores to the scale of urban scores is approximately 2.06 (32.0/15.5). It is also close to the dispersion ratio observed in the previous Schematic Plots.

The preceding analysis shows that METHOD=MAD, METHOD=IQR, and METHOD=ABW all provide better dispersion ratios than does METHOD=STD.

You can recompute the standard deviation after deleting the outlier from the original data set for comparison. The following statements create a DATA set NoOutlier that excludes the outlier from the TotalScores data set and invoke PROC STDIZE with METHOD=STD.

```
data NoOutlier;
   set totalscores;
   if (total = 64) then delete;
run;

proc stdize data=NoOutlier method=std pstat;
   title2 'after removing outlier, METHOD=STD';
   var total;
   by Type;
run;
```



```
                         High School Scores Data
                    after removing outlier, METHOD=STD

---------------------------------- Type=rural ----------------------------------

                           The STDIZE Procedure

               Location and Scale Measures

               Location = mean     Scale = standard deviation

               Name         Location           Scale           N

               total      167.050000       41.956713          20


                         High School Scores Data
                    after removing outlier, METHOD=STD

---------------------------------- Type=urban ----------------------------------

                           The STDIZE Procedure

               Location and Scale Measures

               Location = mean     Scale = standard deviation

               Name         Location           Scale           N

               total      158.000000       22.088207          19
```

Figure 59.7. After Deleting the Outlier, Location and Scale Measures Table when METHOD=STD

Figure 59.7 displays the location and scale measures after deleting the outlier. The lack of resistance of the standard deviation to outliers is clearly illustrated: if you delete the outlier, the sample standard deviation of urban scores changes from 30.07 to 22.09. The new ratio of the scale of rural scores to the scale of urban scores is approximately 1.90 (41.96/22.09).

Syntax

The following statements are available in the STDIZE procedure.

PROC STDIZE $<$ *options* $>$ **;**

BY *variables* **;**
FREQ *variable* **;**
LOCATION *variables* **;**
SCALE *variables* **;**
VAR *variables* **;**
WEIGHT *variable* **;**

The PROC STDIZE statement is required. The BY, LOCATION, FREQ, VAR, SCALE, and WEIGHT statements are described in alphabetical order following the PROC STDIZE statement.

PROC STDIZE Statement

PROC STDIZE $<$ *options* $>$ **;**

The PROC STDIZE statement invokes the procedure. You can specify the following options in the PROC STDIZE statement.

Table 59.1. Summary of PROC STDIZE Statement Options

Task	**Options**	**Description**
Specify standardization methods	METHOD=	specifies the name of the standardization method
	INITIAL=	specifies the method for computing initial estimates for the A estimates
Unstandardize variables	UNSTD	unstandardizes variables when you also specify the METHOD=IN option
Process missing values	NOMISS	omits observations with any missing values from computation
	MISSING=	specifies the method or a numeric value for replacing missing values
	REPLACE	replaces missing data by zero in the standardized data
	REPONLY	replaces missing data by the location measure (does not standardize the data)
Specify data set details	DATA=	specifies the input data set
	OUT=	specifies the output data set
	OUTSTAT=	specifies the output statistic data set
Specify computational settings	VARDEF=	specifies the variances divisor

Table 59.1. (continued)

Task	Options	Description
	NMARKERS=	specifies the number of markers when you also specify PCTLMTD=ONEPASS
	MULT=	specifies the constant to multiply each value by after standardizing
	ADD=	specifies the constant to add to each value after standardizing and multiplying by the value specified in the MULT= option
	FUZZ=	specifies the relative fuzz factor for writing the output
Specify percentiles	PCTLDEF=	specifies the definition of percentiles when you also specify the PCTLMTD=ORD_STAT option
	PCTLMTD=	specifies the method used to estimate percentiles
	PCTLPTS=	writes observations containing percentiles to the data set specified in the OUTSTAT= option
Normalize scale estimators	NORM	normalizes the scale estimator to be consistent for the standard deviation of a normal distribution
	SNORM	normalizes the scale estimator to have an expectation of approximately 1 for a standard normal distribution
Specify output	PSTAT	displays the location and scale measures

These options and their abbreviations are described, in alphabetical order, in the remainder of this section.

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=*SAS-data-set*

specifies the input data set to be standardized. If you omit the DATA= option, the most recently created data set is used.

FUZZ=*c*

specifies the relative fuzz factor. The default value is 1E-14. For the OUT= data set, the score is computed as follows:

$$\text{if } |\text{Result}| < \text{Scale} \times \text{Fuzz}, \text{then Result} = 0$$

For the OUTSTAT= data set and the Location and Scale table, the scale and location values are computed as follows:

$$\text{if Scale} < |\text{Location}| \times \text{Fuzz, then Scale} = 0$$

Otherwise,

$$\text{if } |\text{Location}| < \text{Scale} \times \text{Fuzz, then Location} = 0$$

INITIAL=*method*

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

METHOD=*name*

specifies the name of the method for computing location and scale measures. Valid values for *name* are as follows: MEAN, MEDIAN, SUM, EUCLEN, USTD, STD, RANGE, MIDRANGE, MAXABS, IQR, MAD, ABW, AHUBER, AWAVE, AGK, SPACING, L, and IN.

For details on these methods, see the descriptions in the “Standardization Methods” section on page 3134. The default is METHOD=STD.

MISSING= *method*
MISSING= *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 METHOD= option. Specify the MISSING= option when you want to replace missing values with a different value. You can specify any name that is valid in the METHOD= option except the name IN. 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, you can specify the REPONLY option with the MISSING= option to suppress standardization for cases in which you want only to replace missing values.

MULT= *c*

specifies a constant, *c*, by which to multiply each value after standardizing. The default value is 1.

NMARKERS= *n*

specifies the number of markers used when you specify the one-pass algorithm (PCTLMTD=ONEPASS). The value n must be greater than or equal to 5. The default value is 105.

NOMISS

omits observations with missing values for any of the analyzed variables from calculation of the location and scale measures. If you omit the NOMISS option, all nonmissing values are used.

NORM

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

OUT=*SAS-data-set*

specifies the name of the SAS data set created by PROC STDIZE. The output data set is a copy of the DATA= data set except that the analyzed variables have been standardized. Note that analyzed variables are those specified in the VAR statement or, if there is no VAR statement, all numeric variables not listed in any other statement. See the section “Output Data Sets” on page 3139 for more information.

If you want to create a permanent SAS data set, you must specify a two-level name. (Refer to “SAS Files” in *SAS Language Reference: Concepts* for more information on permanent SAS data sets.)

If you omit the OUT= option, PROC STDIZE creates an output data set named according to the DATAn convention.

OUTSTAT=*SAS-data-set*

specifies the name of the SAS data set containing the location and scale measures and other computed statistics. See the section “Output Data Sets” on page 3139 for more information.

PCTLDEF= *percentiles*

specifies which of five definitions is used to calculate percentiles when you specify the option PCTLMTD=ORD_STAT. By default, PCTLDEF=5.

Note that the option PCTLMTD=ONEPASS implies a specification of PCTLDEF=5. See the section “Computational Methods for the PCTLDEF= Option” on page 3137 for details on the PCTLDEF= option.

PCTLMTD=ORD_STAT
PCTLMTD=ONEPASS | P2

specifies the method used to estimate percentiles. Specify the PCTLMTD=ORD_STAT option to compute the percentiles by the order statistics method. The PCTLMTD=ONEPASS option modifies an algorithm invented by Jain and Chlamtac (1985). See the “Computing Quantiles” section on page 3137 for more details on this algorithm.

PCTLPTS= *n*

writes percentiles to the OUTSTAT= data set. Values of *n* can be any decimal number between 0 and 100, inclusive.

A requested percentile is identified by the _TYPE_ variable in the OUTSTAT= data set with a value of P*n*. For example, suppose you specify the option PCTLPTS=10, 30. The corresponding observations in the OUTSTAT= data set that contain the 10th and the 30th percentiles would then have values _TYPE_=P10 and _TYPE_=P30, respectively.

PSTAT

displays the location and scale measures.

REPLACE

replaces missing data with the value 0 in the standardized data (this value corresponds to the location measure before standardizing). To replace missing data by other values, see the preceding description of the MISSING= option. You cannot specify both the REPLACE and REPONLY options.

REPONLY

replaces missing data only; PROC STDIZE does not standardize the data. Missing values are replaced with the location measure unless you also specify the MISSING=*value* option, in which case missing values are replaced with *value*. You cannot specify both the REPLACE and REPONLY options.

SNORM

normalizes the scale estimator to have an expectation of approximately 1 for a standard normal distribution when you specify the METHOD=SPACING option.

UNSTD
UNSTDIZE

unstandardizes variables when you specify the METHOD=IN(*ds*) option. The location and scale measures, along with constants for addition and multiplication that the unstandardization is based upon, are identified by the _TYPE_ variable in the ds data set.

The ds data set must have a _TYPE_ variable and contain the following two observations: a _TYPE_= ‘LOCATION’ observation and a _TYPE_= ‘SCALE’ observation. The variable _TYPE_ can also contain the optional observations, ‘ADD’ and ‘MULT’; if these observations are not found in the ds data set, the constants specified in the ADD= and MULT= options (or their default values) are used for unstandardization.

See the “OUTSTAT= Data Set” section on page 3139 for details on the statistics that each value of _TYPE_ represents. The formula used for unstandardization is as follows: If the final output value from the previous standardization is calculated as

$$result = add + multiply \times \frac{(original - location)}{scale}$$

then the original value is reconstructed as

$$original = scale \times \frac{(result - add)}{multiply} + location$$

VARDEF= DF
VARDEF= N
VARDEF= WDF
VARDEF= WEIGHT | WGT

specifies the divisor to be used in the calculation of variances. By default, VARDEF=DF. The values and associated divisors are as follows.

Value	**Divisor**	**Formula**
DF	degrees of freedom	$n - 1$
N	number of observations	n
WDF	sum of weights minus 1	$(\sum_i w_i) - 1$
WEIGHT | WGT	sum of weights	$\sum_i w_i$

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC STDIZE to obtain separate standardization for observations in groups defined by the BY variables.

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

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

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

When you specify the option METHOD=IN(ds), the following rules are applied to BY-group processing:

- If the ds data set does not contain any of the BY variables, the entire DATA= data set is standardized by the location and scale measures (along with the constants for addition and multiplication) in the ds data set.
- If the ds data set contains some, but not all, of the BY variables or if some BY variables do not have the same type or length in the ds data set that they have in the DATA= data set, PROC STDIZE displays an error message and stops.
- If all of the BY variables appear in the ds data set with the same type and length as in the DATA= data set, each BY group in the DATA= data set is standardized using the location and scale measures (along with the constants for addition and multiplication) from the corresponding BY group in the ds data set. The BY groups in the ds data set must be in the same order as they appear in the DATA= data set. All BY groups in the DATA= data set must also appear in the ds data set. If you do not specify the NOTSORTED option, some BY groups can appear in the ds data set but not in the DATA= data set; such BY groups are not used in standardizing data.

FREQ Statement

FREQ | **FREQUENCY** *variable* **;**

If one variable in the input data set represents the frequency of occurrence for other values in the observation, specify the variable name in a FREQ statement. PROC STDIZE 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.

LOCATION Statement

LOCATION *variables* **;**

The LOCATION statement specifies a list of numeric variables that contain location measures in the input data set specified by the METHOD=IN option.

SCALE Statement

SCALE *variables* **;**

The SCALE statement specifies the list of numeric variables containing scale measures in the input data set specified by the METHOD=IN option.

VAR Statement

VAR | **VARIABLES** *variables* **;**

The VAR statement lists numeric variables to be standardized. If you omit the VAR statement, all numeric variables not listed in the BY, FREQ, and WEIGHT statements are used.

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. 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 only when you specify the option METHOD=MEAN, METHOD=SUM, METHOD=EUCLEN, METHOD=USTD, METHOD=STD, METHOD=AGK, or METHOD=L.

PROC STDIZE uses the value of the WEIGHT variable w_i, as follows.

The sample mean and (uncorrected) sample variances are computed as

$$\overline{x}_w = \sum_i w_i x_i / \sum_i w_i$$

$$u{s_w}^2 = \sum_i w_i {x_i}^2 / d$$

$${s_w}^2 = \sum_i w_i (x_i - \overline{x}_w)^2 / d$$

where w_i is the weight value of the ith observation, x_i is the value of the ith observation, and d is the divisor controlled by the VARDEF= option (see the VARDEF= option for details).

PROC STDIZE uses the value of the WEIGHT variable to calculate the following statistics:

MEAN the weighted mean, $\overline{x}_w$

SUM the weighted sum, $\sum_i w_i x_i$

USTD the weighted uncorrected standard deviation, $\sqrt{us_w^2}$

STD the weighted standard deviation, $\sqrt{s_w^2}$

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 16, “The ACECLUS Procedure,” for information on 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 27, “The FASTCLUS Procedure,” for information on how the WEIGHT variable is used to compute weighted cluster means. Note that the number of clusters is always 1.

Details

Standardization Methods

The following table lists standardization methods and their corresponding location and scale measures available with the METHOD= option.

Table 59.2. Available Standardization Methods

Method	**Location**	**Scale**
MEAN	mean	1
MEDIAN	median	1
SUM	0	sum
EUCLEN	0	Euclidean length
USTD	0	standard deviation about origin
STD	mean	standard deviation
RANGE	minimum	range
MIDRANGE	midrange	range/2
MAXABS	0	maximum absolute value
IQR	median	interquartile range
MAD	median	median absolute deviation from median
ABW(c)	biweight 1-step M-estimate	biweight A-estimate
AHUBER(c)	Huber 1-step M-estimate	Huber A-estimate
AWAVE(c)	Wave 1-step M-estimate	Wave A-estimate
AGK(p)	mean	AGK estimate (ACECLUS)
SPACING(p)	mid minimum-spacing	minimum spacing
L(p)	L(p)	L(p)
IN(ds)	read from data set	read from data set

For METHOD=ABW(c), METHOD=AHUBER(c), or METHOD=AWAVE(c), c is a positive numeric tuning constant.

For METHOD=AGK(p), p is a numeric constant giving the proportion of pairs to be included in the estimation of the within-cluster variances.

For METHOD=SPACING(p), p is a numeric constant giving the proportion of data to be contained in the spacing.

For METHOD=L(p), p is a numeric constant greater than or equal to 1 specifying the power to which differences are to be raised in computing an L(p) or Minkowski metric.

For METHOD=IN(ds), ds is the name of a SAS data set that meets either one of the following two conditions:

- contains a _TYPE_ variable. The observation that contains the location measure corresponds to the value _TYPE_= 'LOCATION' and the observation that contains the scale measure corresponds to the value _TYPE_= 'SCALE'. You can also use a data set created by the OUTSTAT= option from another PROC STDIZE statement as the ds data set. See the section "Output Data Sets" on page 3139 for the contents of the OUTSTAT= data set.
- contains the location and scale variables specified by the LOCATION and SCALE statements.

PROC STDIZE reads in the location and scale variables in the ds data set by first looking for the _TYPE_ variable in the ds data set. If it finds this variable, PROC STDIZE continues to search for all variables specified in the VAR statement. If it does not find the _TYPE_ variable, PROC STDIZE searches for the location variables specified in the LOCATION statement and the scale variables specified in the SCALE statement.

For robust estimators, refer to Goodall (1983) and Iglewicz (1983). The MAD method has the highest breakdown point (50%), but it is somewhat inefficient. The ABW, AHUBER, and AWAVE methods provide a good compromise between breakdown and efficiency. The L(p) location estimates are increasingly robust as p drops from 2 (corresponding to least squares, or mean estimation) to 1 (corresponding to least absolute value, or median estimation). However, the L(p) scale estimates are not robust.

The SPACING method is robust to both outliers and clustering (Jannsen et al. 1995) and is, therefore, a good choice for cluster analysis or nonparametric density estimation. The mid-minimum spacing method estimates the mode for small p. The AGK method is also robust to clustering and more efficient than the SPACING method, but it is not as robust to outliers and takes longer to compute. If you expect g clusters, the argument to METHOD=SPACING or METHOD=AGK should be $\frac{1}{g}$ or less. The AGK method is less biased than the SPACING method for small samples. As a general guide, it is reasonable to use AGK for samples of size 100 or less and SPACING for samples of size 1000 or more, with the treatment of intermediate sample sizes depending on the available computer resources.

Computation of the Statistics

Formulas for statistics of METHOD=MEAN, METHOD=MEDIAN, METHOD=SUM, METHOD=USTD, METHOD=STD, METHOD=RANGE, and METHOD=IQR are given in the chapter on elementary statistics procedures in the *SAS Procedures Guide*.

Note that the computations of median and upper and lower quartiles depend on the PCTLMTD= option.

The other statistics listed in Table 59.2, except for METHOD=IN, are described as follows:

EUCLEN
: Euclidean length.
$\sqrt{\sum_{i=1}^{n} {x_i}^2}$ where x_i is the ith observation and n is the total number of observations in the sample.

L(p)
: Minkowski metric. This metric is documented as the LEAST=p option in the PROC FASTCLUS statement of the FASTCLUS procedure (see Chapter 27, “The FASTCLUS Procedure”).

If you specify METHOD=L(p) in the PROC STDIZE statement, your results are similar to those obtained from PROC FASTCLUS if you specify the LEAST=p option with MAXCLUS=1 (and use the default values of the MAXITER= option). The difference between the two types of calculations concerns the maximum number of iterations. In PROC STDIZE, it is a criteria for convergence on all variables; In PROC FASTCLUS, it is a criteria for convergence on a single variable.

The location and scale measures for L(p) are output to the OUTSEED= data set in PROC FASTCLUS.

MIDRANGE
: $(\text{maximum} + \text{minimum})/2$

ABW(c)
: Tukey’s biweight. Refer to Goodall (1983, pp. 376–378, p. 385) for the biweight 1-step M-estimate. Also refer to Iglewicz (1983, pp. 416–418) for the biweight A-estimate.

AHUBER(c)
: Hubers. Refer to Goodall (1983, pp. 371–374) for the Huber 1-step M-estimate. Also refer to Iglewicz (1983, pp. 416–418) for the Huber A-estimate of scale.

AWAVE(c)
: Andrews’ Wave. Refer to Goodall (1983, p. 376) for the Wave 1-step M-estimate. Also refer to Iglewicz (1983, pp. 416 –418) for the Wave A-estimate of scale.

AGK(p)
: the noniterative univariate form of the estimator described by Art, Gnanadesikan, and Kettenring (1982)

The AGK estimate is documented in the section on the METHOD= option in the PROC ACECLUS statement of the ACECLUS procedure (also see the “Background” section on page 304 in Chapter 16, “The ACECLUS Procedure”). Specifying METHOD=AGK(p) in the PROC STDIZE statement is the same as specifying METHOD=COUNT and P=p in the PROC ACECLUS statement.

SPACING(p) the absolute difference between two data values. The minimum spacing for a proportion p is the minimum absolute difference between two data values that contain a proportion p of the data between them. The mid minimum-spacing is the mean of these two data values.

Computing Quantiles

PROC STDIZE offers two methods for computing quantiles: the one-pass approach and the order-statistics approach (like that used in the UNIVARIATE procedure).

The one-pass approach used in PROC STDIZE modifies the P^2 algorithm for histograms proposed by Jain and Chlamtac (1985). The primary difference comes from the movement of markers. The one-pass method allows a marker to move to the right (or left) by more than one position (to the largest possible integer) as long as it does not result in two markers being in the same position. The modification is necessary in order to incorporate the FREQ variable.

You may obtain inaccurate results if you use the one-pass approach to estimate quantiles beyond the quartiles (that is, when you estimate quantiles $<$ P25 or $>$ P75). A large sample size (10,000 or more) is often required if the tail quantiles (quantiles <= P10 or >= P90) are requested. Note that, for variables with highly skewed or heavy-tailed distributions, tail quantile estimates may be inaccurate.

The order-statistics approach for estimating quantiles is faster than the one-pass method but requires that the entire data set be stored in memory. The accuracy in estimating the quantiles is comparable for both methods when the requested percentiles are between the lower and upper quartiles. The default is PCTLMTD=ORD_STAT if enough memory is available; otherwise, PCTLMTD=ONEPASS.

Computational Methods for the PCTLDEF= Option

You can specify one of five methods for computing quantile statistics when you use the order-statistics approach (PCTLMTD=ORD_STAT); otherwise, the PCTLDEF=5 method is used when you use the one-pass approach (PCTLMTD=ONEPASS).

Let n be the number of nonmissing values for a variable, and let $x_1, x_2, \ldots, x_n$ represent the ordered values of the variable. For the tth percentile, let $p = t/100$. In the following definitions numbered 1, 2, 3, and 5, let

$$np = j + g$$

where j is the integer part and g is the fractional part of np. For definition 4, let

$$(n+1)p = j + g$$

Given the preceding definitions, the tth percentile, y, is defined as follows:

PCTLDEF=1 weighted average at x_{np}

$$y = (1-g)x_j + gx_{j+1}$$

where x_0 is taken to be x_1

PCTLDEF=2 observation numbered closest to np

$$y = x_i$$

where i is the integer part of $np + 1/2$ if $g \neq 1/2$. If $g = 1/2$, then
$y = x_j$ if j is even, or
$y = x_{j+1}$ if j is odd

PCTLDEF=3 empirical distribution function

$$y = x_j \text{ if } g = 0$$

$$y = x_{j+1} \text{ if } g > 0$$

PCTLDEF=4 weighted average aimed at $x_{p(n+1)}$

$$y = (1-g)x_j + gx_{j+1}$$

where x_{n+1} is taken to be x_n

PCTLDEF=5 empirical distribution function with averaging

$$y = (x_j + x_{j+1})/2 \text{ if } g = 0$$

$$y = x_{j+1} \text{ if } g > 0$$

Missing Values

Missing values can be replaced by the location measure or by any specified constant (see the REPLACE option and the MISSING= option). You can also suppress standardization if you want only to replace missing values (see the REPONLY option).

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

Output Data Sets

OUT= 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 or, if there is no VAR statement, all numeric variables not listed in any other statement.

OUTSTAT= Data Set

The new data set contains the following variables:

- the BY variables, if any
- _TYPE_, a character variable
- the analyzed variables

Each observation in the new data set contains a type of statistic as indicated by the _TYPE_ variable. The values of the _TYPE_ variable are as follows:

TYPE	**Contents**
LOCATION	location measure of each variable
SCALE	scale measure of each variable
ADD	constant specified in the ADD= option. This value is the same for each variable.
MULT	constant specified in the MULT= option. This value is the same for each variable.
N	total number of nonmissing positive frequencies of each variable
NORM	norm measure of each variable. This observation is produced only when you specify the NORM option with METHOD=AGK, METHOD=IQR, METHOD=MAD, or METHOD=SPACING or when you specify the SNORM option with METHOD=SPACING.
P*n*	percentiles of each variable, as specified by the PCTLPTS= option. The argument *n* is any real number such that $0 \le n \le 100$.

Displayed Output

If you specify the PSTAT option, PROC STDIZE displays the following statistics for each variable:

- the name of the variable, Name
- the location estimate, Location
- the scale estimate, Scale
- the norm estimate, Norm (when you specify the NORM option with METHOD=AGK, METHOD=IQR, METHOD=MAD, or METHOD=SPACING or when you specify the SNORM option with METHOD=SPACING)
- the total nonmissing positive frequencies, N

ODS Table Names

PROC STDIZE assigns a name to the single table it creates. You can use this name to reference the table when using the Output Delivery System (ODS) to select output or create an output data set. For more information on ODS, see Chapter 15, “Using the Output Delivery System.”

Table 59.3. ODS Table Produced in PROC STDIZE

ODS Table Name	Description	Option
Statistics	Location and Scale Measures	PSTAT

Examples

Example 59.1. Standardization of Variables in Cluster Analysis

To illustrate the effect of standardization in cluster analysis, this example uses the Fish data set described in the “Getting Started” section of Chapter 27, “The FASTCLUS Procedure.” The numbers are measurements taken on 159 fish caught off the coast of Finland; this data set is available from the Data Archive of the *Journal of Statistics Education*. The complete data set is displayed in Chapter 60, “The STEPDISC Procedure.”

The species (Bream, Parkki, Pike, Perch, Roach, Smelt, and Whitefish), weight, three different length measurements (measured from the nose of the fish to the beginning of its tail, the notch of its tail, and the end of its tail), height, and width of each fish are recorded. The height and width are recorded as percentages of the third length variable.

Several new variables are created in the Fish data set: Weight3, Height, Width, and logLengthRatio. The weight of a fish indicates its size—a heavier Tuna tends to be larger than a lighter Tuna. To get a one dimensional measure of the size of a fish, take the cubic root of the weight (Weight3). The variables Height, Width, Length1, Length2, and Length3 are rescaled in order to adjust for dimensionality. The logLengthRatio variable measures the tail length.

Because the new variables Weight3–logLengthRatio depend on the variable Weight, observations with missing values for Weight are not added to the data set. Consequently, there are 157 observations in the SAS data set Fish.

Before you perform a cluster analysis on coordinate data, it is necessary to consider scaling or transforming the variables since variables with large variances tend to have a larger effect on the resulting clusters than those with small variances.

This example uses three different approaches to standardize or transform the data prior to the cluster analysis. The first approach uses several standardization methods provided in the STDIZE procedure. However, since standardization is not always appropriate prior to the clustering (refer to Milligan and Cooper, 1987, for a Monte Carlo study on various methods of variable standardization), the second approach performs the cluster analysis with no standardization. The third approach invokes the ACECLUS procedure to transform the data into a within-cluster covariance matrix.

The clustering is performed by the FASTCLUS procedure to find seven clusters. Note that the variables Length2 and Length3 are eliminated from this analysis since they both are significantly and highly correlated with the variable Length1. The correlation coefficients are 0.9958 and 0.9604, respectively. An output data set is created, and the FREQ procedure is invoked to compare the clusters with the species classification.

The DATA step is as follows:

```
proc format;
   value specfmt
      1='Bream'
      2='Roach'
      3='Whitefish'
      4='Parkki'
      5='Perch'
      6='Pike'
      7='Smelt';
data Fish (drop=HtPct WidthPct);
   title 'Fish Measurement Data';
   input Species Weight Length1 Length2 Length3 HtPct
         WidthPct @@;

   if Weight <= 0 or Weight=. then delete;
   Weight3=Weight**(1/3);
   Height=HtPct*Length3/(Weight3*100);
   Width=WidthPct*Length3/(Weight3*100);
   Length1=Length1/Weight3;
   Length2=Length2/Weight3;
   Length3=Length3/Weight3;
   logLengthRatio=log(Length3/Length1);

   format Species specfmt.;
   symbol = put(Species, specfmt2.);
   datalines;
1  242.0 23.2 25.4 30.0 38.4 13.4
1  290.0 24.0 26.3 31.2 40.0 13.8
1  340.0 23.9 26.5 31.1 39.8 15.1
1  363.0 26.3 29.0 33.5 38.0 13.3
 ... [155 more records]
;
run;
```

The following macro, Std, standardizes the Fish data. The macro reads a single argument, mtd, which selects the METHOD= specification to be used in PROC STDIZE.

```
/*--- macro for standardization ---*/

%macro Std(mtd);
title2 "Data is standardized by PROC STDIZE with
        METHOD= &mtd";
   proc stdize data=fish out=sdzout method=&mtd;
      var Length1 logLengthRatio Height Width Weight3;
   run;
%mend Std;
```

The following macro, FastFreq, includes a PROC FASTCLUS statement for performing cluster analysis and a PROC FREQ statement for cross-tabulating species with the cluster membership information that is derived from the previous PROC FASTCLUS statement. The macro reads a single argument, ds, which selects the input data set to be used in PROC FASTCLUS.

```
/*--- macro for clustering and cross-tabulating ---*/
/*--- cluster membership with species           ---*/
%macro FastFreq(ds);
   proc fastclus data=&ds out=clust maxclusters=7 maxiter=100 noprint;
      var Length1 logLengthRatio Height Width Weight3;
   run;

   proc freq data=clust;
      tables species*cluster;
   run;
%mend FastFreq;
```

The following analysis, (labeled ‘Approach 1’) includes 18 different methods of standardization followed by clustering. Since there is a large amount of output from this approach, only results from METHOD=STD, METHOD=RANGE, METHOD=AGK(.14), and METHOD=SPACING(.14) are shown. The following statements produce Output 59.1.1 through Output 59.1.4.

```
/************************************************************/
/*                                                          */
/*     Approach 1: data is standardized by PROC STDIZE      */
/*                                                          */
/************************************************************/

%Std(MEAN);
%FastFreq(sdzout);

%Std(MEDIAN);
%FastFreq(sdzout);

%Std(SUM);
%FastFreq(sdzout);

%Std(EUCLEN);
%FastFreq(sdzout);
```

```
%Std(USTD);
%FastFreq(sdzout);

%Std(STD);
%FastFreq(sdzout);

%Std(RANGE);
%FastFreq(sdzout);

%Std(MIDRANGE);
%FastFreq(sdzout);

%Std(MAXABS);
%FastFreq(sdzout);

%Std(IQR);
%FastFreq(sdzout);

%Std(MAD);
%FastFreq(sdzout);

%Std(AGK(.14));
%FastFreq(sdzout);

%Std(SPACING(.14));
%FastFreq(sdzout);

%Std(ABW(5));
%FastFreq(sdzout);

%Std(AWAVE(5));
%FastFreq(sdzout);

%Std(L(1));
%FastFreq(sdzout);

%Std(L(1.5));
%FastFreq(sdzout);

%Std(L(2));
%FastFreq(sdzout);
```

Output 59.1.1. Data is standardized by PROC STDIZE with METHOD=STD

```
                              Fish Measurement Data
                  Data is standardized by PROC STDIZE with METHOD= STD

                                The FREQ Procedure

                             Table of Species by CLUSTER

          Species     CLUSTER(Cluster)

          Frequency |
          Percent   |
          Row Pct   |
          Col Pct   |       1|       2|       3|       4|       5|       6|       7|  Total
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Bream     |      0 |      0 |      0 |      0 |      0 |     34 |      0 |     34
                    |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |  21.66 |   0.00 |  21.66
                    |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 | 100.00 |   0.00 |
                    |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 | 100.00 |   0.00 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Roach     |      0 |      0 |      0 |      0 |      0 |      0 |     19 |     19
                    |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |  12.10 |  12.10
                    |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 | 100.00 |
                    |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |  38.00 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Whitefish |      0 |      2 |      0 |      1 |      0 |      0 |      3 |      6
                    |   0.00 |   1.27 |   0.00 |   0.64 |   0.00 |   0.00 |   1.91 |   3.82
                    |   0.00 |  33.33 |   0.00 |  16.67 |   0.00 |   0.00 |  50.00 |
                    |   0.00 |  10.53 |   0.00 |   7.69 |   0.00 |   0.00 |   6.00 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Parkki    |      0 |      0 |      0 |      0 |     11 |      0 |      0 |     11
                    |   0.00 |   0.00 |   0.00 |   0.00 |   7.01 |   0.00 |   0.00 |   7.01
                    |   0.00 |   0.00 |   0.00 |   0.00 | 100.00 |   0.00 |   0.00 |
                    |   0.00 |   0.00 |   0.00 |   0.00 | 100.00 |   0.00 |   0.00 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Perch     |      0 |     17 |      0 |     12 |      0 |      0 |     27 |     56
                    |   0.00 |  10.83 |   0.00 |   7.64 |   0.00 |   0.00 |  17.20 |  35.67
                    |   0.00 |  30.36 |   0.00 |  21.43 |   0.00 |   0.00 |  48.21 |
                    |   0.00 |  89.47 |   0.00 |  92.31 |   0.00 |   0.00 |  54.00 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Pike      |     17 |      0 |      0 |      0 |      0 |      0 |      0 |     17
                    |  10.83 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |  10.83
                    | 100.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
                    | 100.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Smelt     |      0 |      0 |     13 |      0 |      0 |      0 |      1 |     14
                    |   0.00 |   0.00 |   8.28 |   0.00 |   0.00 |   0.00 |   0.64 |   8.92
                    |   0.00 |   0.00 |  92.86 |   0.00 |   0.00 |   0.00 |   7.14 |
                    |   0.00 |   0.00 | 100.00 |   0.00 |   0.00 |   0.00 |   2.00 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Total           17       19       13       13       11       34       50      157
                       10.83    12.10     8.28     8.28     7.01    21.66    31.85   100.00
```

Output 59.1.2. Data is standardized by PROC STDIZE with METHOD=RANGE

```
                         Fish Measurement Data
            Data is standardized by PROC STDIZE with METHOD= RANGE

                           The FREQ Procedure

                       Table of Species by CLUSTER

      Species     CLUSTER(Cluster)

      Frequency |
      Percent   |
      Row Pct   |
      Col Pct   |       1|       2|       3|       4|       5|       6|       7|  Total
      ----------+--------+--------+--------+--------+--------+--------+--------+
      Bream     |      0 |      0 |     34 |      0 |      0 |      0 |      0 |     34
                |   0.00 |   0.00 |  21.66 |   0.00 |   0.00 |   0.00 |   0.00 |  21.66
                |   0.00 |   0.00 | 100.00 |   0.00 |   0.00 |   0.00 |   0.00 |
                |   0.00 |   0.00 | 100.00 |   0.00 |   0.00 |   0.00 |   0.00 |
      ----------+--------+--------+--------+--------+--------+--------+--------+
      Roach     |      0 |      0 |      0 |     19 |      0 |      0 |      0 |     19
                |   0.00 |   0.00 |   0.00 |  12.10 |   0.00 |   0.00 |   0.00 |  12.10
                |   0.00 |   0.00 |   0.00 | 100.00 |   0.00 |   0.00 |   0.00 |
                |   0.00 |   0.00 |   0.00 |  61.29 |   0.00 |   0.00 |   0.00 |
      ----------+--------+--------+--------+--------+--------+--------+--------+
      Whitefish |      0 |      0 |      0 |      3 |      3 |      0 |      0 |      6
                |   0.00 |   0.00 |   0.00 |   1.91 |   1.91 |   0.00 |   0.00 |   3.82
                |   0.00 |   0.00 |   0.00 |  50.00 |  50.00 |   0.00 |   0.00 |
                |   0.00 |   0.00 |   0.00 |   9.68 |  13.04 |   0.00 |   0.00 |
      ----------+--------+--------+--------+--------+--------+--------+--------+
      Parkki    |      0 |      0 |      0 |      0 |      0 |     11 |      0 |     11
                |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   7.01 |   0.00 |   7.01
                |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 | 100.00 |   0.00 |
                |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 | 100.00 |   0.00 |
      ----------+--------+--------+--------+--------+--------+--------+--------+
      Perch     |      0 |      0 |      0 |      9 |     20 |      0 |     27 |     56
                |   0.00 |   0.00 |   0.00 |   5.73 |  12.74 |   0.00 |  17.20 |  35.67
                |   0.00 |   0.00 |   0.00 |  16.07 |  35.71 |   0.00 |  48.21 |
                |   0.00 |   0.00 |   0.00 |  29.03 |  86.96 |   0.00 | 100.00 |
      ----------+--------+--------+--------+--------+--------+--------+--------+
      Pike      |     17 |      0 |      0 |      0 |      0 |      0 |      0 |     17
                |  10.83 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |  10.83
                | 100.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
                | 100.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
      ----------+--------+--------+--------+--------+--------+--------+--------+
      Smelt     |      0 |     14 |      0 |      0 |      0 |      0 |      0 |     14
                |   0.00 |   8.92 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   8.92
                |   0.00 | 100.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
                |   0.00 | 100.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
      ----------+--------+--------+--------+--------+--------+--------+--------+
      Total           17       14       34       31       23       11       27      157
                   10.83     8.92    21.66    19.75    14.65     7.01    17.20   100.00
```

Output 59.1.3. Data is standardized by PROC STDIZE with METHOD=AGK(.14)

```
                         Fish Measurement Data
            Data is standardized by PROC STDIZE with METHOD= AGK(.14)

                           The FREQ Procedure

                       Table of Species by CLUSTER

         Species     CLUSTER(Cluster)

         Frequency |
         Percent   |
         Row Pct   |
         Col Pct   |       1|       2|       3|       4|       5|       6|       7|  Total
         ----------+--------+--------+--------+--------+--------+--------+--------+
         Bream     |      0 |      0 |     34 |      0 |      0 |      0 |      0 |     34
                   |   0.00 |   0.00 |  21.66 |   0.00 |   0.00 |   0.00 |   0.00 |  21.66
                   |   0.00 |   0.00 | 100.00 |   0.00 |   0.00 |   0.00 |   0.00 |
                   |   0.00 |   0.00 | 100.00 |   0.00 |   0.00 |   0.00 |   0.00 |
         ----------+--------+--------+--------+--------+--------+--------+--------+
         Roach     |      0 |      0 |      0 |     17 |      0 |      0 |      2 |     19
                   |   0.00 |   0.00 |   0.00 |  10.83 |   0.00 |   0.00 |   1.27 |  12.10
                   |   0.00 |   0.00 |   0.00 |  89.47 |   0.00 |   0.00 |  10.53 |
                   |   0.00 |   0.00 |   0.00 |  73.91 |   0.00 |   0.00 |   5.71 |
         ----------+--------+--------+--------+--------+--------+--------+--------+
         Whitefish |      0 |      0 |      0 |      3 |      0 |      3 |      0 |      6
                   |   0.00 |   0.00 |   0.00 |   1.91 |   0.00 |   1.91 |   0.00 |   3.82
                   |   0.00 |   0.00 |   0.00 |  50.00 |   0.00 |  50.00 |   0.00 |
                   |   0.00 |   0.00 |   0.00 |  13.04 |   0.00 |  13.04 |   0.00 |
         ----------+--------+--------+--------+--------+--------+--------+--------+
         Parkki    |     11 |      0 |      0 |      0 |      0 |      0 |      0 |     11
                   |   7.01 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   7.01
                   | 100.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
                   | 100.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
         ----------+--------+--------+--------+--------+--------+--------+--------+
         Perch     |      0 |      0 |      0 |      3 |      0 |     20 |     33 |     56
                   |   0.00 |   0.00 |   0.00 |   1.91 |   0.00 |  12.74 |  21.02 |  35.67
                   |   0.00 |   0.00 |   0.00 |   5.36 |   0.00 |  35.71 |  58.93 |
                   |   0.00 |   0.00 |   0.00 |  13.04 |   0.00 |  86.96 |  94.29 |
         ----------+--------+--------+--------+--------+--------+--------+--------+
         Pike      |      0 |      0 |      0 |      0 |     17 |      0 |      0 |     17
                   |   0.00 |   0.00 |   0.00 |   0.00 |  10.83 |   0.00 |   0.00 |  10.83
                   |   0.00 |   0.00 |   0.00 |   0.00 | 100.00 |   0.00 |   0.00 |
                   |   0.00 |   0.00 |   0.00 |   0.00 | 100.00 |   0.00 |   0.00 |
         ----------+--------+--------+--------+--------+--------+--------+--------+
         Smelt     |      0 |     14 |      0 |      0 |      0 |      0 |      0 |     14
                   |   0.00 |   8.92 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   8.92
                   |   0.00 | 100.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
                   |   0.00 | 100.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
         ----------+--------+--------+--------+--------+--------+--------+--------+
         Total           11       14       34       23       17       23       35      157
                       7.01     8.92    21.66    14.65    10.83    14.65    22.29   100.00
```

Output 59.1.4. Data is standardized by PROC STDIZE with METHOD=SPACING(.14)

```
                              Fish Measurement Data
                Data is standardized by PROC STDIZE with METHOD= SPACING(.14)

                                 The FREQ Procedure

                             Table of Species by CLUSTER

          Species     CLUSTER(Cluster)

          Frequency |
          Percent   |
          Row Pct   |
          Col Pct   |       1|       2|       3|       4|       5|       6|       7|  Total
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Bream     |      0 |      0 |      0 |      0 |      0 |      0 |     34 |     34
                    |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |  21.66 |  21.66
                    |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 | 100.00 |
                    |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 | 100.00 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Roach     |      0 |      0 |      0 |     17 |      0 |      2 |      0 |     19
                    |   0.00 |   0.00 |   0.00 |  10.83 |   0.00 |   1.27 |   0.00 |  12.10
                    |   0.00 |   0.00 |   0.00 |  89.47 |   0.00 |  10.53 |   0.00 |
                    |   0.00 |   0.00 |   0.00 |  85.00 |   0.00 |   5.26 |   0.00 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Whitefish |      3 |      0 |      0 |      3 |      0 |      0 |      0 |      6
                    |   1.91 |   0.00 |   0.00 |   1.91 |   0.00 |   0.00 |   0.00 |   3.82
                    |  50.00 |   0.00 |   0.00 |  50.00 |   0.00 |   0.00 |   0.00 |
                    |  13.04 |   0.00 |   0.00 |  15.00 |   0.00 |   0.00 |   0.00 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Parkki    |      0 |      0 |     11 |      0 |      0 |      0 |      0 |     11
                    |   0.00 |   0.00 |   7.01 |   0.00 |   0.00 |   0.00 |   0.00 |   7.01
                    |   0.00 |   0.00 | 100.00 |   0.00 |   0.00 |   0.00 |   0.00 |
                    |   0.00 |   0.00 | 100.00 |   0.00 |   0.00 |   0.00 |   0.00 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Perch     |     20 |      0 |      0 |      0 |      0 |     36 |      0 |     56
                    |  12.74 |   0.00 |   0.00 |   0.00 |   0.00 |  22.93 |   0.00 |  35.67
                    |  35.71 |   0.00 |   0.00 |   0.00 |   0.00 |  64.29 |   0.00 |
                    |  86.96 |   0.00 |   0.00 |   0.00 |   0.00 |  94.74 |   0.00 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Pike      |      0 |     17 |      0 |      0 |      0 |      0 |      0 |     17
                    |   0.00 |  10.83 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |  10.83
                    |   0.00 | 100.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
                    |   0.00 | 100.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Smelt     |      0 |      0 |      0 |      0 |     14 |      0 |      0 |     14
                    |   0.00 |   0.00 |   0.00 |   0.00 |   8.92 |   0.00 |   0.00 |   8.92
                    |   0.00 |   0.00 |   0.00 |   0.00 | 100.00 |   0.00 |   0.00 |
                    |   0.00 |   0.00 |   0.00 |   0.00 | 100.00 |   0.00 |   0.00 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Total           23       17       11       20       14       38       34      157
                       14.65    10.83     7.01    12.74     8.92    24.20    21.66   100.00
```

The following analysis (labeled ‘Approach 2’) applies the cluster analysis directly to the original data. The following statements produce Output 59.1.5.

```
/*************************************************************/
/*                                                           */
/*         Approach 2: data is untransformed                 */
/*                                                           */
/*************************************************************/

title2 'Data is untransformed';
%FastFreq(fish);
```

Output 59.1.5. Untransformed Data

```
                              Fish Measurement Data
                              Data is untransformed

                               The FREQ Procedure

                          Table of Species by CLUSTER

          Species     CLUSTER(Cluster)

          Frequency |
          Percent   |
          Row Pct   |
          Col Pct   |       1|       2|       3|       4|       5|       6|       7|  Total
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Bream     |     13 |      0 |      0 |      0 |      0 |      0 |     21 |     34
                    |   8.28 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |  13.38 |  21.66
                    |  38.24 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |  61.76 |
                    |  44.83 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |  47.73 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Roach     |      3 |      4 |      0 |      0 |     12 |      0 |      0 |     19
                    |   1.91 |   2.55 |   0.00 |   0.00 |   7.64 |   0.00 |   0.00 |  12.10
                    |  15.79 |  21.05 |   0.00 |   0.00 |  63.16 |   0.00 |   0.00 |
                    |  10.34 |  25.00 |   0.00 |   0.00 |  30.77 |   0.00 |   0.00 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Whitefish |      3 |      0 |      0 |      0 |      0 |      0 |      3 |      6
                    |   1.91 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   1.91 |   3.82
                    |  50.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |  50.00 |
                    |  10.34 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   6.82 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Parkki    |      2 |      3 |      0 |      0 |      6 |      0 |      0 |     11
                    |   1.27 |   1.91 |   0.00 |   0.00 |   3.82 |   0.00 |   0.00 |   7.01
                    |  18.18 |  27.27 |   0.00 |   0.00 |  54.55 |   0.00 |   0.00 |
                    |   6.90 |  18.75 |   0.00 |   0.00 |  15.38 |   0.00 |   0.00 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Perch     |      8 |      9 |      0 |      1 |     20 |      0 |     18 |     56
                    |   5.10 |   5.73 |   0.00 |   0.64 |  12.74 |   0.00 |  11.46 |  35.67
                    |  14.29 |  16.07 |   0.00 |   1.79 |  35.71 |   0.00 |  32.14 |
                    |  27.59 |  56.25 |   0.00 |   6.67 |  51.28 |   0.00 |  40.91 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Pike      |      0 |      0 |     10 |      0 |      1 |      4 |      2 |     17
                    |   0.00 |   0.00 |   6.37 |   0.00 |   0.64 |   2.55 |   1.27 |  10.83
                    |   0.00 |   0.00 |  58.82 |   0.00 |   5.88 |  23.53 |  11.76 |
                    |   0.00 |   0.00 | 100.00 |   0.00 |   2.56 | 100.00 |   4.55 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Smelt     |      0 |      0 |      0 |     14 |      0 |      0 |      0 |     14
                    |   0.00 |   0.00 |   0.00 |   8.92 |   0.00 |   0.00 |   0.00 |   8.92
                    |   0.00 |   0.00 |   0.00 | 100.00 |   0.00 |   0.00 |   0.00 |
                    |   0.00 |   0.00 |   0.00 |  93.33 |   0.00 |   0.00 |   0.00 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Total           29       16       10       15       39        4       44      157
                       18.47    10.19     6.37     9.55    24.84     2.55    28.03   100.00
```

The following analysis (labeled ‘Approach 3’) transforms the original data with the ACECLUS procedure and creates a TYPE=ACE output data set that is used as an input data set for the cluster analysis. The following statements produce Output 59.1.6.

```
/************************************************************/
/*                                                          */
/*    Approach 3: data is transformed by PROC ACECLUS       */
/*                                                          */
/************************************************************/

title2 ’Data is transformed by PROC ACECLUS’;
proc aceclus data=fish out=ace p=.02 noprint;
   var Length1 logLengthRatio Height Width Weight3;
run;
%FastFreq(ace);
```

Output 59.1.6. Data is transformed by PROC ACECLUS

```
                           Fish Measurement Data
                    Data is transformed by PROC ACECLUS

                            The FREQ Procedure

                        Table of Species by CLUSTER

          Species     CLUSTER(Cluster)

          Frequency |
          Percent   |
          Row Pct   |
          Col Pct   |       1|       2|       3|       4|       5|       6|       7|  Total
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Bream     |     13 |      0 |      0 |      0 |      0 |      0 |     21 |     34
                    |   8.28 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |  13.38 |  21.66
                    |  38.24 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |  61.76 |
                    |  44.83 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |  47.73 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Roach     |      3 |      4 |      0 |      0 |     12 |      0 |      0 |     19
                    |   1.91 |   2.55 |   0.00 |   0.00 |   7.64 |   0.00 |   0.00 |  12.10
                    |  15.79 |  21.05 |   0.00 |   0.00 |  63.16 |   0.00 |   0.00 |
                    |  10.34 |  25.00 |   0.00 |   0.00 |  30.77 |   0.00 |   0.00 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Whitefish |      3 |      0 |      0 |      0 |      0 |      0 |      3 |      6
                    |   1.91 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   1.91 |   3.82
                    |  50.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |  50.00 |
                    |  10.34 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   6.82 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Parkki    |      2 |      3 |      0 |      0 |      6 |      0 |      0 |     11
                    |   1.27 |   1.91 |   0.00 |   0.00 |   3.82 |   0.00 |   0.00 |   7.01
                    |  18.18 |  27.27 |   0.00 |   0.00 |  54.55 |   0.00 |   0.00 |
                    |   6.90 |  18.75 |   0.00 |   0.00 |  15.38 |   0.00 |   0.00 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Perch     |      8 |      9 |      0 |      1 |     20 |      0 |     18 |     56
                    |   5.10 |   5.73 |   0.00 |   0.64 |  12.74 |   0.00 |  11.46 |  35.67
                    |  14.29 |  16.07 |   0.00 |   1.79 |  35.71 |   0.00 |  32.14 |
                    |  27.59 |  56.25 |   0.00 |   6.67 |  51.28 |   0.00 |  40.91 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Pike      |      0 |      0 |     10 |      0 |      1 |      4 |      2 |     17
                    |   0.00 |   0.00 |   6.37 |   0.00 |   0.64 |   2.55 |   1.27 |  10.83
                    |   0.00 |   0.00 |  58.82 |   0.00 |   5.88 |  23.53 |  11.76 |
                    |   0.00 |   0.00 | 100.00 |   0.00 |   2.56 | 100.00 |   4.55 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Smelt     |      0 |      0 |      0 |     14 |      0 |      0 |      0 |     14
                    |   0.00 |   0.00 |   0.00 |   8.92 |   0.00 |   0.00 |   0.00 |   8.92
                    |   0.00 |   0.00 |   0.00 | 100.00 |   0.00 |   0.00 |   0.00 |
                    |   0.00 |   0.00 |   0.00 |  93.33 |   0.00 |   0.00 |   0.00 |
          ----------+--------+--------+--------+--------+--------+--------+--------+
          Total           29       16       10       15       39        4       44      157
                       18.47    10.19     6.37     9.55    24.84     2.55    28.03   100.00
```

Table 59.4 displays a table summarizing each classification results. In this table, the first column represents the standardization method, the second column represents the number of clusters that the 7 species are classified into, and the third column represents the total number of observations that are misclassified.

Table 59.4. Summary of Clustering Results

Method of Standardization	Number of Clusters	Misclassification
MEAN	5	71
MEDIAN	5	71
SUM	6	51
EUCLEN	6	45
USTD	6	45
STD	5	33
RANGE	7	32
MIDRANGE	7	32
MAXABS	7	26
IQR	5	28
MAD	4	35
ABW(5)	6	34
AWAVE(5)	6	29
AGK(.14)	7	28
SPACING(.14)	7	25
L(1)	6	41
L(1.5)	5	33
L(2)	5	33
untransformed	5	71
PROC ACECLUS	5	71

Consider the results displayed in Output 59.1.1. In that analysis, the method of standardization is STD, and the number of clusters and the number of misclassifications are computed as shown in Table 59.5.

Table 59.5. Computations of numbers of clusters and misclassification when standardization method is STD

Species	Cluster Number	Misclassification in Each Species
Bream	6	0
Roach	7	0
Whitefish	7	3
Parkki	5	0
Perch	7	29
Pike	1	0
Smelt	3	1

In Output 59.1.1, the Bream species is classified as cluster 6 since all 34 Bream fish are categorized into cluster 6 with no misclassification. A similar pattern is seen with the Roach, Parkki, Pike, and Smelt species.

For the Whitefish species, two fish are categorized into cluster 2, one fish is categorized into cluster 4, and three fish are categorized into cluster 7. Because the majority of this species is categorized into cluster 7, it is recorded in Table 59.5 as being classified as cluster 7 with 3 misclassifications. A similar pattern is seen with the Perch species: it is classified as cluster 7 with 29 misclassifications.

In summary, when the standardization method is STD, seven species of fish are classified into only 5 clusters and the total number of misclassified observations is 33.

The result of this analysis demonstrates that when variables are standardized by the STDIZE procedure with methods including RANGE, MIDRANGE, MAXABS, AGK(.14), and SPACING(.14), the FASTCLUS procedure produces the correct number of clusters and less misclassification than it does when other standardization methods are used. The SPACING method attains the best result, probably because the variables Length1 and Height both exhibit marked groupings (bimodality) in their distributions.

References

Art, D., Gnanadesikan, R., and Kettenring, R. (1982), "Data-based Metrics for Cluster Analysis," *Utilitas Mathematica*, 21A, 75–99.

Goodall, C. (1983), "M-Estimators of Location: An Outline of Theory," in Hoaglin, D.C., Mosteller, M. and Tukey, J.W., eds., *Understanding Robust and Exploratory Data Analysis*, New York: John Wiley & Sons, Inc.

Iglewicz, B. (1983), "Robust Scale Estimators and Confidence Intervals for Location," in Hoaglin, D.C., Mosteller, M. and Tukey, J.W., eds., *Understanding Robust and Exploratory Data Analysis*, New York: John Wiley & Sons, Inc.

Jannsen, P., Marron, J.S., Veraverbeke, N, and Sarle, W.S. (1995), "Scale Measures for Bandwidth Selection," *J. of Nonparametric Statistics*, 5(4), 359–380.

Jain R. and Chlamtac I. (1985), "The P^2 Algorithm for Dynamic Calculation of Quantiles and Histograms Without Storing Observations," *Communications of the ACM*, 28(10), 1076-1085.

Journal of Statistics Education, "Fish Catch Data Set," [http://www.stat.ncsu.edu/info/jse], accessed 4 December 1997.

Milligan, G.W. and Cooper, M.C. (1987), "A Study of Variable Standardization," *College of Administrative Science Working Paper Series*, 87–63, Columbus, OH: The Ohio State University.

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

SAS/STAT® User's Guide, Version 8

ISBN 1–58025–494–2

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

1st printing, October 1999

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