SAS/STAT® 14.2 User's Guide
The BOXPLOT Procedure

This document is an individual chapter from *SAS/STAT® 14.2 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2016. *SAS/STAT® 14.2 User's Guide*. Cary, NC: SAS Institute Inc.

SAS/STAT® 14.2 User's Guide

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

November 2016

Chapter 28
The BOXPLOT Procedure

Contents

Overview: BOXPLOT Procedure

The BOXPLOT procedure creates side-by-side box-and-whiskers plots of measurements organized in groups. A box-and-whiskers plot displays the mean, quartiles, and minimum and maximum observations for a group. Throughout this chapter, this type of plot, which can contain one or more box-and-whiskers plots, is referred to as a *box plot*.

The PLOT statement of the BOXPLOT procedure produces a box plot. You can specify more than one PLOT statement to produce multiple box plots. You can use options in the PLOT statement to do the following:

- control the style of the box-and-whiskers plots
- specify one of several methods for calculating quantile statistics (percentiles)
- add block legends and symbol markers to reveal stratification in data
- display vertical and horizontal reference lines
- control axis values and labels
- overlay the box plot with plots of additional variables
- control the layout and appearance of the plot

The INSET and INSETGROUP statements produce boxes or tables (referred to as *insets*) of summary statistics or other data on a box plot. An INSET statement produces an inset of statistics pertaining to the entire box plot. An INSETGROUP statement produces an inset containing statistics calculated separately for each group. An INSET or INSETGROUP statement by itself does not produce a display; it must be used with a PLOT statement.

You can use options in an INSET or INSETGROUP statement to control insets in these ways:

- specify the position of the inset
- specify a header for the inset
- specify graphical enhancements, such as background colors, text colors, text height, text font, and drop shadows

Traditional Graphics and ODS Graphics

The BOXPLOT procedure can produce two kinds of graphical output:

- traditional graphics
- ODS Statistical Graphics output

Traditional graphics are saved in graphics catalogs with entry type GRSEG. Their appearance is controlled by global statements such as the GOPTIONS, AXIS, and SYMBOL statements (as described in *SAS/GRAPH: Reference*) and numerous specialized PLOT statement options. You must have a SAS/GRAPH® license to produce traditional graphics.

ODS Statistical Graphics (or ODS Graphics for short) is an extension to the Output Delivery System (ODS). Graphs are produced in standard image file formats (such as PNG) instead of graphics catalogs, and the details of their appearance and layout are controlled by ODS styles and templates. When ODS Graphics is enabled (for example, with the ODS GRAPHICS ON statement) PROC BOXPLOT produces ODS Graphics output. Otherwise, it produces traditional graphics. See Chapter 21, "Statistical Graphics Using ODS," for a thorough discussion of ODS Graphics.

Global graphics statements (GOPTIONS, AXIS, and SYMBOL, for example) and PLOT statement options that specify details of graph appearance (such as CBOXFILL= and FONT=) are ignored when ODS Graphics is enabled. Some PLOT statement options do affect ODS Graphics output, as indicated in the section "PLOT Statement Options" on page 1139.

See the section "Getting Started: BOXPLOT Procedure" on page 1123 for examples producing box plots via the traditional graphics system and ODS Graphics.

NOTE: Prior to SAS 9.2, traditional graphics produced by PROC BOXPLOT were extremely basic by default. Producing attractive graphical output required the careful selection of colors, fonts, and other elements, which were specified via SAS/GRAPH statements and PLOT statement options. Beginning with SAS 9.2, the default appearance of traditional box plots is governed by the prevailing ODS style, which automatically produces attractive, consistent output. You can specify the NOGSTYLE system option to prevent the ODS style from affecting the appearance of traditional graphs.

Getting Started: BOXPLOT Procedure

This section introduces the BOXPLOT procedure with simple examples demonstrating commonly used options. Complete syntax for the BOXPLOT procedure is presented in the section "Syntax: BOXPLOT Procedure" on page 1131, and advanced examples are presented in the section "Examples: BOXPLOT Procedure" on page 1183.

Creating Box Plots from Raw Data

A petroleum company uses a turbine to heat water into steam that is pumped into the ground to make oil less viscous and easier to extract. This process occurs 20 times daily, and the amount of power (in kilowatts) used to heat the water to the desired temperature is recorded. The following statements create a SAS data set called Turbine that contains the power output measurements for 10 nonconsecutive days:

```
data Turbine;
   informat Day date7.;
   format Day date5.;
   label KWatts='Average Power Output';
   input Day @;
   do i=1 to 10;
```

```
      input KWatts @;
      output;
   end;
   drop i;
   datalines;
05JUL94 3196 3507 4050 3215 3583 3617 3789 3180 3505 3454
05JUL94 3417 3199 3613 3384 3475 3316 3556 3607 3364 3721
06JUL94 3390 3562 3413 3193 3635 3179 3348 3199 3413 3562
06JUL94 3428 3320 3745 3426 3849 3256 3841 3575 3752 3347
07JUL94 3478 3465 3445 3383 3684 3304 3398 3578 3348 3369
07JUL94 3670 3614 3307 3595 3448 3304 3385 3499 3781 3711
08JUL94 3448 3045 3446 3620 3466 3533 3590 3070 3499 3457
08JUL94 3411 3350 3417 3629 3400 3381 3309 3608 3438 3567
11JUL94 3568 2968 3514 3465 3175 3358 3460 3851 3845 2983
11JUL94 3410 3274 3590 3527 3509 3284 3457 3729 3916 3633
12JUL94 3153 3408 3741 3203 3047 3580 3571 3579 3602 3335
12JUL94 3494 3662 3586 3628 3881 3443 3456 3593 3827 3573
13JUL94 3594 3711 3369 3341 3611 3496 3554 3400 3295 3002
13JUL94 3495 3368 3726 3738 3250 3632 3415 3591 3787 3478
14JUL94 3482 3546 3196 3379 3559 3235 3549 3445 3413 3859
14JUL94 3330 3465 3994 3362 3309 3781 3211 3550 3637 3626
15JUL94 3152 3269 3431 3438 3575 3476 3115 3146 3731 3171
15JUL94 3206 3140 3562 3592 3722 3421 3471 3621 3361 3370
18JUL94 3421 3381 4040 3467 3475 3285 3619 3325 3317 3472
18JUL94 3296 3501 3366 3492 3367 3619 3550 3263 3355 3510
;
```

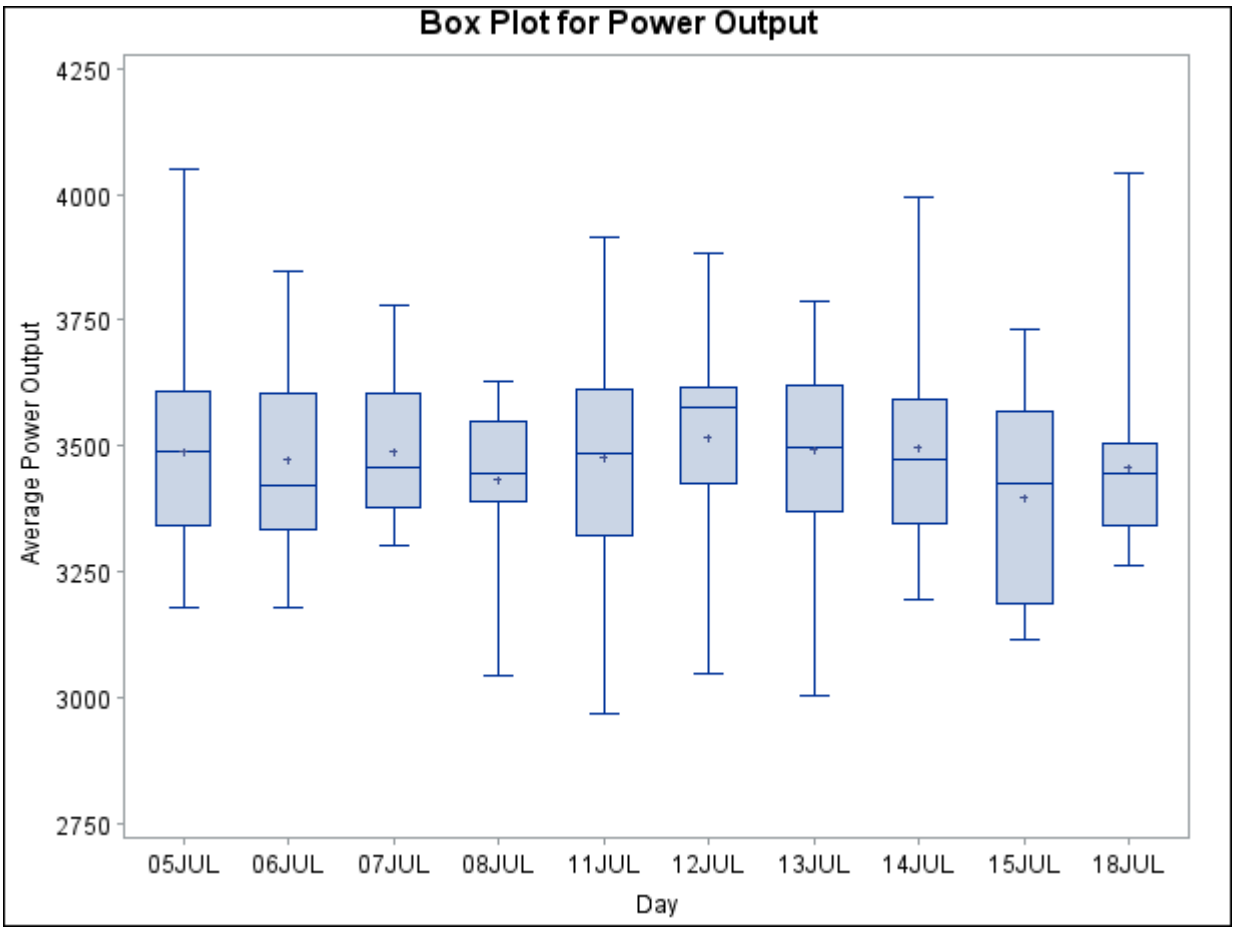
In the data set Turbine, each observation contains the date and the power output for a single heating. The first 20 observations contain the outputs for the first day, the second 20 observations contain the outputs for the second day, and so on. Because the variable Day classifies the observations into groups, it is referred to as the *group variable*. The variable KWatts contains the output measurements and is referred to as the *analysis variable*.

The following statements create a box plot showing the distribution of power output for each day:

```
ods graphics off;
title 'Box Plot for Power Output';
proc boxplot data=Turbine;
   plot KWatts*Day;
run;
```

The input data set Turbine is specified with the DATA= option in the PROC BOXPLOT statement. The PLOT statement requests a box-and-whiskers plot for each group of data. After the keyword PLOT, you specify the analysis variable (in this case, KWatts), followed by an asterisk and the group variable (Day). The ODS GRAPHICS OFF statement specified before the PROC BOXPLOT statement disables ODS Graphics, so the box plot is produced using traditional graphics. The box plot is shown in Figure 28.1.

Figure 28.1 Box Plot for Power Output Data



The box plot displayed in Figure 28.1 represents summary statistics for the analysis variable KWatts. Each of the 10 box-and-whiskers plots describes the variable KWatts for a particular day. The plot elements and the statistics they represent are as follows:

The length of the box represents the interquartile range (the distance between the 25th and 75th percentiles).

The symbol in the box interior represents the group mean.

The horizontal line in the box interior represents the group median.

The vertical lines (called *whiskers*) issuing from the box extend to the group minimum and maximum values.

Creating Box Plots from Summary Data

The previous example illustrates how you can create box plots from raw data. However, in some applications the data are provided as summary statistics. This example illustrates how you can use the BOXPLOT procedure with data of this type.

The following statements create the data set Oilsum, which provides the data from the preceding example in summarized form:

```
data Oilsum;
   input Day KWattsL KWatts1 KWattsX KWattsM
             KWatts3 KWattsH KWattsS KWattsN;
   informat Day date7. ;
   format Day date5. ;
   label Day    ='Date of Measurement'
         KWattsL='Minimum Power Output'
         KWatts1='25th Percentile'
         KWattsX='Average Power Output'
         KWattsM='Median Power Output'
         KWatts3='75th Percentile'
         KWattsH='Maximum Power Output'
         KWattsS='Standard Deviation of Power Output'
         KWattsN='Group Sample Size';
   datalines;
05JUL94 3180 3340.0 3487.40 3490.0 3610.0 4050 220.3 20
06JUL94 3179 3333.5 3471.65 3419.5 3605.0 3849 210.4 20
07JUL94 3304 3376.0 3488.30 3456.5 3604.5 3781 147.0 20
08JUL94 3045 3390.5 3434.20 3447.0 3550.0 3629 157.6 20
11JUL94 2968 3321.0 3475.80 3487.0 3611.5 3916 258.9 20
12JUL94 3047 3425.5 3518.10 3576.0 3615.0 3881 211.6 20
13JUL94 3002 3368.5 3492.65 3495.5 3621.5 3787 193.8 20
14JUL94 3196 3346.0 3496.40 3473.5 3592.5 3994 212.0 20
15JUL94 3115 3188.5 3398.50 3426.0 3568.5 3731 199.2 20
18JUL94 3263 3340.0 3456.05 3444.0 3505.5 4040 173.5 20
;
```

Oilsum contains exactly one observation for each group. Note that, as in the previous example, the groups are indexed by the variable Day. A listing of Oilsum is shown in Figure 28.2.

Figure 28.2 The Summary Data Set Oilsum

Box Plot for Power Output

Day	KWattsL	KWatts1	KWattsX	KWattsM	KWatts3	KWattsH	KWattsS	KWattsN
05JUL	3180	3340.0	3487.40	3490.0	3610.0	4050	220.3	20
06JUL	3179	3333.5	3471.65	3419.5	3605.0	3849	210.4	20
07JUL	3304	3376.0	3488.30	3456.5	3604.5	3781	147.0	20
08JUL	3045	3390.5	3434.20	3447.0	3550.0	3629	157.6	20
11JUL	2968	3321.0	3475.80	3487.0	3611.5	3916	258.9	20
12JUL	3047	3425.5	3518.10	3576.0	3615.0	3881	211.6	20
13JUL	3002	3368.5	3492.65	3495.5	3621.5	3787	193.8	20
14JUL	3196	3346.0	3496.40	3473.5	3592.5	3994	212.0	20
15JUL	3115	3188.5	3398.50	3426.0	3568.5	3731	199.2	20
18JUL	3263	3340.0	3456.05	3444.0	3505.5	4040	173.5	20

There are eight summary variables in Oilsum:

KWattsL contains the group minima (low values).

KWatts1 contains the 25th percentile (first quartile) for each group.

KWattsX contains the group means.

KWattsM contains the group medians.

KWatts3 contains the 75th percentile (third quartile) for each group.

KWattsH contains the group maxima (high values).

KWattsS contains the group standard deviations.

KWattsN contains the group sizes.

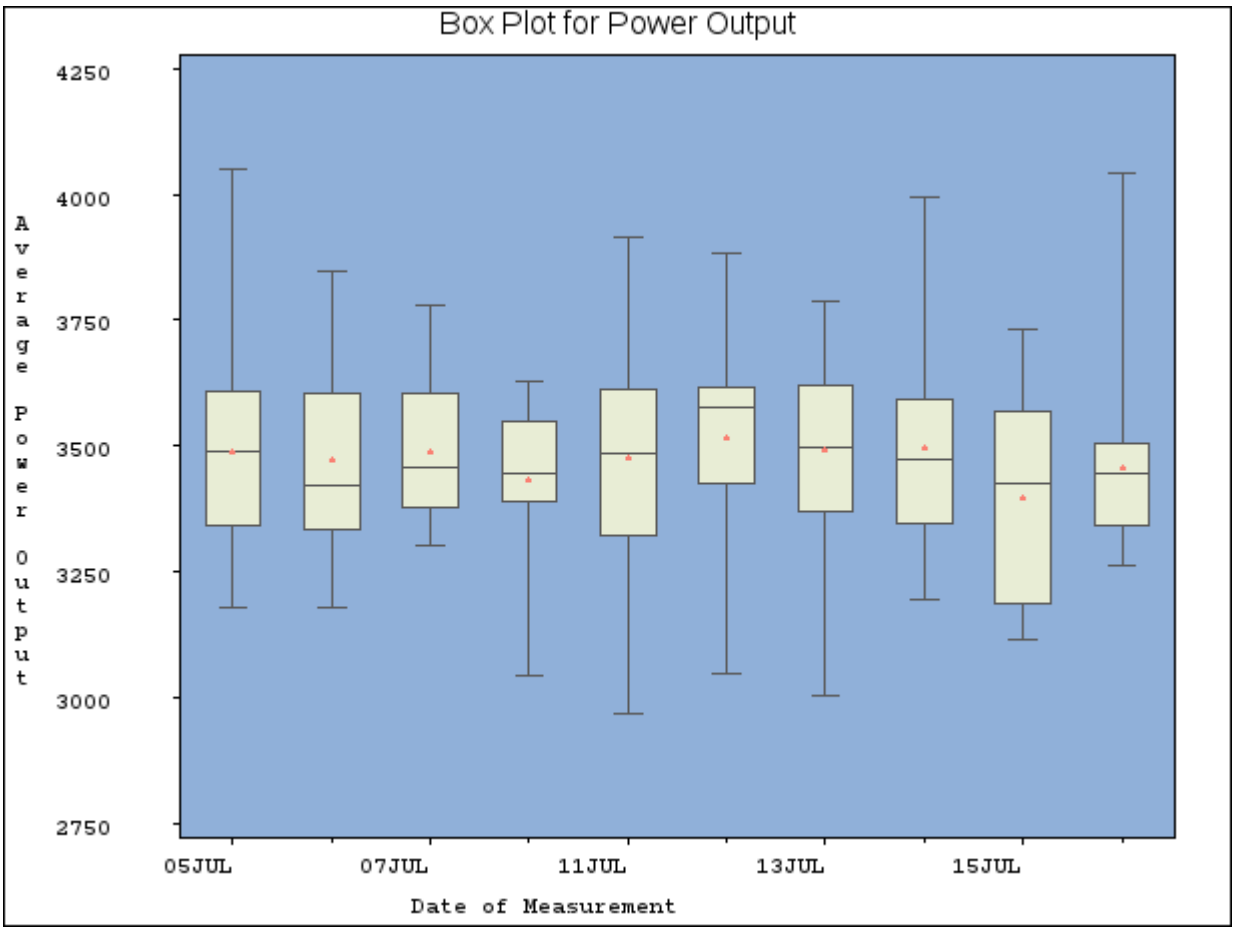
You can use this data set as input to the BOXPLOT procedure by specifying it with the HISTORY= option in the PROC BOXPLOT statement. Detailed requirements for HISTORY= data sets are presented in the section "HISTORY= Data Set" on page 1167.

The following statements produce a box plot of the summary data from the Oilsum data set:

```
options nogstyle;
title 'Box Plot for Power Output';
symbol value=dot color=salmon;
proc boxplot history=Oilsum;
   plot KWatts*Day / cframe   = vligb
                     cboxes   = dagr
                     cboxfill = ywh;
run;
options gstyle;
goptions reset=symbol;
```

The NOGSTYLE system option causes PROC BOXPLOT not to use ODS style information when it produces a traditional graphics box plot. Instead, the SYMBOL statement and options specified after the slash (/) in the PLOT statement control its appearance. The CFRAME= option specifies the background color of the graph frame, the CBOXES= option specifies the color of the box outlines and whiskers, and the CBOXFILL= option specifies the color of the box interiors. The GSTYLE system option restores the use of ODS styles for subsequent traditional graphics output. For more information about SYMBOL statements and the colors specified in the PLOT statement options, see *SAS/GRAPH: Reference*. The resulting box plot is shown in Figure 28.3.

Figure 28.3 Traditional Graphics Box Plot with NOGSTYLE



Saving Summary Data with Outliers

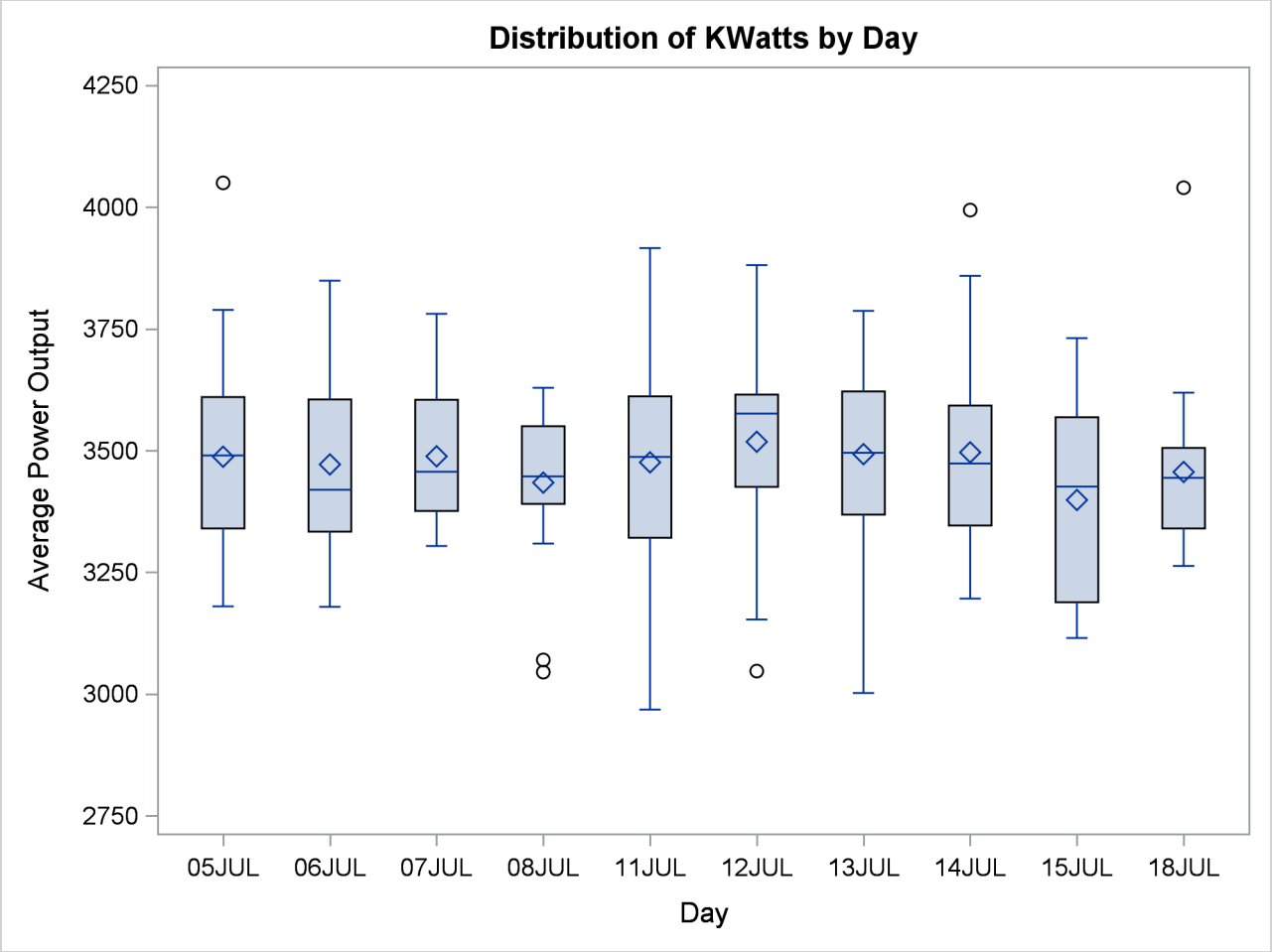
In a *schematic* box plot, outlier values within a group are plotted as separate points beyond the whiskers of the box-and-whiskers plot. See the section "Styles of Box Plots" on page 1168 and the description of the BOXSTYLE= option for a complete description of schematic box plots.

The following statements use the BOXSTYLE= option to produce a schematic box plot of the data from the Turbine data set. The OUTBOX= option creates a summary data set named OilSchematic. The ODS GRAPHICS ON statement specified before the PROC BOXPLOT statement enables ODS Graphics, so the box plot is created using ODS Graphics instead of traditional graphics.

```
title 'Schematic Box Plot for Power Output';
ods graphics on;
proc boxplot data=Turbine;
   plot KWatts*Day / boxstyle = schematic
                     outbox   = OilSchematic;
run;
```

The schematic box plot is shown in Figure 28.4. Note the outliers plotted for several of the groups.

Figure 28.4 Schematic Box Plot of Power Output



Whereas the Oilsum data set from the section "Creating Box Plots from Summary Data" on page 1125 contains a *variable* for each summary statistic and one observation per group, the OUTBOX= data set OilSchematic contains one *observation* for each summary statistic in each group. The _TYPE_ variable identifies the statistic and the _VALUE_ variable contains its value. In addition, the OilSchematic data set contains an observation recording each outlier value for each group. Figure 28.5 shows a partial listing of the OilSchematic data set.

Figure 28.5 The Summary Data Set OilSchematic

Schematic Box Plot for Power Output

Day	_VAR_	_TYPE_	_VALUE_
05JUL	KWatts	N	20.00
05JUL	KWatts	MIN	3180.00
05JUL	KWatts	Q1	3340.00
05JUL	KWatts	MEAN	3487.40
05JUL	KWatts	MEDIAN	3490.00
05JUL	KWatts	Q3	3610.00
05JUL	KWatts	MAX	4050.00
05JUL	KWatts	STDDEV	220.26
05JUL	KWatts	HIWHISKR	3789.00
05JUL	KWatts	HIGH	4050.00
06JUL	KWatts	N	20.00
06JUL	KWatts	MIN	3179.00
06JUL	KWatts	Q1	3333.50
06JUL	KWatts	MEAN	3471.65
06JUL	KWatts	MEDIAN	3419.50
06JUL	KWatts	Q3	3605.00
06JUL	KWatts	MAX	3849.00
06JUL	KWatts	STDDEV	210.43
07JUL	KWatts	N	20.00
07JUL	KWatts	MIN	3304.00
07JUL	KWatts	Q1	3376.00
07JUL	KWatts	MEAN	3488.30
07JUL	KWatts	MEDIAN	3456.50
07JUL	KWatts	Q3	3604.50
07JUL	KWatts	MAX	3781.00
07JUL	KWatts	STDDEV	147.02
08JUL	KWatts	N	20.00
08JUL	KWatts	MIN	3045.00
08JUL	KWatts	Q1	3390.50
08JUL	KWatts	MEAN	3434.20
08JUL	KWatts	MEDIAN	3447.00
08JUL	KWatts	Q3	3550.00
08JUL	KWatts	MAX	3629.00
08JUL	KWatts	STDDEV	157.64
08JUL	KWatts	LOWHISKR	3309.00
08JUL	KWatts	LOW	3070.00
08JUL	KWatts	LOW	3045.00
11JUL	KWatts	N	20.00
11JUL	KWatts	MIN	2968.00
11JUL	KWatts	Q1	3321.00

Observations with the _TYPE_ variable values “HIGH” and “LOW” contain outlier values. If you want to use a summary data set to re-create a schematic box plot, you *must* create an OUTBOX= data set in order to save the outlier data.

Syntax: BOXPLOT Procedure

The following statements are available in the BOXPLOT procedure:

PROC BOXPLOT *options* **;**
BY *variables* **;**
ID *variables* **;**
INSET *keywords* < / *options* > **;**
INSETGROUP *keywords* < / *options* > **;**
PLOT *analysis-variable group-variable* < **(***block-variables***)** > < **=***symbol-variable* > < / *options* > **;**

Both the PROC BOXPLOT and PLOT statements are required. You can specify any number of PLOT statements within a single PROC BOXPLOT invocation.

PROC BOXPLOT Statement

PROC BOXPLOT *options* **;**

The PROC BOXPLOT statement invokes the BOXPLOT procedure. Table 28.1 summarizes the *options* available in the PROC BOXPLOT statement.

Table 28.1 PROC BOXPLOT Statement Options

Statement	Description
ANNOTATE=	Enhances traditional graphics box plots
BOX=	Names an input data set containing group summary statistics and outlier values
DATA=	Names an input data set containing raw data to be analyzed
GOUT=	Specifies the SAS catalog in which to save traditional graphics output
HISTORY=	Names an input data set containing group summary statistics

The following *options* can appear in the PROC BOXPLOT statement.

ANNOTATE=*SAS-data-set*

ANNO=*SAS-data-set*

specifies an ANNOTATE= type data set, as described in *SAS/GRAPH: Reference*, which enhances traditional graphics box plots requested in subsequent PLOT statements. **NOTE:** The ANNOTATE= option is ignored when ODS Graphics is enabled.

BOX=*SAS-data-set*

names an input data set containing group summary statistics and outlier values. Typically, this data set is created as an OUTBOX= data set in a previous run of PROC BOXPLOT. Each group summary statistic or outlier value is recorded in a separate observation in a BOX= data set, so there are multiple observations per group. You cannot use a BOX= data set together with a DATA= or HISTORY= data set. If you do not specify one of these input data sets, the procedure uses the most recently created SAS data set as a DATA= data set.

DATA=*SAS-data-set*
names an input data set containing raw data to be analyzed. You cannot use a DATA= data set together with a BOX= or HISTORY= data set. If you do not specify one of these input data sets, the procedure uses the most recently created SAS data set as a DATA= data set.

GOUT=*< libref. >output catalog*
specifies the SAS catalog in which to save traditional graphics output that is produced by the BOXPLOT procedure. If you omit the libref, PROC BOXPLOT looks for the catalog in the temporary library called WORK and creates the catalog if it does not exist. **NOTE:** The GOUT= option is ignored when ODS Graphics is enabled.

HISTORY=*SAS-data-set*

HIST=*SAS-data-set*
names an input data set containing group summary statistics. Typically, this data set is created as an OUTHISTORY= data set in a previous run of PROC BOXPLOT, but it can also be created using a SAS summarization procedure such as the MEANS procedure. The HISTORY= data set can contain only one observation for each value of the group variable. You cannot use a HISTORY= data set with a DATA= or BOX= data set. If you do not specify one of these three input data sets, PROC BOXPLOT uses the most recently created data set as a DATA= data set.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC BOXPLOT 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 for the BOXPLOT 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*.

ID Statement

ID *variables* **;**

The ID statement specifies variables used to identify observations. The ID variables must be variables in the input data set.

If you specify the keyword SCHEMATICID or SCHEMATICIDFAR with the BOXSTYLE= option, the value of an ID variable is used to label each extreme observation. When you specify a BOX= data set, the label values come from the variable _ID_, if it is present in the data set. When you specify a DATA= or HISTORY= input data set, or a BOX= data set that does not contain the variable _ID_, the labels come from the first variable listed in the ID statement. If ID statement is specified, the outliers are not labeled.

INSET Statement

INSET *keywords* </ *options*> **;**

A PLOT statement in the BOXPLOT procedure can be followed by a series of INSET and INSETGROUP statements. Each INSET statement in that series produces one inset in the box plot produced by the preceding PLOT statement. If the box plot occupies multiple panels, the inset appears on each panel.

The data requested using the *keywords* are displayed in the order in which they are specified. Summary statistics requested with an INSET statement are calculated using the observations in all groups.

keywords identify summary statistics or other data to be displayed in the inset. By default, inset statistics are identified with appropriate labels, and numeric values are printed using appropriate formats. However, you can provide customized labels and formats. You provide the customized label by specifying the *keyword* for that statistic followed by an equal sign (=) and the label in quotes. Labels can have up to 24 characters. You provide the numeric format in parentheses after the *keyword*. Note that if you specify both a label and a format for a statistic, the label must appear before the format.

The available *keywords* are listed in Table 28.2.

options control the appearance of the inset. Most of these *options* apply only to traditional graphics and are ignored when ODS Graphics is enabled. Table 28.3 summarizes the *options* available in the INSET statement. It also lists *options* and identifies those that are valid when ODS Graphics is enabled. Complete descriptions for each *option* follow.

Table 28.2 INSET Statement Keywords

Keyword	Description
DATA=	(label, value) pairs from *SAS-data-set*
MEAN	mean of all observations
MIN	minimum observed value
MAX	maximum observed value
NMIN	minimum group size
NMAX	maximum group size
NOBS	number of observations in box plot
STDDEV	pooled standard deviation

The DATA= *keyword* specifies a SAS data set containing (label, value) pairs to be displayed in an inset. The data set must contain the variables _LABEL_ and _VALUE_. _LABEL_ is a character variable of up to 24 characters whose values provide labels for inset entries. _VALUE_ can be character or numeric, and provides values displayed in the inset. The label and value from each observation in the DATA= data set occupy one line in the inset.

The *pooled standard deviation* requested with the STDDEV keyword is defined as

$$s_p = \sqrt{\frac{\sum_{i=1}^{N} s_i^2 . n_i \quad 1/}{\sum_{i=1}^{N} . n_i \quad 1/}}$$

where N is the number of groups, n_i is the size of the ith group, and s_i^2 is the variance of the ith group.

Table 28.3 INSET Statement Options

Option	Description	ODS Graphics
CFILL=	Specifies color of inset background	
CFILLH=	Specifies color of inset header background	
CFRAME=	Specifies color of inset frame	
CHEADER=	Specifies color of inset header text	
CSHADOW=	Specifies color of inset drop shadow	
CTEXT=	Specifies color of inset text	
DATA	Specifies data units for POSITION=.x; y/ co-ordinates	
FONT=	Specifies font of inset text	
FORMAT=	Specifies format of values in inset	X
HEADER=	Specifies inset header text	X
HEIGHT=	Specifies height of inset and header text	
NOFRAME	Suppresses frame around inset	X
POSITION=	Specifies position of inset	X
REFPOINT=	Specifies reference point of inset positioned with POSITION=.x; y/ coordinates	

Following are descriptions of the *options* that you can specify in the INSET statement after a slash (/). Only those *options* marked with † are applicable when ODS Graphics is enabled.

CFILL=*color* **| BLANK**
specifies the color of the inset background (including the header background if you do not specify the CFILLH= option).

If you do not specify the CFILL= option, then by default the background is empty. This means that items that overlap the inset (such as box-and-whiskers plots or reference lines) show through the inset. If you specify any value for the CFILL= option, then overlapping items no longer show through the inset. Specify CFILL=BLANK to leave the background uncolored and also to prevent items from showing through the inset.

CFILLH=*color*
specifies the color of the header background. By default, if you do not specify a CFILLH= color, the CFILL= color is used.

CFRAME=*color*
specifies the color of the frame around the inset. By default, the frame is the same color as the axis of the plot.

CHEADER=*color*
specifies the color of the header text. By default, if you do not specify a CHEADER= color, the INSET statement CTEXT= color is used.

CSHADOW=*color*

CS=*color*
specifies the color of the drop shadow. If you do not specify the CSHADOW= option, a drop shadow is not displayed.

CTEXT=*color*

CT=*color*
specifies the color of the text in the inset. By default, the inset text color is the same as the other text in the box plot.

DATA
specifies that data coordinates be used in positioning the inset with the POSITION= option. The DATA option is available only when you specify POSITIOND .x; y/, and it must be placed immediately after the coordinates .x; y/. See the entry for the POSITION= option.

FONT=*font*
specifies the font of the text.

† FORMAT=*format*
specifies a format for all the values displayed in an inset. If you specify a format for a particular statistic, then this format overrides the format you specified with the FORMAT= option.

† HEADER='*string***'**
specifies the header text. The *string* can be up to 40 characters. If you do not specify the HEADER= option, no header line appears in the inset.

HEIGHT=*value*
specifies the height of the inset and header text.

† **NOFRAME**
suppresses the frame drawn around the inset.

† **POSITION=***position*

† **POS=***position*
determines the position of the inset. The *position* can be a compass point keyword, a margin keyword, or (for traditional graphics) a pair of coordinates .x; y/. You can specify coordinates in axis percent units or axis data units. For more information, see the section "Positioning Insets" on page 1172. By default, POSITION=NW, which positions the inset in the upper-left (northwest) corner of the plot.

REFPOINT=BR | BL | TR | TL

RP=BR | BL | TR | TL
specifies the reference point for an inset that is positioned by a pair of coordinates with the POSITION= option. Use the REFPOINT= option with POSITION= coordinates. The REFPOINT= option specifies which corner of the inset frame you want positioned at coordinates .x; y/. The keywords BL, BR, TL, and TR represent bottom left, bottom right, top left, and top right, respectively. The default is REFPOINT=BL.

If you specify the position of the inset as a compass point or margin keyword, the REFPOINT= option is ignored.

INSETGROUP Statement

INSETGROUP *keywords* </ *options*> **;**

A PLOT statement in the BOXPLOT procedure can be followed by a series of INSET and INSETGROUP statements. Each INSETGROUP statement in that series displays statistics associated with individual groups in the box plot produced by the preceding PLOT statement. **NOTE:** The INSETGROUP statement is ignored when you specify the HORIZONTAL option with ODS Graphics enabled. No more than two INSETGROUP statements can be associated with a given PLOT statement: one that displays group statistics above the box plot and one that displays group statistics below it. The data requested using the *keywords* are displayed in the order in which they are specified.

keywords identify summary statistics to be displayed in the insets. By default, inset statistics are identified with appropriate labels, and numeric values are printed using appropriate formats. However, you can provide customized labels and formats. You provide the customized label by specifying the *keyword* for that statistic followed by an equal sign (=) and the label in quotes. Labels can have up to 24 characters. You provide the numeric format in parentheses after the *keyword*. Note that if you specify both a label and a format for a statistic, the label must appear before the format. The *keywords* are listed in Table 28.4.

options control the appearance of the insets. Table 28.5 lists all the *options* in the INSETGROUP statement. Complete descriptions for each *option* follow.

Table 28.4 INSETGROUP Statement Keywords

Keyword	Description
MEAN	group mean
MIN	group minimum value or low whisker value
MAX	group maximum value or high whisker value
N	number of observations in group
NHIGH	number of outliers above upper fence
NLOW	number of outliers below lower fence
NOUT	total number of outliers in group
Q1	first quartile of group values
Q2	second quartile of group values
Q3	third quartile of group values
RANGE	range of group values
STDDEV	group standard deviation

NOTE: The NHIGH, NLOW, and NOUT keywords are not supported when ODS Graphics is enabled,

Table 28.5 summarizes the *options* available in the INSETGROUP statement. All of these *options* apply to traditional graphics only. They are ignored when ODS Graphics is enabled.

Table 28.5 INSETGROUP Statement Options

Option	Description
CFILL=	Specifies color of inset background
CFILLH=	Specifies color of inset header background
CFRAME=	Specifies color of inset frame
CHEADER=	Specifies color of inset header text
CTEXT=	Specifies color of inset text
FONT=	Specifies font of inset text
FORMAT=	Specifies format of values in inset
HEADER=	Specifies inset header text
HEIGHT=	Specifies height of inset and header text
NOFRAME	Suppresses frame around inset
POSITION=	Specifies position of inset

Following are descriptions of the *options* that you can specify in the INSETGROUP statement after a slash (/).

CFILL=*color*
specifies the color of the inset background (including the header background if you do not specify the CFILLH= option). If you do not specify the CFILL= option, then by default the background is empty.

CFILLH=*color*
specifies the color of the header background. By default, if you do not specify a CFILLH= color, the CFILL= color is used.

CFRAME=*color*
specifies the color of the frame around the inset. By default, the frame is the same color as the axis of the plot.

CHEADER=*color*
specifies the color of the header text. By default, if you do not specify a CHEADER= color, the CTEXT= color is used.

CTEXT=*color*

CT=*color*
specifies the color of the inset text. By default, the inset text color is the same as the other text in the plot.

FONT=*font*
specifies the font of the inset text. By default, the font is SIMPLEX.

FORMAT=*format*
specifies a format for all the values displayed in an inset. If you specify a format for a particular statistic, then this format overrides the format you specified with the FORMAT= option.

HEADER='*string***'**
specifies the header text. The *string* can be up to 40 characters. If you do not specify the HEADER= option, no header line appears in the inset.

HEIGHT=*value*
specifies the height of the inset and header text.

NOFRAME
suppresses the frame drawn around the inset.

POSITION=*position*

POS=*position*
determines the position of the inset. Valid positions are TOP, TOPOFF, AXIS, and BOTTOM. By default, POSITION=TOP.

Position Keyword	Description
TOP	top of plot, immediately above axis frame
TOPOFF	top of plot, offset from axis frame
AXIS	bottom of plot, immediately above horizontal axis
BOTTOM	bottom of plot, below horizontal axis label

PLOT Statement

PLOT (*analysis-variables***)** *group-variable* < **(***block-variables***)** > < **=***symbol-variable* >
< / *options* > **;**

You can specify multiple PLOT statements after the PROC BOXPLOT statement. The components of the PLOT statement are as follows:

analysis-variables identify one or more variables to be analyzed. An analysis variable is required. If you specify more than one analysis variable, enclose the list in parentheses. For example, the following statements request distinct box plots for the variables Weight, Length, and Width:

```
proc boxplot data=Summary;
   plot (Weight Length Width)*Day;
run;
```

group-variable specifies the variable that identifies groups in the data. The group variable is required. In the preceding PLOT statement, Day is the group variable.

block-variables specify optional variables that group the data into blocks of consecutive groups. These blocks are labeled in a legend, and each block variable provides one level of labels in the legend.

symbol-variable specifies an optional variable whose levels (unique values) determine the symbol marker used to plot the means. Distinct symbol markers are displayed for points corresponding to the various levels of the symbol variable. You can specify the symbol markers with SYMBOL*n* statements (see *SAS/GRAPH: Reference* for complete details).

options enhance the appearance of the box plot, request additional analyses, save results in data sets, and so on. Complete descriptions of each option follow.

PLOT Statement Options

Many PLOT statement *options* apply only to traditional graphics and are ignored when ODS Graphics is enabled. Table 28.6 summarizes the *options* available in the PLOT statement. It also lists *options* by function and indicates which are applicable with ODS Graphics.

Table 28.6 PLOT Statement Options

Option	Description	ODS Graphics
Options for Controlling Box Appearance		
BOXCONNECT=	Connects features of adjacent box-and-whiskers plots with line segments	X
BOXSTYLE=	Specifies style of box-and-whiskers plots	X
BOXWIDTH=	Specifies width of box-and-whiskers plots	
BOXWIDTHSCALE=	Specifies that widths of box-and-whiskers plots vary proportionately to group size	X
CBOXES=	Specifies color for outlines of box-and-whiskers plots	

Table 28.6 *continued*

Option	Description	ODS Graphics
CBOXFILL=	Specifies fill color for interior of box-and-whiskers plots	
IDCOLOR=	Specifies outlier symbol color in schematic box-and-whiskers plots	
IDCTEXT=	Specifies outlier label color in schematic box-and-whiskers plots	
IDFONT=	Specifies outlier label font in schematic box-and-whiskers plots	
IDHEIGHT=	Specifies outlier label height in schematic box-and-whiskers plots	
IDSYMBOL=	Specifies outlier symbol in schematic box-and-whiskers plots	
IDSYMBOLHEIGHT=	Specifies outlier symbol height in schematic box-and-whiskers plots	
LBOXES=	Specifies line types for outlines of box-and-whiskers plots	
NOSERIFS	Eliminates serifs from whiskers of box-and-whiskers plots	X
NOTCHES	Specifies that box-and-whiskers plots be notched	X
PCTLDEF=	Specifies percentile definition used for box-and-whiskers plots	X
WHISKERPERCENTILE=	Specifies that box whiskers be drawn to percentile values	X
Options for Plotting and Labeling Points		
ALLLABEL=	Labels means of box-and-whiskers plots	
CLABEL=	Specifies color for labels requested with ALLLABEL= option	
CCONNECT=	Specifies color for line segments requested with BOXCONNECT= option	
LABELANGLE=	Specifies angle for labels requested with ALLLABEL= option	
SYMBOLLEGEND=	Specifies LEGEND statement for levels of symbol variable	
SYMBOLORDER=	Specifies order in which symbols are assigned for levels of symbol variable	
Reference Line Options		
CHREF=	Specifies color for lines requested by HREF= option	
CVREF=	Specifies color for lines requested by VREF= option	
FRONTREF	Draws reference lines in front of boxes	
HREF=	Requests reference lines perpendicular to horizontal axis	X
HREFLABELS=	Specifies labels for HREF= lines	X
HREFLABPOS=	Specifies position of HREFLABELS= labels	
LHREF=	Specifies line type for HREF= lines	
LVREF=	Specifies line type for VREF= lines	
NOBYREF	Specifies that reference line information in a data set be applied uniformly to plots created for all BY groups	X
VREF=	Requests reference lines perpendicular to vertical axis	X
VREFLABELS=	Specifies labels for VREF= lines	X

Table 28.6 *continued*

Option	Description	ODS Graphics
VREFLABPOS=	Specifies position of VREFLABELS= labels	
Block Variable Legend Options		
BLOCKLABELPOS=	Specifies position of label for block variable legend	
BLOCKLABTYPE=	Specifies text size of block variable legend	
BLOCKPOS=	Specifies vertical position of block variable legend	X
BLOCKREP	Repeats identical consecutive labels in block variable legend	X
CBLOCKLAB=	Specifies colors for filling frames enclosing block variable labels	
CBLOCKVAR=	Specifies colors for filling background of block variable legend	
Axis and Axis Label Options		
CAXIS=	Specifies color for axis lines and tick marks	
CFRAME=	Specifies fill color for frame for plot area	
CONTINUOUS	Produces horizontal axis for continuous group variable values (traditional graphics only)	
CTEXT=	Specifies color for tick mark values and axis labels	
HAXIS=	Specifies major tick mark values for horizontal axis	
HEIGHT=	Specifies height of axis label and axis legend text	
HMINOR=	Specifies number of minor tick marks between major tick marks on horizontal axis	
HOFFSET=	Specifies length of offset at both ends of horizontal axis	
NOHLABEL	Suppresses horizontal axis label	X
NOTICKREP	Specifies that only first occurrence of repeated, adjacent character group values be labeled on horizontal axis	
NOVANGLE	Requests vertical axis labels that are strung out vertically	
SKIPHLABELS=	Specifies thinning factor for tick mark labels on horizontal axis	
TURNHLABELS	Requests horizontal tick labels that are strung out vertically	
VAXIS=	Specifies major tick mark values for vertical axis	X
VFORMAT=	Specifies format for vertical axis tick marks	X
VMINOR=	Specifies number of minor tick marks between major tick marks on vertical axis	
VOFFSET=	Specifies length of offset at both ends of vertical axis	
VZERO	Forces origin to be included in vertical axis	
WAXIS=	Specifies width of axis lines	
Input Data Set Options		
MISSBREAK	Specifies that a missing value between identical character group values signify the start of a new group	X
Output Data Set Options		
OUTBOX=	Produces an output data set containing group summary statistics and outlier values	X

Table 28.6 *continued*

Option	Description	ODS Graphics
OUTHISTORY=	Produces an output data set containing group summary statistics	X
Graphical Enhancement Options		
ANNOTATE=	Specifies annotate data set that adds features to box plot	
BWSLEGEND	Displays a legend identifying the function of group size specified with BOXWIDTHSCALE= option	
DESCRIPTION=	Specifies string that appears in description field of PROC GREPLAY master menu for traditional graphics box plot	
FONT=	Specifies font for labels and legends on plots	
HORIZONTAL	Requests a horizontal box plot with ODS Graphics	X
HTML=	Specifies URLs to be associated with box-and-whiskers plots	
NAME=	Specifies name that appears in name field of PROC GREPLAY master menu for traditional graphics box plot	
NLEGEND	Requests legend displaying group sizes	
OUTHIGHHTML=	Specifies URLs to be associated with high outliers on box-and-whiskers plots	
OUTLOWHTML=	Specifies URLs to be associated with low outliers on box-and-whiskers plots	
PAGENUM=	Specifies form of label used in pagination	
PAGENUMPOS=	Specifies position of page number requested with PAGENUM= option	
Grid Options		
CGRID=	Specifies color for grid requested with ENDGRID or GRID option	
ENDGRID	Adds grid after last box-and-whiskers plot	
GRID	Adds grid to box plot	X
LENDGRID=	Specifies line type for grid requested with ENDGRID option	
LGRID=	Specifies line type for grid requested with GRID option	
WGRID=	Specifies width of grid lines	
Plot Layout Options		
INTERVAL=	Specifies natural time interval between consecutive group positions when time, date, or datetime format is associated with numeric group variable	
INTSTART=	Specifies first major tick mark value on horizontal axis when date, time, or datetime format is associated with numeric group variable	
MAXPANELS=	Specifies maximum number of panels used for box plot	X
NOCHART	Suppresses creation of box plot	X
NOFRAME	Suppresses frame for plot area	
NPANELPOS=	Specifies number of group positions per panel	X
REPEAT	Repeats last group position on panel as first group position of next panel	X

Table 28.6 *continued*

Option	Description	ODS Graphics
TOTPANELS=	Specifies number of panels to be used to display box plot	X
Overlay Options		
CCOVERLAY=	Specifies colors for line segments connecting points on overlays	
COVERLAY=	Specifies colors for points on overlays	
LOVERLAY=	Specifies line types for line segments connecting points on overlays	
NOOVERLAYLEGEND	Suppresses overlay legend	X
OVERLAY=	Specifies variables to be plotted on overlays	X
OVERLAYHTML=	Specifies URLs to be associated with overlay plot points	
OVERLAYID=	Specifies labels for overlay plot points	
OVERLAYLEGLAB=	Specifies label for overlay legend	X
OVERLAYSYM=	Specifies symbols used for overlays	
OVERLAYSYMHT=	Specifies heights for overlay symbols	
WOVERLAY=	Specifies widths for line segments connecting points on overlays	
Clipping Options		
CCLIP=	Specifies color for plot symbol for clipped points	
CLIPFACTOR=	Determines extent to which extreme values are clipped	X
CLIPLEGEND=	Specifies text for clipping legend	X
CLIPLEGPOS=	Specifies position of clipping legend	
CLIPSUBCHAR=	Specifies substitution character for CLIPLEGEND= text	X
CLIPSYMBOL=	Specifies plot symbol for clipped points	
CLIPSYMBOLHT=	Specifies symbol marker height for clipped points	
COVERLAYCLIP=	Specifies color for clipped points on overlays	
OVERLAYCLIPSYM=	Specifies symbol for clipped points on overlays	
OVERLAYCLIPSYMHT=	Specifies symbol height for clipped points on overlays	
Options for Box Plots Produced Using Styles		
BLOCKVAR=	Groups block legends whose backgrounds are filled with colors from style	X
BOXES=	Groups boxes whose outlines are drawn with colors from style	
BOXFILL=	Groups boxes that are filled with colors from style	
Options for ODS Graphics Output		
ODSFOOTNOTE=	Specifies a footnote displayed in ODS Graphics output	X
ODSFOOTNOTE2=	Specifies a secondary footnote displayed in ODS Graphics output	X
ODSTITLE=	Specifies a title displayed in ODS Graphics output	X
ODSTITLE2=	Specifies a secondary title displayed in ODS Graphics output	X

Following are explanations of the *options* you can specify in the PLOT statement after a slash (/). Only those *options* marked with † are applicable when ODS Graphics is enabled.

ALLLABEL=VALUE | (*variable*)
labels the point plotted for the mean of each box-and-whiskers plot with the mean (when ALLLABEL=VALUE) or with the value of the ALLLABEL=*variable* from the input data set.

ANNOTATE=*SAS-data-set*
specifies an ANNOTATE= type data set, as described in *SAS/GRAPH: Reference*.

BLOCKLABELPOS=ABOVE | LEFT
specifies the position of a block variable label in the block legend. The keyword ABOVE places the label immediately above the legend, and LEFT places the label to the left of the legend. Use the keyword LEFT with labels that are short enough to fit in the margin of the plot; otherwise, they are truncated. The default keyword is ABOVE.

BLOCKLABTYPE=SCALED | TRUNCATED | *height*
specifies how lengthy block variable values are treated when there is insufficient space to display them in the block legend. If you specify BLOCKLABTYPE=SCALED, the values are uniformly reduced in height so that they fit. If you specify BLOCKLABTYPE=TRUNCATED, lengthy values are truncated on the right until they fit. You can also specify a text height in vertical percent screen units for the values. By default, lengthy values are not displayed. For more information, see the section "Displaying Blocks of Data" on page 1177.

† BLOCKPOS=*n*
specifies the vertical position of the legend for the values of the block variables. Values of *n* and the corresponding positions are as follows. By default, BLOCKPOS=1.

n	Legend Position
1	top of plot, offset from axis frame
2	top of plot, immediately above axis frame
3	bottom of plot, immediately above horizontal axis
4	bottom of plot, below horizontal axis label

† BLOCKREP
specifies that block variable values for all groups be displayed. By default, only the first block variable value in any block is displayed, and repeated block variable values are not displayed.

† BLOCKVAR=*variable* **| (***variable-list***)**
specifies variables whose values are used to assign colors for filling the background of the legend associated with block variables. A list of BLOCKVAR= variables must be enclosed in parentheses. BLOCKVAR= variables are matched with block variables by their order in the respective variable lists. While the values of a CBLOCKVAR= variable are color names, values of a BLOCKVAR= variable are used to group block legends for assigning fill colors from the ODS style. Block legends with the same BLOCKVAR= variable value are filled with the same color.

† BOXCONNECT=MEAN | MEDIAN | MAX | MIN | Q1 | Q3

† BOXCONNECT

specifies that the points in adjacent box-and-whiskers plots representing group means, medians, maximum values, minimum values, first quartiles, or third quartiles be connected with line segments. If the BOXCONNECT option is specified without a *keyword* identifying the points to be connected, group means are connected. By default, no points are connected.

BOXES=(*variable***)**

specifies a variable whose values are used to assign colors for the outlines of box-and-whiskers plots. While the values of a CBOXES= variable are color names, values of the BOXES= variable are used to group box-and-whiskers plots for assigning outline colors from the ODS style. The outlines of box-and-whiskers plots of groups with the same BOXES= variable value are drawn using the same color.

BOXFILL=(*variable***)**

specifies a variable whose values are used to assign fill colors for box-and-whiskers plots. While the values of a CBOXFILL= variable are color names, values of the BOXFILL= variable are used to group box-and-whiskers plots for assigning fill colors from the ODS style. Box-and-whiskers plots of groups with the same BOXFILL= variable value are filled with the same color.

† BOXSTYLE=*keyword*

specifies the style of the box-and-whiskers plots displayed. If you specify BOXSTYLE=SKELETAL, the whiskers are drawn from the edges of the box to the extreme values of the group. This plot is sometimes referred to as a skeletal box-and-whiskers plot. By default, the whiskers are drawn with serifs. You can specify the NOSERIFS option to draw the whiskers without serifs.

In the following descriptions, the terms *fence* and *far fence* refer to the distance from the first and third quartiles (25th and 75th percentiles, respectively), expressed in terms of the interquartile range (IQR). For example, the lower fence is located at 1:5 IQR below the 25th percentile; the upper fence is located at 1:5 IQR above the 75th percentile. Similarly, the lower far fence is located at 3 IQR below the 25th percentile; the upper far fence is located at 3 IQR above the 75th percentile.

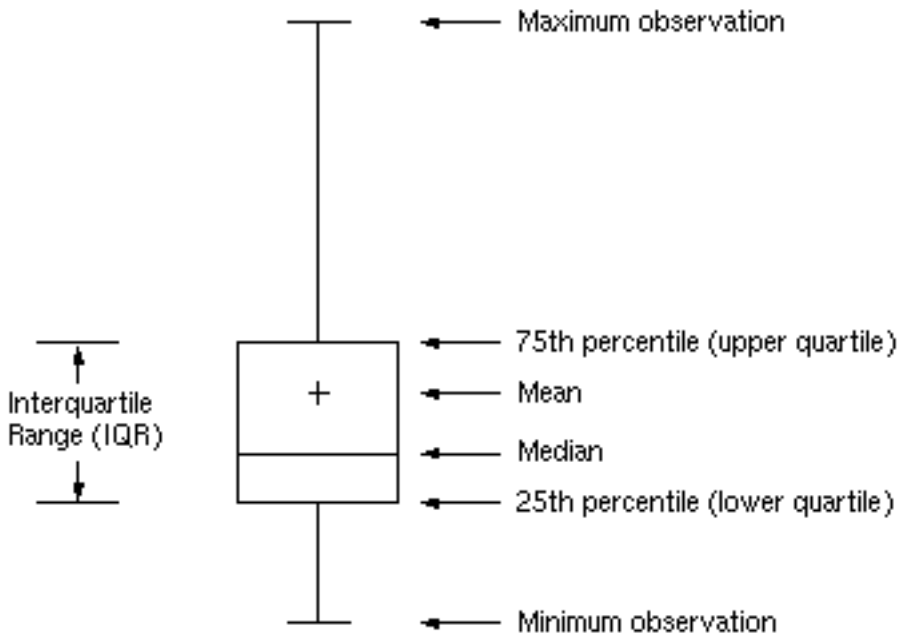
If you specify BOXSTYLE=SCHEMATIC, a whisker is drawn from the upper edge of the box to the largest observed value less than or equal to the upper fence, and another is drawn from the lower edge of the box to the smallest observed value greater than or equal to the lower fence. Serifs are added to the whiskers by default. Observations outside the fences are identified with a special symbol. For traditional graphics you can specify the shape and color for this symbol with the IDSYMBOL= and IDCOLOR= options. The default symbol is a square. This type of plot corresponds to the schematic box-and-whiskers plot described in Chapter 2 of Tukey (1977). See Figure 28.8 and the discussion in the section "Styles of Box Plots" on page 1168 for more information.

If you specify BOXSTYLE=SCHEMATICID, a schematic box-and-whiskers plot is displayed in which an ID variable value is used to label the symbol marking each observation outside the upper and lower fences. A BOX= data set can contain a variable named _ID_ that is used as the ID variable. Otherwise, the first variable listed in the ID statement provides the labels.

If you specify BOXSTYLE=SCHEMATICIDFAR, a schematic box-and-whiskers plot is displayed in which the value of the ID variable is used to label the symbol marking each observation outside the lower and upper far fences. Observations between the fences and the far fences are identified with a symbol but are not labeled with the ID variable.

Figure 28.6 illustrates the elements of a skeletal box-and-whiskers plot.

Figure 28.6 Skeletal Box-and-Whiskers Plot



The skeletal style of the box-and-whiskers plot shown in Figure 28.6 is the default.

BOXWIDTH=*value*
specifies the width (in horizontal percent screen units) of the box-and-whiskers plots.

† BOXWIDTHSCALE=*value*
specifies that the widths of the box-and-whisker plots are to vary according to a particular function of the group size *n*. The function, f .n/, is determined by the specified *value* (0).

If you specify a positive *value*, f .n/ D n^{value}. In particular, if you specify BOXWIDTHSCALE=1, f .n/ D n. If you specify BOXWIDTHSCALE=0.5, f .n/ D $\sqrt{n}$, as described by McGill, Tukey, and Larsen (1978).

If you specify BOXWIDTHSCALE=0, f .n/ D log.n/.

The box widths vary between minimum (w_{min}) and maximum (w_{max}) widths that are determined by the output destination. The width of the *i*th box is

$$w_i \text{ D } w_{min} \text{ C } .w_{max} \quad w_{min}/\frac{f\ .n_i/ \quad f\ .n_{min}/}{f\ .n_{max}/ \quad f\ .n_{min}/}$$

where n_{min} is the minimum group size and n_{max} is the maximum group size.

By default, the box widths are constant.

See Example 28.5 for an illustration of the BOXWIDTHSCALE= option.

You can specify the BWSLEGEND option to display a legend that identifies f .n/.

BWSLEGEND
displays a legend identifying the function of group size *n* specified with the BOXWIDTHSCALE= option. No legend is displayed if all group sizes are equal. The BWSLEGEND option is not applicable unless you also specify the BOXWIDTHSCALE= option.

CAXIS=*color*

CAXES=*color*

CA=*color*
specifies the color for the axes and tick marks. This option overrides any COLOR= specifications in an AXIS statement.

CBLOCKLAB=*color* **| (***color-list***)**
specifies fill colors for the frames that enclose the block variable labels in a block legend. By default, these areas are not filled. Colors in the CBLOCKLAB= list are matched with block variables in the order in which they appear in the PLOT statement.

CBLOCKVAR=*variable* **| (***variable-list***)**
specifies variables whose values are colors for filling the background of the legend associated with block variables. CBLOCKVAR= variables are matched with block variables by their order in the respective variable lists. Each CBLOCKVAR= variable must be a character variable of no more than eight characters in the input data set, and its values must be valid SAS/GRAPH color names (see *SAS/GRAPH: Reference* for complete details). A list of CBLOCKVAR= variables must be enclosed in parentheses.

The procedure matches the CBLOCKVAR= variables with block variables in the order specified. That is, each block legend is filled with the color value of the CBLOCKVAR= variable of the first observation in each block. In general, values of the *i*th CBLOCKVAR= variable are used to fill the block of the legend corresponding to the *i*th block variable.

By default, fill colors are not used for the block variable legend. The CBLOCKVAR= option is available only when block variables are used in the PLOT statement.

CBOXES=*color* **| (***variable***)**
specifies the colors for the outlines of the box-and-whiskers plots created with the PLOT statement. You can use one of the following approaches:

You can specify CBOXES=*color* to provide a single outline color for all the box-and-whiskers plots.

You can specify CBOXES=(*variable*) to provide a distinct outline color for each box-and-whiskers plot as the value of the variable. The variable must be a character variable of up to eight characters in the input data set, and its values must be valid SAS/GRAPH color names (see *SAS/GRAPH: Reference* for complete details). The outline color of the plot displayed for a particular group is the value of the variable in the observations corresponding to this group. Note that if there are multiple observations per group in the input data set, the values of the variable should be identical for all the observations in a given group.

CBOXFILL=*color* **| (***variable***)**
specifies the interior fill colors for the box-and-whiskers plots. You can use one of the following approaches:

You can specify CBOXFILL=*color* to provide a single color for all of the box-and-whiskers plots.

You can specify CBOXFILL=(*variable*) to provide a distinct color for each box-and-whiskers plot as the value of the variable. The variable must be a character variable of up to eight characters in the input data set, and its values must be valid SAS/GRAPH color names (or the value EMPTY, which you can use to suppress color filling). See *SAS/GRAPH: Reference* for complete details. The interior color of the box displayed for a particular group is the value of the variable in the observations corresponding to this group. Note that if there are multiple observations per group in the input data set, the values of the variable should be identical for all the observations in a given group.

By default, the interiors are not filled.

CCLIP=*color*
specifies a color for the plotting symbol that is specified with the CLIPSYMBOL= option to mark clipped values. The default color is the color specified in the COLOR= option in the SYMBOL1 statement.

CCONNECT=*color*
specifies the color for line segments connecting points on the plot. The default color is the color specified in the COLOR= option in the SYMBOL1 statement. This option is not applicable unless you also specify the BOXCONNECT= option.

CCOVERLAY=(*color-list***)**
specifies the colors for line segments connecting points on overlay plots. Colors in the CCOVERLAY= list are matched with variables in the corresponding positions in the OVERLAY= list. By default, points are connected by line segments of the same color as the plotted points. You can specify the value NONE to suppress the line segments connecting points of an overlay plot.

CFRAME=*color*
specifies the color for filling the rectangle enclosed by the axes and the frame. By default, this area is not filled. The CFRAME= option cannot be used in conjunction with the NOFRAME option.

CGRID=*color*
specifies the color for the grid requested by the ENDGRID or GRID option. By default, the grid is the same color as the axes.

CHREF=*color*
specifies the color for the lines requested by the HREF= option.

CLABEL=*color*
specifies the color for labels produced by the ALLLABEL= option. The default color is the CTEXT= color.

† **CLIPFACTOR=***factor*
requests clipping of extreme values on the box plot. The *factor* that you specify determines the extent to which these values are clipped, and it must be greater than 1.

For examples of the CLIPFACTOR= option, see Figure 28.17 and Figure 28.18. Related clipping options are CCLIP=, CLIPLEGEND=, CLIPLEGPOS=, CLIPSUBCHAR=, and CLIPSYMBOL=.

† CLIPLEGEND='*label***'**
specifies the *label* for the legend that indicates the number of clipped boxes when the CLIPFACTOR= option is used. The *label* must be no more than 16 characters and must be enclosed in quotes. For an example, see Figure 28.18.

CLIPLEGPOS= TOP | BOTTOM
specifies the position for the legend that indicates the number of clipped boxes when the CLIPFACTOR= option is used. The keyword TOP or BOTTOM positions the legend at the top or bottom of the chart, respectively. Do not specify CLIPLEGPOS=TOP together with the BLOCKPOS=1 or BLOCKPOS=2 option. By default, CLIPLEGPOS=BOTTOM.

† CLIPSUBCHAR='*character***'**
specifies a substitution character (such as '#') for the label provided with the CLIPLEGEND= option. The substitution character is replaced with the number of boxes that are clipped. For example, suppose that the following statements produce a chart in which three boxes are clipped:

```
proc boxplot data=Pistons;
   plot Diameter*Hour /
      clipfactor  = 1.5
      cliplegend  = 'Boxes clipped=#'
      clipsubchar = '#' ;
run;
```

Then the clipping legend displayed on the chart will be "Boxes clipped=3".

CLIPSYMBOL=*symbol*
specifies a plot symbol used to identify clipped points on the chart and in the legend when the CLIPFACTOR= option is used. You should use this option in conjunction with the CLIPFACTOR= option. The default *symbol* is CLIPSYMBOL=SQUARE.

CLIPSYMBOLHT=*value*
specifies the height for the symbol marker used to identify clipped points on the chart when the CLIPFACTOR= option is used. The default is the height specified with the H= option in the SYMBOL statement.

For general information about clipping options, see the section "Clipping Extreme Values" on page 1179.

CONTINUOUS
specifies that numeric group variable values be treated as continuous values. By default, the values of a numeric group variable are considered discrete values unless the HAXIS= option is specified. **NOTE:** The CONTINUOUS option is not supported for ODS Graphics output. For more information, see the discussion in the section "Continuous Group Variables" on page 1170.

COVERLAY=(*color-list***)**
specifies the colors used to plot overlay variables. Colors in the COVERLAY= list are matched with variables in the corresponding positions in the OVERLAY= list.

COVERLAYCLIP=*color*
specifies the color used to plot clipped values on overlay plots when the CLIPFACTOR= option is used.

CTEXT=*color*
specifies the color for tick mark values and axis labels. The default color is the color specified in the CTEXT= option in the most recent GOPTIONS statement.

CVREF=*color*
specifies the color for the lines requested by the VREF= option.

DESCRIPTION='*string***'**

DES='*string***'**
specifies a description of a box plot produced with traditional graphics. The description appears in the PROC GREPLAY master menu and can be no longer than 256 characters. The default description is the analysis variable name.

ENDGRID
adds a grid to the rightmost portion of the plot, beginning with the first labeled major tick mark position that follows the last box-and-whiskers plot. You can use the HAXIS= option to force space to be added to the horizontal axis.

FONT=*font*
specifies a font for labels and legends. You can also specify fonts for axis labels in an AXIS statement. The FONT= font takes precedence over the FTEXT= font specified in the GOPTIONS statement. See *SAS/GRAPH: Reference* for more information about the GOPTIONS statement.

FRONTREF
draws reference lines specified with the HREF= and VREF= options in front of box-and-whiskers plots. By default, reference lines are drawn behind the box-and-whiskers plots and can be obscured by filled boxes.

† GRID
adds a grid to the box plot. Grid lines are horizontal lines positioned at labeled major tick marks, and they cover the length and height of the plotting area.

HAXIS=*value-list* **| AXIS***n*
specifies tick mark values for the horizontal (group) axis. If the group variable is numeric, the values must be numeric and equally spaced. If the group variable is character, values must be quoted strings of up to 16 characters. Optionally, you can specify an axis name defined in a previous AXIS statement. See *SAS/GRAPH: Reference* for more information about the AXIS statement.

If you are producing traditional graphics, specifying the HAXIS= option with a numeric group variable causes the group variable values to be treated as continuous values. For more information, see the description of the CONTINUOUS option and the discussion in the section "Continuous Group Variables" on page 1170. Numeric values can be given in an explicit or implicit list. If a date, time, or datetime format is associated with a numeric group variable, SAS datetime literals can be used. Examples of HAXIS= lists follow:

```
haxis=0 2 4 6 8 10
haxis=0 to 10 by 2
```

haxis='LT12A' 'LT12B' 'LT12C' 'LT15A' 'LT15B' 'LT15C'

haxis='20MAY88'D to '20AUG88'D by 7

haxis='01JAN88'D to '31DEC88'D by 30

If the group variable is numeric, the HAXIS= list must span the group variable values. If the group variable is character, the HAXIS= list must include all of the group variable values. You can add group positions to the box plot by specifying HAXIS= values that are not group variable values.

If you specify a large number of HAXIS= values, some of these can be thinned to avoid collisions between tick mark labels. To avoid thinning, use one of the following methods.

Shorten values of the group variable by eliminating redundant characters. For example, if your group variable has values LOT1, LOT2, LOT3, and so on, you can use the SUBSTR function in a DATA step to eliminate LOT from each value, and you can modify the horizontal axis label to indicate that the values refer to lots.

Use the TURNHLABELS option to turn the labels vertically.

Use the NPANELPOS= option to force fewer group positions per panel.

HEIGHT=*value*

specifies the height (in vertical screen percent units) of the text for axis labels and legends. This value takes precedence over the HTEXT= value specified in the GOPTIONS statement. This option is recommended for use with fonts specified with the FONT= option or with the FTEXT= option in the GOPTIONS statement. See *SAS/GRAPH: Reference* for complete information about the GOPTIONS statement.

HMINOR=*n*

HM=*n*

specifies the number of minor tick marks between major tick marks on the horizontal axis. Minor tick marks are not labeled. The default is HMINOR=0.

HOFFSET=*value*

specifies the length (in percent screen units) of the offset at both ends of the horizontal axis. You can eliminate the offset by specifying HOFFSET=0.

† HORIZONTAL

produces a horizontal box plot, with group variable values on the vertical axis and analysis variable values on the horizontal axis. The HORIZONTAL option is supported only with ODS Graphics. **NOTE:** When you specify the HORIZONTAL option, any INSETGROUP statements associated with the PLOT statement are ignored.

† HREF=*value-list*

HREF=*SAS-data-set*

draws reference lines perpendicular to the horizontal (group) axis on the box plot. You can use this option in the following ways:

You can specify the values for the lines with an HREF= list. If the group variable is numeric, the values must be numeric. If the group variable is character, the values must be quoted strings of up to 16 characters. If the group variable is formatted, the values must be given as internal values. Examples of HREF= values follow:

```
href=5
href=5 10 15 20 25 30
href='Shift 1' 'Shift 2' 'Shift 3'
```

You can specify reference line values as the values of a variable named _REF_ in an HREF= data set. The type and length of _REF_ must match those of the group variable specified in the PLOT statement. Optionally, you can provide labels for the lines as values of a variable named _REFLAB_, which must be a character variable of up to 16 characters. If you want distinct reference lines to be displayed in plots for different analysis variables specified in the PLOT statement, you must include a character variable named _VAR_, whose values are the analysis variable names. If you do not include the variable _VAR_, all of the lines are displayed in all of the plots. Each observation in an HREF= data set corresponds to a reference line. If BY variables are used in the input data set, the same BY variable structure must be used in the reference line data set unless you specify the NOBYREF option.

Unless the CONTINUOUS or HAXIS= option is specified, numeric group variable values are treated as discrete values, and only HREF= values matching these discrete values are valid. Other values are ignored.

† HREFLABELS='*label1***' ... '***labeln***'**

† HREFLABEL='*label1***' ... '***labeln***'**

† HREFLAB='*label1***' ... '***labeln***'**

specifies labels for the reference lines requested by the HREF= option. The number of labels must equal the number of lines. Enclose each label in quotes. Labels can be up to 16 characters.

HREFLABPOS=*n*

specifies the vertical position of the HREFLABELS= label, as described in the following table. By default, n=2.

HREFLABPOS=	Label Position
1	along top of plot area
2	staggered from top to bottom of plot area
3	along bottom of plot area
4	staggered from bottom to top of plot area

HTML=*variable*

specifies uniform resource locators (URLs) as values of the specified character variable (or formatted values of a numeric variable). These URLs are associated with box-and-whiskers plots when graphics output is directed into HTML. The value of the HTML= variable should be the same for each observation with a given value of the group variable.

IDCOLOR=*color*

specifies the color of the symbol marker used to identify outliers in schematic box-and-whiskers plots (that is, when you specify the keyword SCHEMATIC, SCHEMATICID, or SCHEMATICIDFAR with the BOXSTYLE= option). The default color is the color specified with the CBOXES= option.

IDCTEXT=*color*
specifies the color for the text used to label outliers when you specify the keyword SCHEMATICID or SCHEMATICIDFAR with the BOXSTYLE= option. The default value is the color specified with the CTEXT= option.

IDFONT=*font*
specifies the font for the text used to label outliers when you specify the keyword SCHEMATICID or SCHEMATICIDFAR with the BOXSTYLE= option. The default font is SIMPLEX.

IDHEIGHT=*value*
specifies the height for the text used to label outliers when you specify the keyword SCHEMATICID or SCHEMATICIDFAR with the BOXSTYLE= option. The default value is the height specified with the HTEXT= option in the GOPTIONS statement. See *SAS/GRAPH: Reference* for complete information about the GOPTIONS statement.

IDSYMBOL=*symbol*
specifies the symbol marker used to identify outliers in schematic box plots. The default symbol is SQUARE.

IDSYMBOLHEIGHT=*value*
specifies the height of the symbol marker used to identify outliers in schematic box plots.

INTERVAL=DAY | DTDAY | HOUR | MINUTE | MONTH | QTR | SECOND
specifies the natural time interval between consecutive group positions when a time, date, or datetime format is associated with a numeric group variable. By default, the INTERVAL= option uses the number of group positions per panel (screen or page) that you specify with the NPANELPOS= option. The default time interval *keywords* for various time formats are shown in the following table.

Format	Default Keyword	Format	Default Keyword
DATE	**DAY**	MONYY	**MONTH**
DATETIME	**DTDAY**	TIME	**SECOND**
DDMMYY	**DAY**	TOD	**SECOND**
HHMM	**HOUR**	WEEKDATE	**DAY**
HOUR	**HOUR**	WORDDATE	**DAY**
MMDDYY	**DAY**	YYMMDD	**DAY**
MMSS	**MINUTE**	YYQ	**QTR**

You can use the INTERVAL= option to modify the effect of the NPANELPOS= option, which specifies the number of group positions per panel. The INTERVAL= option enables you to match the scale of the horizontal axis to the scale of the group variable without having to associate a different format with the group variable.

For example, suppose that your formatted group values span an overall time interval of 100 days and a DATETIME format is associated with the group variable. Since the default interval for the DATETIME format is DTDAY and since NPANELPOS=25 by default, the plot is displayed with four panels.

Now, suppose that your data span an overall time interval of 100 hours and a DATETIME format is associated with the group variable. The plot for these data is created in a single panel, but the data occupy only a small fraction of the plot since the scale of the data (hours) does not match that of the

horizontal axis (days). If you specify INTERVAL=HOUR, the horizontal axis is scaled for 25 hours, matching the scale of the data, and the plot is displayed with four panels.

You should use the INTERVAL= option only in conjunction with the CONTINUOUS or HAXIS= option, which produces a horizontal axis of continuous group variable values. For more information, see the descriptions of the CONTINUOUS and HAXIS= options, and the discussion in the section "Continuous Group Variables" on page 1170.

INTSTART=*value*

specifies the starting value for a numeric horizontal axis when a date, time, or datetime format is associated with the group variable. If the value specified is greater than the first group variable value, this option has no effect.

LABELANGLE=*angle*

specifies the angle at which labels requested with the ALLLABEL= option are drawn. A positive angle rotates the labels counterclockwise; a negative angle rotates them clockwise. By default, labels are oriented horizontally.

LBOXES=*linetype* **| (***variable***)**

specifies the line types for the outlines of the box-and-whiskers plots. You can use one of the following approaches:

You can specify LBOXES=*linetype* to provide a single linetype for all of the box-and-whiskers plots.

You can specify LBOXES=(*variable*) to provide a distinct line type for each box-and-whiskers plot. The variable must be a numeric variable in the input data set, and its values must be valid SAS/GRAPH linetype values (numbers ranging from 1 to 46). The line type for the plot displayed for a particular group is the value of the variable in the observations corresponding to this group. Note that if there are multiple observations per group in the input data set, the values of the variable should be identical for all of the observations in a given group.

The default value is 1, which produces solid lines. See the description of the SYMBOL statement in *SAS/GRAPH: Reference* for more information about valid linetypes.

LENDGRID=*linetype*

specifies the line type for the grid requested with the ENDGRID option. The default value is 1, which produces a solid line. If you use the LENDGRID= option, you do not need to specify the ENDGRID option. See the description of the SYMBOL statement in *SAS/GRAPH: Reference* for more information about valid linetypes.

LGRID=*linetype*

specifies the line type for the grid requested with the GRID option. The default value is 1, which produces a solid line. If you use the LGRID= option, you do not need to specify the GRID option. See the description of the SYMBOL statement in *SAS/GRAPH: Reference* for more information about valid linetypes.

LHREF=*linetype*

LH=*linetype*

specifies the line type for reference lines requested with the HREF= option. The default value is 2, which produces a dashed line. See the description of the SYMBOL statement in *SAS/GRAPH: Reference* for more information about valid linetypes.

LOVERLAY=(*linetypes***)**
specifies line types for the line segments connecting points on overlay plots. Line types in the LOVERLAY= list are matched with variables in the corresponding positions in the OVERLAY= list.

LVREF=*linetype*

LV=*linetype*
specifies the line type for reference lines requested by the VREF= option. The default value is 2, which produces a dashed line. See the description of the SYMBOL statement in *SAS/GRAPH: Reference* for more information about valid linetypes.

† MAXPANELS=*n*
specifies the maximum number of panels used to display a box plot. By default, *n* = 20.

† MISSBREAK
determines how groups are formed when observations are read from a DATA= data set and a character group variable is provided. When you specify the MISSBREAK option, observations with missing values of the group variable are not processed. Furthermore, the next observation with a nonmissing value of the group variable is treated as the beginning observation of a new group even if this value is identical to the most recent nonmissing group value. In other words, by specifying the option MISSBREAK and by inserting an observation with a missing group variable value into a group of consecutive observations with the same group variable value, you can split the group into two distinct groups of observations.

By default (that is, when you omit the MISSBREAK option), observations with missing values of the group variable are not processed, and all remaining observations with the same consecutive value of the group variable are treated as a single group.

NAME='*string***'**
specifies a name, not more than eight characters long, for a traditional graphics box plot. The name appears in the PROC GREPLAY master menu.

NLEGEND
requests a legend displaying group sizes. If the size is the same for each group, that number is displayed. Otherwise, the minimum and maximum group sizes are displayed.

† NOBYREF
specifies that the reference line information in an HREF= or VREF= data set be applied uniformly to box plots created for all the BY groups in the input data set. If you specify the NOBYREF option, you do not need to provide BY variables in the reference line data set. By default, you must provide BY variables.

† NOCHART
suppresses the creation of the box plot. You typically specify the NOCHART option when you are using the procedure to compute group summary statistics and save them in an output data set.

NOFRAME
suppresses the default frame drawn around the plot.

† NOHLABEL
suppresses the label for the horizontal (group) axis. Use the NOHLABEL option when the meaning of the axis is evident from the tick mark labels, such as when a date format is associated with the group variable.

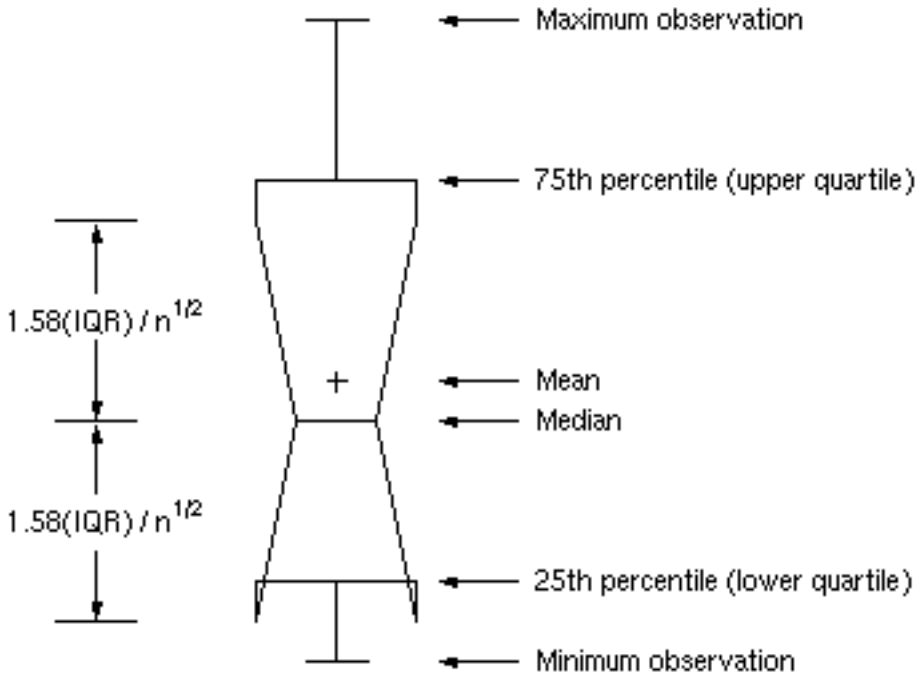
† NOOVERLAYLEGEND
suppresses the legend for overlay plots that is displayed by default when the OVERLAY= option is specified.

† NOSERIFS
eliminates serifs from the whiskers of box-and-whiskers plots.

† NOTCHES
specifies that box-and-whiskers plots be notched. The endpoints of the notches are located at the median plus and minus 1:58.IQR=p n/, where IQR is the interquartile range and *n* is the group size. The medians (central lines) of two box-and-whiskers plots are significantly different at approximately the 0.95 confidence level if the corresponding notches do not overlap.

See McGill, Tukey, and Larsen (1978) for more information. Figure 28.7 illustrates the NOTCHES option. Notice the folding effect at the bottom, which happens when the endpoint of a notch is beyond its corresponding quartile. This situation typically occurs when the group size is small.

Figure 28.7 Box Plot: The NOTCHES Option



NOTICKREP
applies to character-valued group variables and specifies that only the first occurrence of repeated, adjacent group values be labeled on the horizontal axis.

NOVANGLE
requests that the vertical axis label be strung out vertically.

† NPANELPOS=*n*

NPANEL=*n*
specifies the number of group positions per panel. You typically specify the NPANELPOS= option to display more box-and-whiskers plots on a panel than the default number, which is *n* = 25.

You can specify a positive or negative number for *n*. The absolute value of *n* must be at least 5. If *n* is positive, the number of positions is adjusted so that it is approximately equal to *n* and so that all panels display approximately the same number of group positions. If *n* is negative, no balancing is done, and each panel (except possibly the last) displays approximately jnj positions. In this case, the approximation is due only to axis scaling.

You can use the INTERVAL= option to change the effect of the NPANELPOS= option when a date or time format is associated with the group variable. The INTERVAL= option enables you to match the scale of the horizontal axis to the scale of the group variable without having to associate a different format with the group variable.

† ODSFOOTNOTE=FOOTNOTE | FOOTNOTE1 | '*string***'**
adds a footnote to ODS Graphics output. If you specify the FOOTNOTE (or FOOTNOTE1) keyword, the value of SAS FOOTNOTE statement is used the as the graph footnote. If you specify a quoted string, that is used as the footnote. The quoted string can contain any of the following escaped characters, which are replaced with the appropriate values from the analysis:

nn	analysis variable name
nl	analysis variable label (or name if the analysis variable has no label)
nx	group variable name
ns	group variable label (or name if the group variable has no label)

† ODSFOOTNOTE2=FOOTNOTE2 | '*string***'**
adds a secondary footnote to ODS Graphics output. If you specify the FOOTNOTE2 keyword, the value of SAS FOOTNOTE2 statement is used as the secondary graph footnote. If you specify a quoted string, that is used as the secondary footnote. The quoted string can contain any of the following escaped characters, which are replaced with the appropriate values from the analysis:

nn	analysis variable name
nl	analysis variable label (or name if the analysis variable has no label)
nx	group variable name
ns	group variable label (or name if the group variable has no label)

† ODSTITLE=TITLE | TITLE1 | NONE | DEFAULT | LABELFMT | '*string***'**
specifies a title for ODS Graphics output.

TITLE (or TITLE1)	uses the value of SAS TITLE statement as the graph title.
NONE	suppresses all titles from the graph.

DEFAULT uses the default ODS Graphics title (a descriptive title consisting of the plot type and the process variable name.)

LABELFMT uses the default ODS Graphics title with the variable label instead of the variable name.

If you specify a quoted string, that is used as the graph title. The quoted string can contain any of the following escaped characters, which are replaced with the appropriate values from the analysis:

nn analysis variable name

nl analysis variable label (or name if the analysis variable has no label)

nx group variable name

ns group variable label (or name if the group variable has no label)

† ODSTITLE2=TITLE2 | '*string***'**

specifies a secondary title for ODS Graphics output. If you specify the TITLE2 keyword, the value of SAS TITLE2 statement is used as the secondary graph title. If you specify a quoted string, that is used as the secondary title. The quoted string can contain any of the following escaped characters, which are replaced with the appropriate values from the analysis:

nn analysis variable name

nl analysis variable label (or name if the analysis variable has no label)

nx group variable name

ns group variable label (or name if the group variable has no label)

† OUTBOX=*SAS-data-set*

creates an output data set that contains group summary statistics and outlier values for a box plot. You can use an OUTBOX= data set as a BOX= input data set in a subsequent run of the procedure. See the section "OUTBOX= Data Set" on page 1163 for details.

OUTHIGHHTML=*variable*

specifies a variable whose values are URLs to be associated with outlier points above the upper fence on a schematic box plot when graphics output is directed into HTML.

† OUTHISTORY=*SAS-data-set*

creates an output data set that contains the group summary statistics. You can use an OUTHISTORY= data set as a HISTORY= input data set in a subsequent run of the procedure. See the section "OUTHISTORY= Data Set" on page 1164 for details.

OUTLOWHTML=*variable*

specifies a variable whose values are URLs to be associated with outlier points below the lower fence on a schematic box plot when graphics output is directed into HTML.

† OVERLAY=(*variable-list***)**

specifies variables to be plotted as overlays on the box plot. One value for each overlay variable is plotted at each group position. If there are multiple observations with the same group variable value in the input data set, the overlay variable values from the first observation in each group are plotted. By default, the points in an overlay plot are connected with line segments.

OVERLAYCLIPSYM=*symbol*
specifies the symbol used to plot clipped values on overlay plots when the CLIPFACTOR= option is used.

OVERLAYCLIPSYMHT=*value*
specifies the height for the symbol used to plot clipped values on overlay plots when the CLIPFACTOR= option is used.

OVERLAYHTML=(*variable-list***)**
specifies variables whose values are URLs to be associated with points on overlay plots when graphics output is directed into HTML. Variables in the OVERLAYHTML= list are matched with variables in the corresponding positions in the OVERLAY= list.

OVERLAYID=(*variable-list***)**
specifies variables whose formatted values are used to label points on overlays. Variables in the OVERLAYID= list are matched with variables in the corresponding positions in the OVERLAY= list. The value of the OVERLAYID= variable should be the same for each observation with a given value of the group variable.

† OVERLAYLEGLAB='*label***'**
specifies the label displayed to the left of the overlay legend produced by the OVERLAY= option. The label can be up to 16 characters and must be enclosed in quotes. The default label is "Overlays:".

OVERLAYSYM=(*symbol-list***)**
specifies symbols used to plot overlay variables. Symbols in the OVERLAYSYM= list are matched with variables in the corresponding positions in the OVERLAY= list.

OVERLAYSYMHT=(*value-list***)**
specifies the heights of symbols used to plot overlay variables. Symbol heights in the OVERLAYSYMHT= list are matched with variables in the corresponding positions in the OVERLAY= list.

PAGENUM='*string***'**
specifies the form of the label used for pagination. The string can be up to 16 characters, and it must include one or two occurrences of the substitution character '#'. The first '#' is replaced with the page number, and the optional second '#' is replaced with the total number of pages.

The PAGENUM= option is useful when you are working with a large number of groups, resulting in multiple pages of output. For example, suppose that each of the following PLOT statements produces multiple pages:

```
proc boxplot data=Pistons;
   plot Diameter*Hour / pagenum='Page #';
   plot Diameter*Hour / pagenum='Page # of #';
   plot Diameter*Hour / pagenum='#/#';
run;
```

The third page produced by the first statement would be labeled "Page 3". The third page produced by the second statement would be labeled "Page 3 of 5". The third page produced by the third statement would be labeled "3/5".

By default, no page number is displayed.

PAGENUMPOS=TL | TR | BL | BR | TL100 | TR100 | BL0 | BR0
specifies where to position the page number requested with the PAGENUM= option. The keywords TL, TR, BL, and BR correspond to the positions top left, top right, bottom left, and bottom right, respectively. You can use the TL100 and TR100 keywords to ensure that the page number appears at the very top of a page when a title is displayed. The BL0 and BR0 keywords ensure that the page number appears at the very bottom of a page when footnotes are displayed.

The default value is BR.

† **PCTLDEF=***index*
specifies one of five definitions used to calculate percentiles in the construction of box-and-whiskers plots. The index can be 1, 2, 3, 4, or 5. The five corresponding percentile definitions are discussed in the section "Percentile Definitions" on page 1169. The default index is 5.

† **REPEAT**

† **REP**
specifies that the horizontal axis of a plot that spans multiple panels be arranged so that the last group position on a panel is repeated as the first group position on the next panel. The REPEAT option facilitates cutting and pasting panels together. When a SAS DATETIME format is associated with the group variable, the REPEAT option is the default.

SKIPHLABELS=*n*

SKIPHLABEL=*n*
specifies the number *n* of consecutive tick mark labels, beginning with the second tick mark label, that are thinned (not displayed) on the horizontal (group) axis. For example, specifying SKIPHLABEL=1 causes every other label to be skipped. Specifying SKIPHLABEL=2 causes the second and third labels to be skipped, the fifth and sixth labels to be skipped, and so forth.

The default value of the SKIPHLABELS= option is the smallest value *n* for which tick mark labels do not collide. A specified *n* will be overridden to avoid collision. To reduce thinning, you can use the TURNHLABELS option.

SYMBOLLEGEND=LEGEND*n* **| NONE**
controls the legend for the levels of a symbol variable (see Example 28.2). You can specify SYMBOLLEGEND=LEGEND*n*, where *n* is the number of a LEGEND statement defined previously. You can specify SYMBOLLEGEND=NONE to suppress the default legend. See *SAS/GRAPH: Reference* for more information about the LEGEND statement.

SYMBOLORDER=DATA | INTERNAL | FORMATTED

SYMORD=DATA | INTERNAL | FORMATTED
specifies the order in which symbols are assigned for levels of the symbol variable. The DATA keyword assigns symbols to values in the order in which values appear in the input data set. The INTERNAL keyword assigns symbols based on sorted order of internal values of the symbol variable, and the FORMATTED keyword assigns them based on sorted formatted values. The default value is FORMATTED.

† **TOTPANELS=***n*
specifies the total number of panels to be used to display the plot. This option overrides the NPANELPOS= option.

TURNHLABELS

TURNHLABEL

turns the major tick mark labels for the horizontal (group) axis so that they are arranged vertically. By default, labels are arranged horizontally.

Note that arranging the labels vertically might leave insufficient vertical space on the panel for a plot.

† **VAXIS=***value-list*

VAXIS=AXIS*n*

specifies major tick mark values for the vertical axis of a box plot. The values must be listed in increasing order, must be evenly spaced, and must span the range of values displayed in the plot. You can specify the values with an explicit list or with an implicit list, as shown in the following example:

```
proc boxplot;
   plot Width*Hour / vaxis=0 2 4 6 8;
   plot Width*Hour / vaxis=0 to 8 by 2;
run;
```

You can also specify a previously defined AXIS statement with the VAXIS= option.

† **VFORMAT=***format*

specifies a format to be used for displaying tick mark labels on the vertical axis of the box plot.

VMINOR=*n*

VM=*n*

specifies the number of minor tick marks between major tick marks on the vertical axis. Minor tick marks are not labeled. By default, VMINOR=0.

VOFFSET=*value*

specifies the length in percent screen units of the offset at the ends of the vertical axis.

† **VREF=***value-list* **|** *SAS-data-set*

draws reference lines perpendicular to the vertical axis. You can use this option in the following ways:

Specify the values for the lines with a VREF= list:

```
vref=20
vref=20 40 80
```

Specify the values for the lines as the values of a numeric variable named _REF_ in a VREF= data set. Optionally, you can provide labels for the lines as values of a variable named _REFLAB_, which must be a character variable of up to 16 characters. If you want distinct reference lines to be displayed in plots for different analysis variables specified in the PLOT statement, you must include a character variable named _VAR_, whose values are the names of the analysis variables. If you do not include the variable _VAR_, all of the lines are displayed in all of the plots. Each observation in the VREF= data set corresponds to a reference line. If BY variables are used in the input data set, the same BY-variable structure must be used in the VREF= data set unless you specify the NOBYREF option.

† **VREFLABELS='***label1***' . . . '***labeln***'**

† **VREFLABEL='***label1***' . . . '***labeln***'**

† **VREFLAB='***label1***' . . . '***labeln***'**

specifies labels for the reference lines requested by the VREF= option. The number of labels must equal the number of lines. Enclose each label in quotes. Labels can be up to 16 characters.

VREFLABPOS=*n*

specifies the horizontal position of the VREFLABELS= label, as described in the following table. By default, *n* = 1.

n	Label Position
1	left-justified in plot area
2	right-justified in plot area
3	left-justified in right margin

VZERO

forces the origin to be included in the vertical axis for a box plot.

WAXIS=*n*

specifies the width in pixels for the axis and frame lines. By default, *n* = 1.

WGRID=*n*

specifies the width in pixels for grid lines requested with the ENDGRID and GRID options. By default, *n* = 1.

† **WHISKERPERCENTILE=***p*

specifies that the whiskers of the box-and-whisker plots be drawn to the *p*th and (100 *p*)th percentiles. For example, if you specify WHISKERPERCENTILE=10 the whiskers are drawn to the 10th and 90th percentiles. Observations lying beyond the whiskers are outliers and there are no far outliers.

By default, whiskers are drawn to the minimum and maximum data values if the BOXSTYLE= value is SKELETAL (the default), and to the most extreme values within or equal to the lower and upper fences otherwise.

WOVERLAY=(*value-list***)**

specifies the widths in pixels for the line segments connecting points on overlay plots. Widths in the WOVERLAY= list are matched with variables in the corresponding positions in the OVERLAY= list. By default, all overlay widths are 1.

Details: BOXPLOT Procedure

Summary Statistics Represented by Box Plots

Table 28.7 lists the summary statistics represented in each box-and-whiskers plot.

Table 28.7 Summary Statistics Represented by Box Plots

Group Summary Statistic	Feature of Box-and-Whiskers Plot
maximum	endpoint of upper whisker
third quartile (75th percentile)	upper edge of box
median (50th percentile)	line inside box
mean	symbol marker
first quartile (25th percentile)	lower edge of box
minimum	endpoint of lower whisker

Note that you can request different box plot styles, as discussed in the section "Styles of Box Plots" on page 1168, and as illustrated in Example 28.2.

Output Data Sets

OUTBOX= Data Set

The OUTBOX= data set saves group summary statistics and outlier values. The following variables can be saved:

the group variable

the variable _VAR_, containing the analysis variable name

the variable _TYPE_, identifying features of box-and-whiskers plots

the variable _VALUE_, containing values of box-and-whiskers plot features

the variable _ID_, containing labels for outliers

the variable _HTML_, containing URLs associated with plot features

_ID_ is included in the OUTBOX= data set only if the keyword SCHEMATICID or SCHEMATICIDFAR is specified with the BOXSTYLE= option. _HTML_ is present only if one or more of the HTML=, OUTHIGHHTML=, and OUTLOWHTML= options are specified.

Each observation in an OUTBOX= data set records the value of a single feature of one group's box-and-whiskers plot, such as its mean. The _TYPE_ variable identifies the feature whose value is recorded in _VALUE_. Table 28.8 lists valid _TYPE_ variable values.

Table 28.8 Valid _TYPE_ Values in an OUTBOX= Data Set

TYPE	Description
N	group size
MIN	minimum group value
Q1	group first quartile
MEDIAN	group median
MEAN	group mean
Q3	group third quartile
MAX	group maximum value
STDDEV	group standard deviation
LOW	low outlier value
HIGH	high outlier value
LOWHISKR	low whisker value, if different from MIN
HIWHISKR	high whisker value, if different from MAX
FARLOW	low far outlier value
FARHIGH	high far outlier value

Additionally, the following variables, if specified, are included:

block variables

symbol variable

BY variables

ID variables

OUTHISTORY= Data Set

The OUTHISTORY= data set saves group summary statistics. The following variables are saved:

the group variable

group minimum variables named by *analysis-variable* suffixed with *L*

group first-quartile variables named by *analysis-variable* suffixed with 1

group mean variables named by *analysis-variable* suffixed with *X*

group median variables named by *analysis-variable* suffixed with *M*

group third-quartile variables named by *analysis-variable* suffixed with 3

group maximum variables named by *analysis-variable* suffixed with *H*

group standard deviation variables named by *analysis-variable* suffixed with *S*

group size variables named by *analysis-variable* suffixed with *N*

If an analysis variable name has the maximum length of 32 characters, PROC BOXPLOT forms summary statistic names from its first 16 characters, its last 15 characters, and the appropriate suffix.

Group summary variables are created for each analysis variable specified in the PLOT statement. For example, consider the following statements:

```
proc boxplot data=Steel;
   plot (Width Diameter)*Lot / outhistory=Summary;
run;
```

The data set Summary contains variables named Lot, WidthL, Width1, WidthM, WidthX, Width3, WidthH, WidthS, WidthN, DiameterL, Diameter1, DiameterM, DiameterX, Diameter3, DiameterH, DiameterS, and DiameterN.

Additionally, the following variables, if specified, are included:

BY variables

block variables

symbol variable

ID variables

Note that an OUTHISTORY= data set does not contain outlier values, and therefore cannot be used, in general, to save a schematic box plot. You can use an OUTBOX= data set to save a schematic box plot summary.

Input Data Sets

DATA= Data Set

You can read analysis variable measurements from a data set specified with the DATA= option in the PROC BOXPLOT statement. Each analysis variable specified in the PLOT statement must be a SAS variable in the data set. This variable provides measurements that are organized into groups indexed by the group variable. The group variable, specified in the PLOT statement, must also be a SAS variable in the DATA= data set. Each observation in a DATA= data set must contain a value for each analysis variable and a value for the group variable. If the *i*th group contains n_i measurements, there should be n_i consecutive observations for which the value of the group variable is the index of the *i*th group. For example, if each group contains 20 items and there are 30 groups, the DATA= data set should contain 600 observations. Other variables that can be read from a DATA= data set include the following:

block variables

symbol variable

BY variables

ID variables

BOX= Data Set

You can read group summary statistics and outlier information from a BOX= data set specified in the PROC BOXPLOT statement. This enables you to reuse OUTBOX= data sets that have been created in previous runs of the BOXPLOT procedure to reproduce schematic box plots.

A BOX= data set must contain the following variables:

the group variable

VAR, containing the analysis variable name

TYPE, identifying features of box-and-whiskers plots

VALUE, containing values of those features

Each observation in a BOX= data set records the value of a single feature of one group's box-and-whiskers plot, such as its mean. Consequently, a BOX= data set contains multiple observations per group. These must appear consecutively in the BOX= data set.

The _TYPE_ variable identifies the feature whose value is recorded in a given observation. The following table lists valid _TYPE_ variable values.

Table 28.9 Valid _TYPE_ Values in a BOX= Data Set

TYPE	Description
N	group size
MIN	group minimum value
Q1	group first quartile
MEDIAN	group median
MEAN	group mean
Q3	group third quartile
MAX	group maximum value
STDDEV	group standard deviation
LOW	low outlier value
HIGH	high outlier value
LOWHISKR	low whisker value, if different from MIN
HIWHISKR	high whisker value, if different from MAX
FARLOW	low far outlier value
FARHIGH	high far outlier value

The features identified by _TYPE_ values N, MIN, Q1, MEDIAN, MEAN, Q3, and MAX are required for each group.

Other variables that can be read from a BOX= data set include the following:

the variable _ID_, containing labels for outliers

the variable _HTML_, containing URLs to be associated with features on box plots

block variables

symbol variable

BY variables

ID variables

When you specify the keyword SCHEMATICID or SCHEMATICIDFAR with the BOXSTYLE= option, values of _ID_ are used as outlier labels. If _ID_ does not exist in the BOX= data set, the values of the first variable listed in the ID statement are used.

HISTORY= Data Set

You can read group summary statistics from a HISTORY= data set specified in the PROC BOXPLOT statement. This enables you to reuse OUTHISTORY= data sets that have been created in previous runs of the BOXPLOT procedure or to read output data sets created with SAS summarization procedures, such as PROC UNIVARIATE.

Note that a HISTORY= data set does *not* contain outlier information. Therefore, in general you cannot reproduce a schematic box plot from summary statistics saved in an OUTHISTORY= data set. To save and reproduce schematic box plots, use OUTBOX= and BOX= data sets.

A HISTORY= data set must contain the following:

the group variable

a group minimum variable for each analysis variable

a group first-quartile variable for each analysis variable

a group median variable for each analysis variable

a group mean variable for each analysis variable

a group third-quartile variable for each analysis variable

a group maximum variable for each analysis variable

a group standard deviation variable for each analysis variable

a group size variable for each analysis variable

The names of the group summary statistics variables must be the analysis variable name concatenated with the following special suffix characters.

Group Summary Statistic	Suffix Character
group minimum	L
group first quartile	1
group median	M
group mean	X
group third quartile	3
group maximum	H
group standard deviation	S
group size	N

For example, consider the following statements:

```
proc boxplot history=Summary;
   plot (Weight Yieldstrength) * Batch;
run;
```

The data set Summary must include the variables Batch, WeightL, Weight1, WeightM, WeightX, Weight3, WeightH, WeightS, WeightN, YieldstrengthL, Yieldstrength1, YieldstrengthM, YieldstrengthX, Yieldstrength3, YieldstrengthH, YieldstrengthS, and YieldstrengthN.

Note that if you specify an analysis variable whose name contains the maximum of 32 characters, the summary variable names must be formed from the first 16 characters and the last 15 characters of the analysis variable name, suffixed with the appropriate character.

These other variables can be read from a HISTORY= data set:

block variables

symbol variable

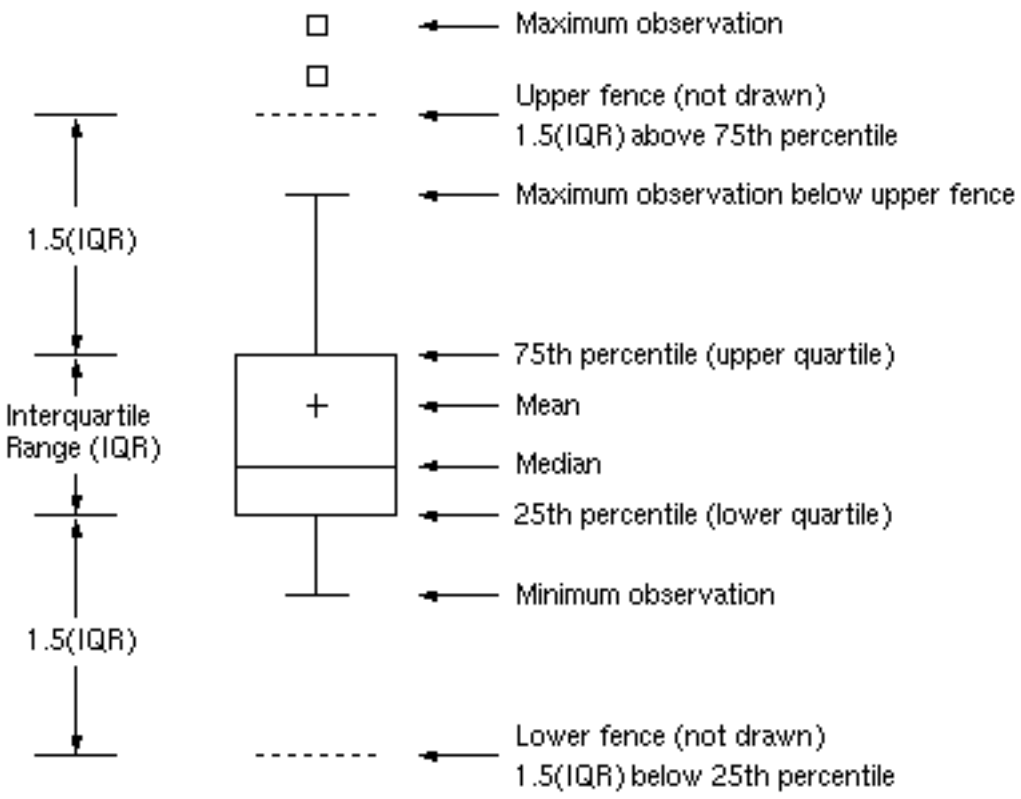
BY variables

ID variables

Styles of Box Plots

A box-and-whiskers plot is displayed for the measurements in each group on the box plot. The *skeletal* style of the box-and-whiskers plot shown in Figure 28.6 is the default. You can produce a *schematic* box plot by specifying the BOXSTYLE=SCHEMATIC option in the PLOT statement. Figure 28.8 illustrates a typical schematic box plot and the locations of the fences (which are not displayed in actual output). See the description of the BOXSTYLE= option for complete details.

Figure 28.8 Schematic Box-and-Whiskers Plot



You can draw connecting lines between adjacent box-and-whiskers plots by using the BOXCONNECT=*keyword* option. For example, BOXCONNECT=MEAN connects the points representing the means of adjacent groups. Other available keywords are MIN, Q1, MEDIAN, Q3, and MAX. Specifying BOXCONNECT without a keyword is equivalent to specifying BOXCONNECT=MEAN. You can specify the color for the connecting lines with the CCONNECT= option.

Percentile Definitions

You can use the PCTLDEF= option to specify one of five definitions for computing quantile statistics (percentiles). Suppose that n is the number of nonmissing values for a variable and that $x_1; x_2; : : : ; x_n$ represent the ordered values of the analysis variable. For the tth percentile, set $p D t=100$.

For the following definitions numbered 1, 2, 3, and 5, express np as

$$np D j C g$$

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

$$.n C 1/p D j C g$$

The *t*th percentile (call it *y*) can be 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 $np + 1/2$

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

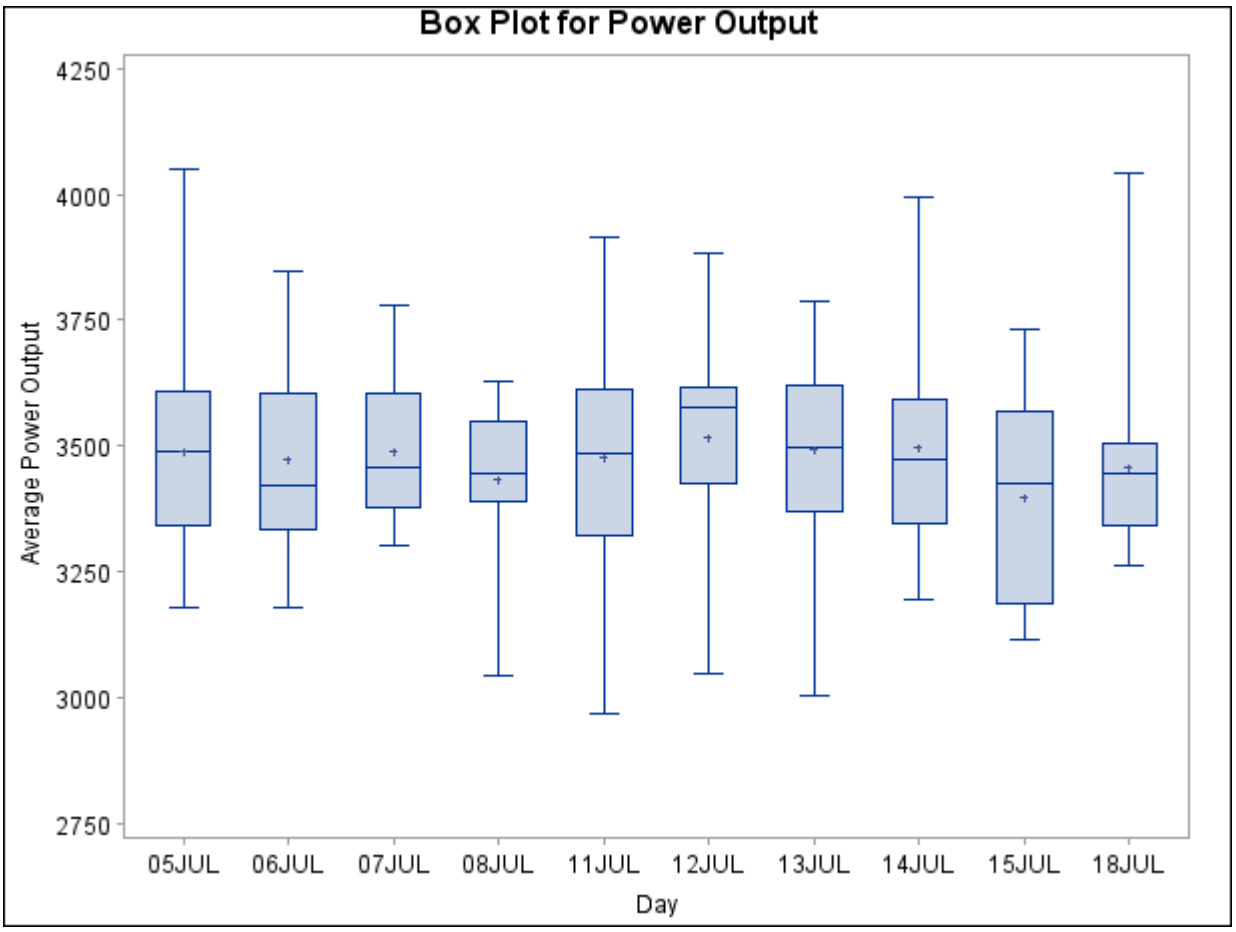
An observation read from an input data set is not analyzed if the value of the group variable is missing. For a particular analysis variable, an observation read from a DATA= data set is not analyzed if the value of the analysis variable is missing.

Continuous Group Variables

By default, the PLOT statement treats numerical group variable values as *discrete* values and spaces the boxes evenly on the plot. The following statements produce the box plot in Figure 28.9:

```
ods graphics off;
title 'Box Plot for Power Output';
proc boxplot data=Turbine;
   plot KWatts*Day;
run;
```

Figure 28.9 Box Plot with Discrete Group Variable



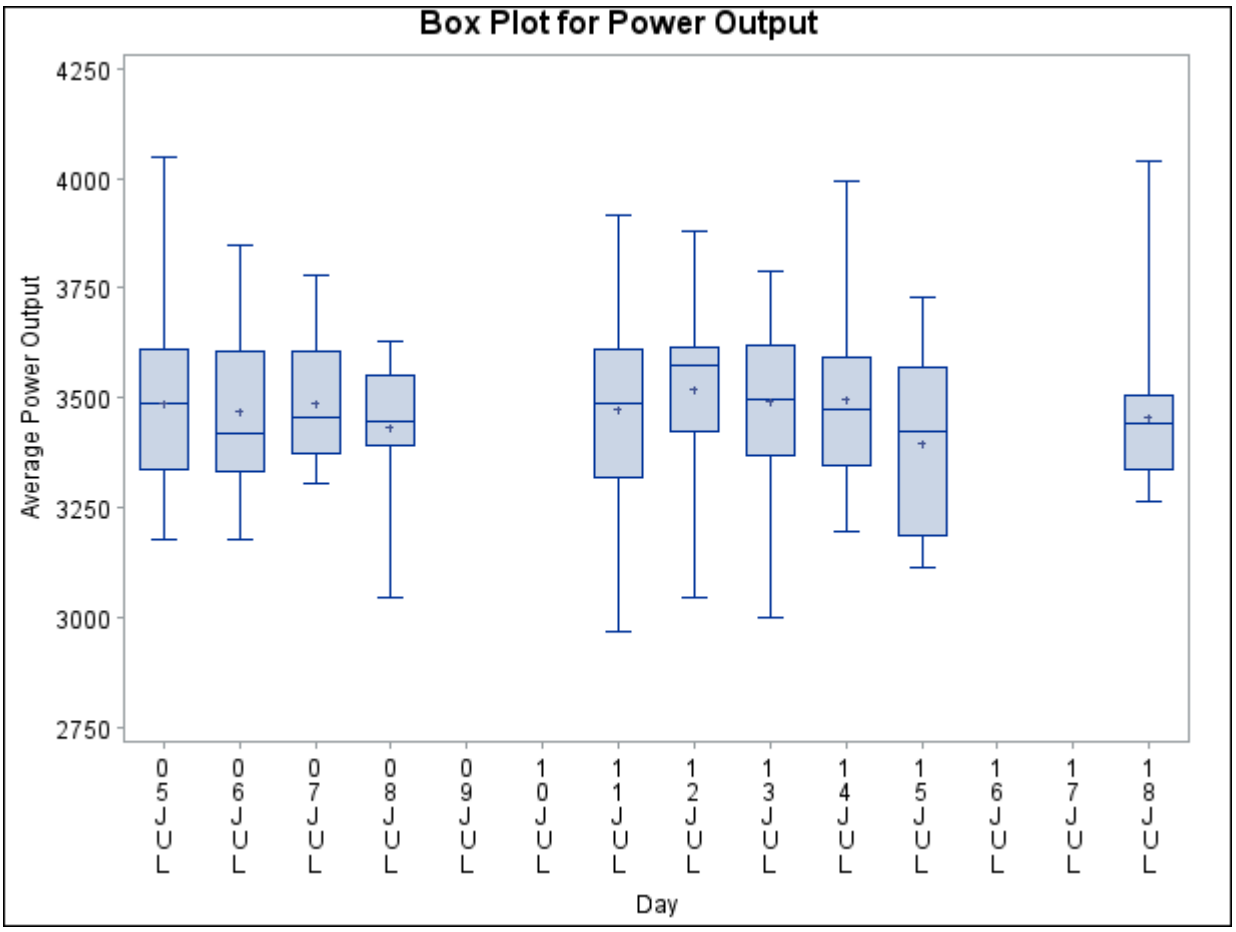
The labels on the horizontal axis in Figure 28.9 do not represent 10 consecutive days, but the box-and-whiskers plots are evenly spaced.

In order to treat the group variable as *continuous*, you can specify the CONTINUOUS or HAXIS= option when producing traditional graphics. Either option produces a box plot with a horizontal axis scaled for continuous group variable values. (ODS Graphics does not support a continuous group axis.)

The following statements produce the plot shown in Figure 28.10. The TURNHLABELS option orients the horizontal axis labels vertically so there is room to display them all.

```
title 'Box Plot for Power Output';
proc boxplot data=Turbine;
   plot KWatts*Day / turnhlabels
                     continuous;
run;
```

Figure 28.10 Box Plot with Continuous Group Variable



Note that the tick values on the horizontal axis represent consecutive days and that no box-and-whiskers plots are displayed for days when no turbine data were collected.

Positioning Insets

This section provides details on three different methods of positioning INSET boxes by using the POSITION= option. With the POSITION= option, you can specify the following:

compass points

keywords for margin positions

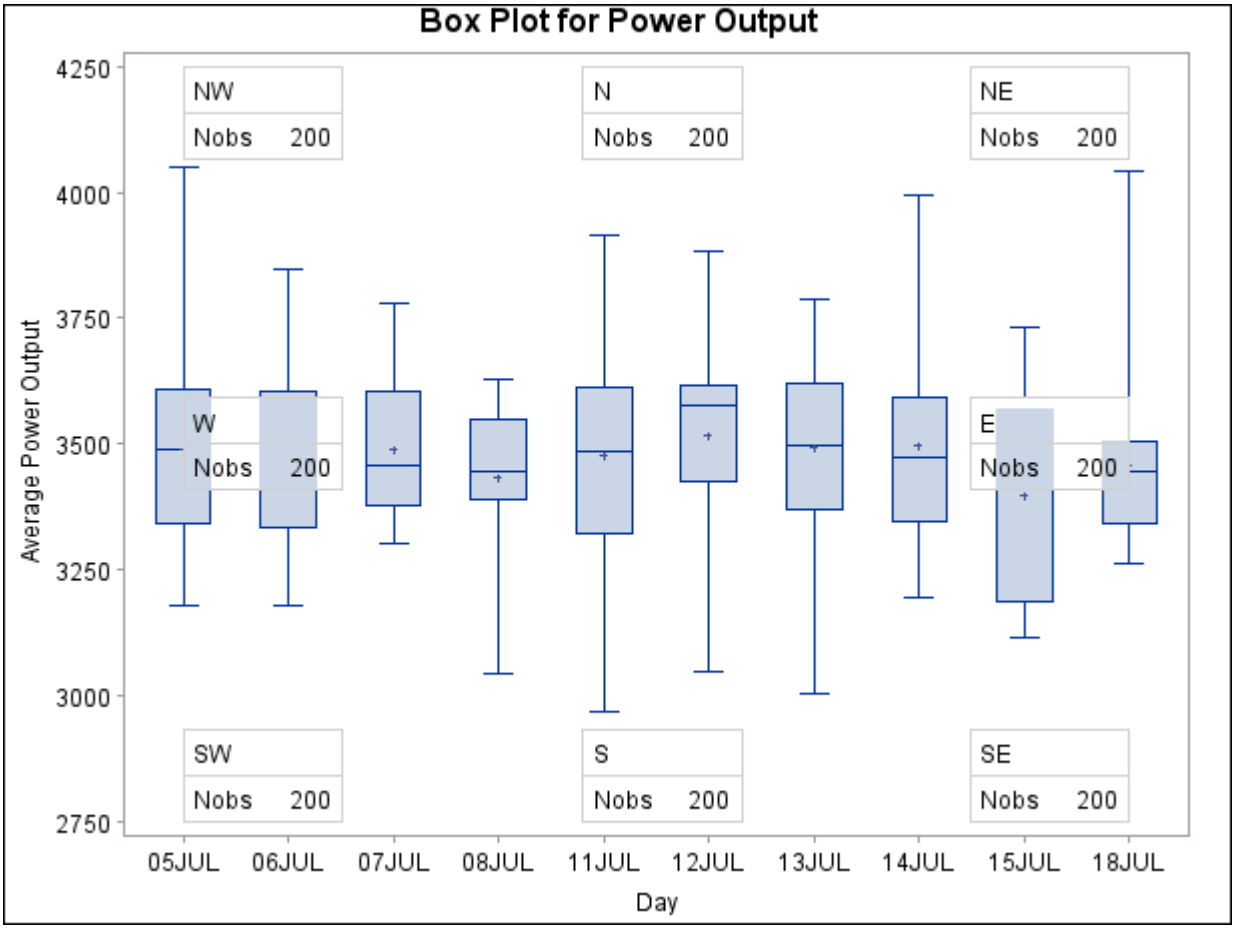
coordinates in data units or percent axis units

Positioning the Inset Using Compass Points

You can specify the eight compass points (N, NE, E, SE, S, SW, W, and NW) as keywords for the POSITION= option. The default inset position is NW. The following statements create the display in Figure 28.11, which illustrates all eight compass positions:

```
title 'Box Plot for Power Output';
proc boxplot data=Turbine;
   plot KWatts*Day;
   inset nobs / height=2.5 cfill=blank header='NW' pos=nw;
   inset nobs / height=2.5 cfill=blank header='N ' pos=n ;
   inset nobs / height=2.5 cfill=blank header='NE' pos=ne;
   inset nobs / height=2.5 cfill=blank header='E ' pos=e ;
   inset nobs / height=2.5 cfill=blank header='SE' pos=se;
   inset nobs / height=2.5 cfill=blank header='S ' pos=s ;
   inset nobs / height=2.5 cfill=blank header='SW' pos=sw;
   inset nobs / height=2.5 cfill=blank header='W ' pos=w ;
run;
```

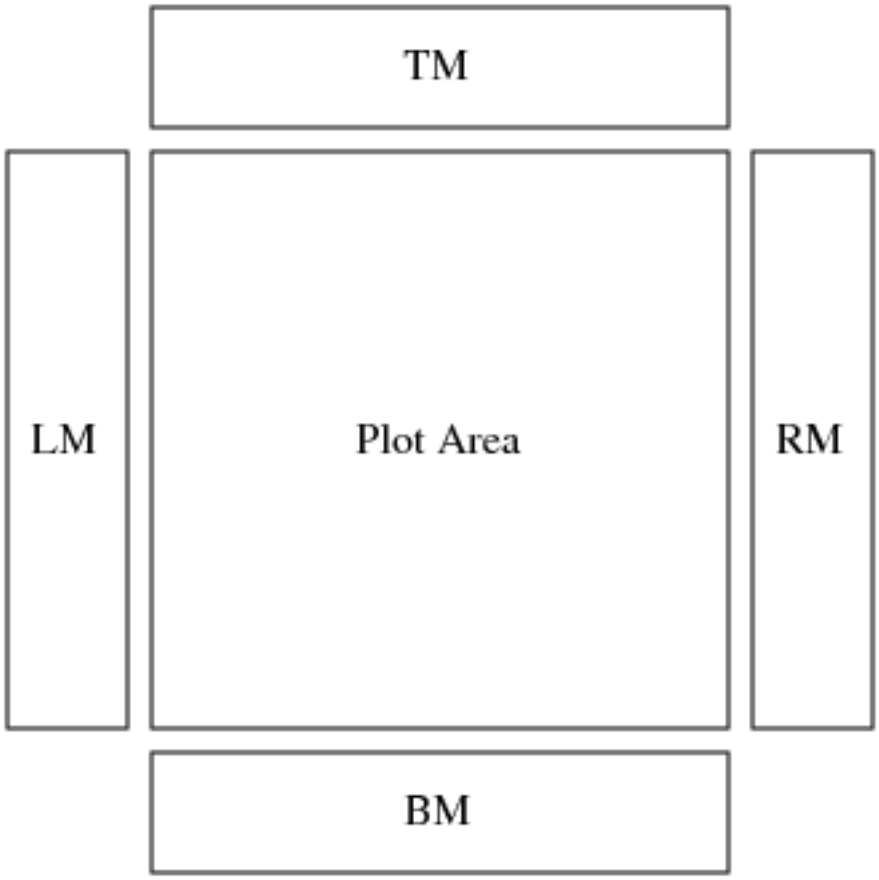
Figure 28.11 Insets Positioned Using Compass Points



Positioning the Inset in the Margins

You can also use the INSET statement to position an inset in one of the four margins surrounding the plot area by using the margin keyword LM, RM, TM, or BM, as illustrated in Figure 28.12.

Figure 28.12 Positioning Insets in the Margins



For an example of an inset placed in the top margin, see Output 28.1.1. Margin positions are recommended for insets containing a large number of statistics. If you attempt to display a lengthy inset in the interior of the plot, it is likely that the inset will collide with the data display.

Positioning the Inset Using Coordinates

You can also specify the position of an inset with coordinates by using the POSITIOND .x; y/ option. You can specify coordinates in axis percent units (the default) or in axis data units.

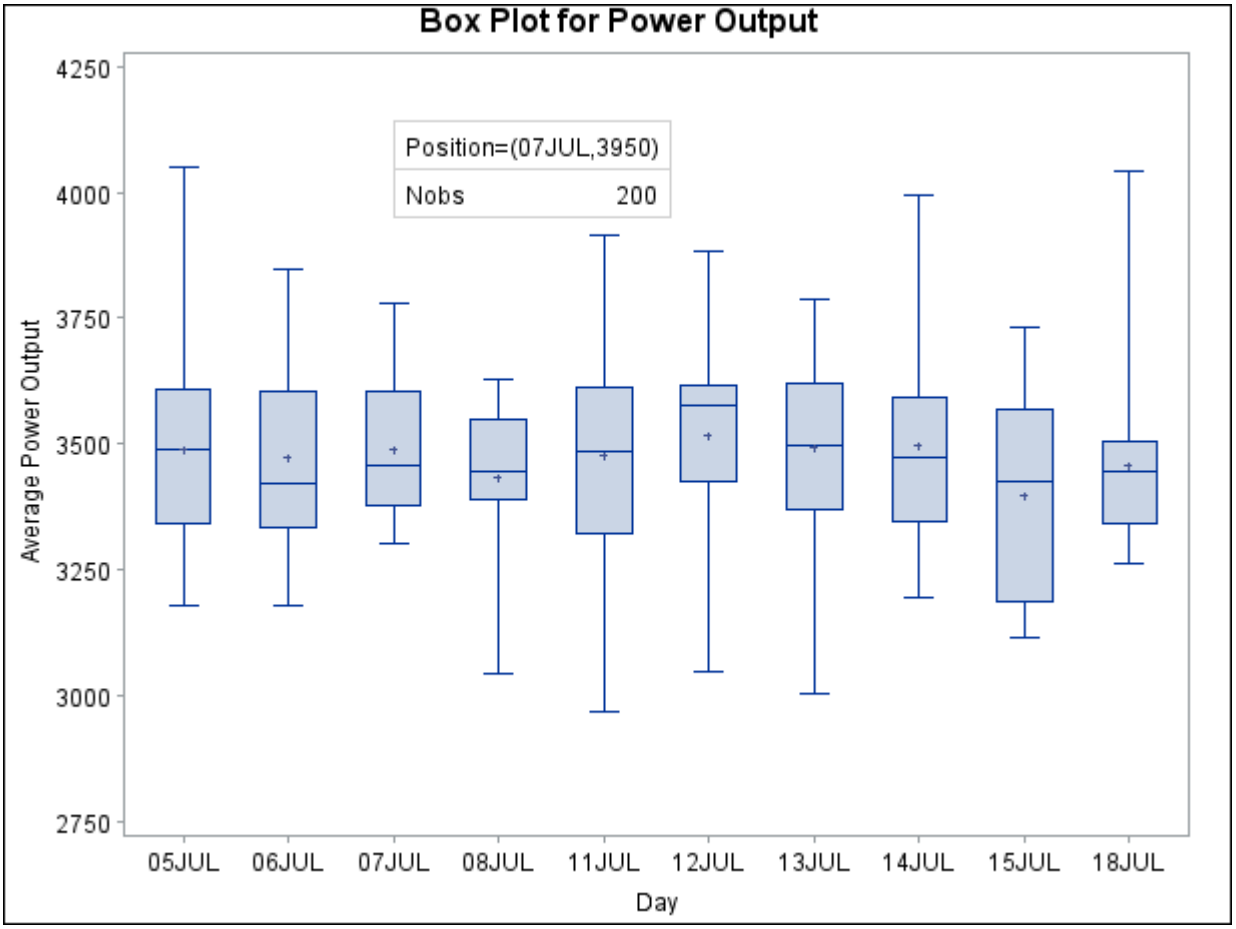
Data Unit Coordinates

If you specify the DATA option immediately following the coordinates, the inset is positioned using axis data units. For example, the following statements place the bottom-left corner of the inset at 07JUL on the horizontal axis and 3950 on the vertical axis:

```
title 'Box Plot for Power Output';
proc boxplot data=Turbine;
   plot KWatts*Day;
   inset nobs /
      header   = 'Position=(07JUL,3950)'
      position = ('07JUL94'd, 3950) data;
run;
```

The box plot is displayed in Figure 28.13. By default, the specified coordinates determine the position of the bottom-left corner of the inset. You can change this reference point with the REFPOINT= option, as in the next example.

Figure 28.13 Inset Positioned Using Data Unit Coordinates



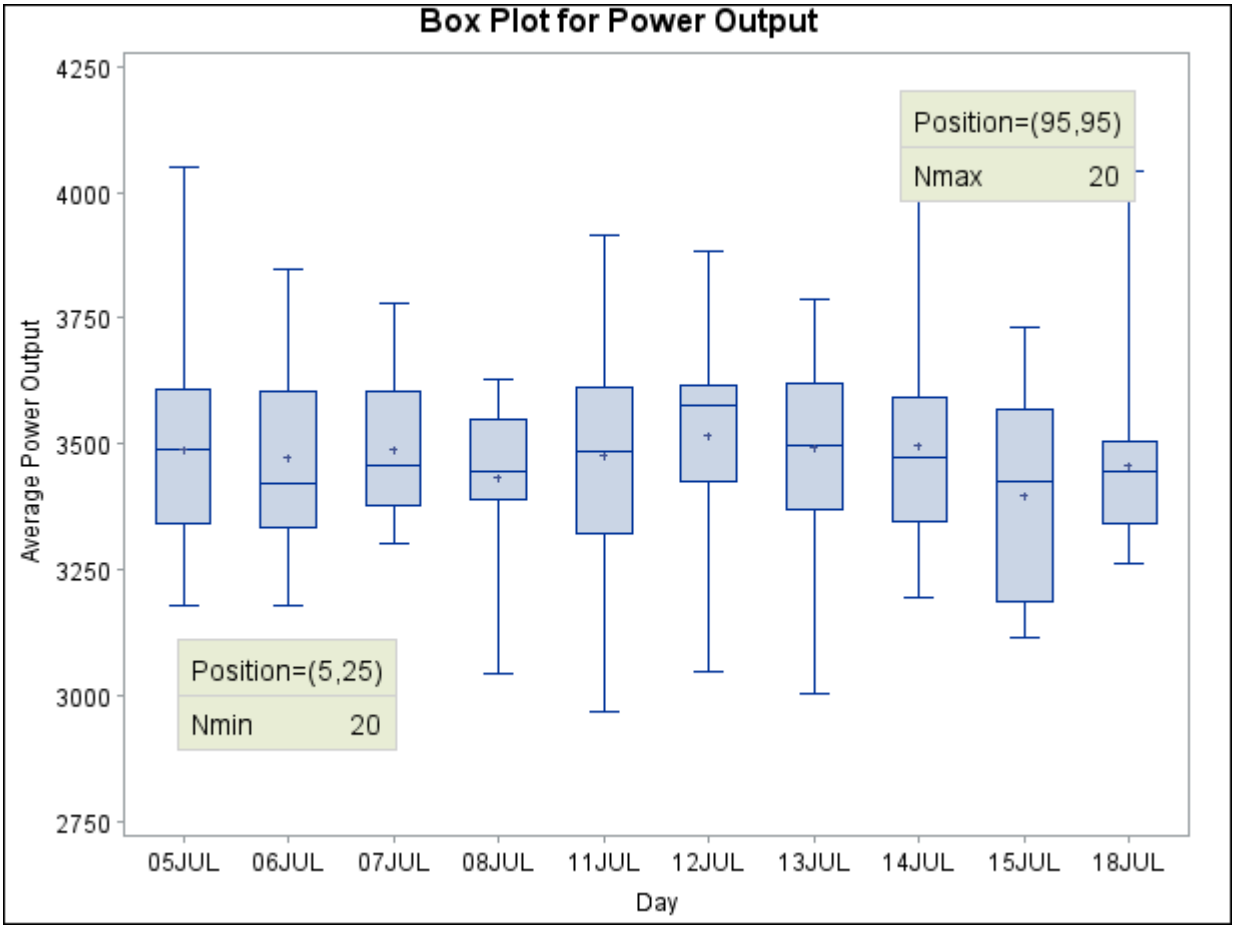
Axis Percent Unit Coordinates

If you do not use the DATA option, the inset is positioned using axis percent units. The coordinates of the bottom-left corner of the display are .0; 0/, while the coordinates of the top-right corner are .100; 100/. For example, the following statements create a box plot with two insets, both positioned using coordinates in axis percent units:

```
title 'Box Plot for Power Output';
proc boxplot data=Turbine;
   plot KWatts*Day;
   inset nmin / position = (5,25)
                header   = 'Position=(5,25)'
                height   = 3
                cfill    = ywh
                refpoint = tl;
   inset nmax / position = (95,95)
                header   = 'Position=(95,95)'
                height   = 3
                cfill    = ywh
                refpoint = tr;
run;
```

The display is shown in Figure 28.14. Notice that the REFPOINT= option is used to determine which corner of the inset is placed at the coordinates specified with the POSITION= option. The first inset has REFPOINT=TL, so the top-left corner of the inset is positioned 5% of the way across the horizontal axis and 25% of the way up the vertical axis. The second inset has REFPOINT=TR, so the top-right corner of the inset is positioned 95% of the way across the horizontal axis and 95% of the way up the vertical axis. Note also that coordinates in axis percent units must be *between* 0 and 100.

Figure 28.14 Inset Positioned Using Axis Percent Unit Coordinates



Displaying Blocks of Data

To display data organized in blocks of consecutive observations, specify one or more *block variables* in parentheses after the group variable in the PLOT statement. The block variables must be variables in the input data set. The BOXPLOT procedure displays a legend identifying blocks of consecutive observations with identical values of the block variables. The legend displays one track of values for each block variable containing formatted values of the block variable.

The values of a block variable must be the same for all observations with the same value of the group variable. In other words, groups must be nested within blocks determined by block variables.

The following statements create a SAS data set containing diameter measurements for a part produced on three different machines:

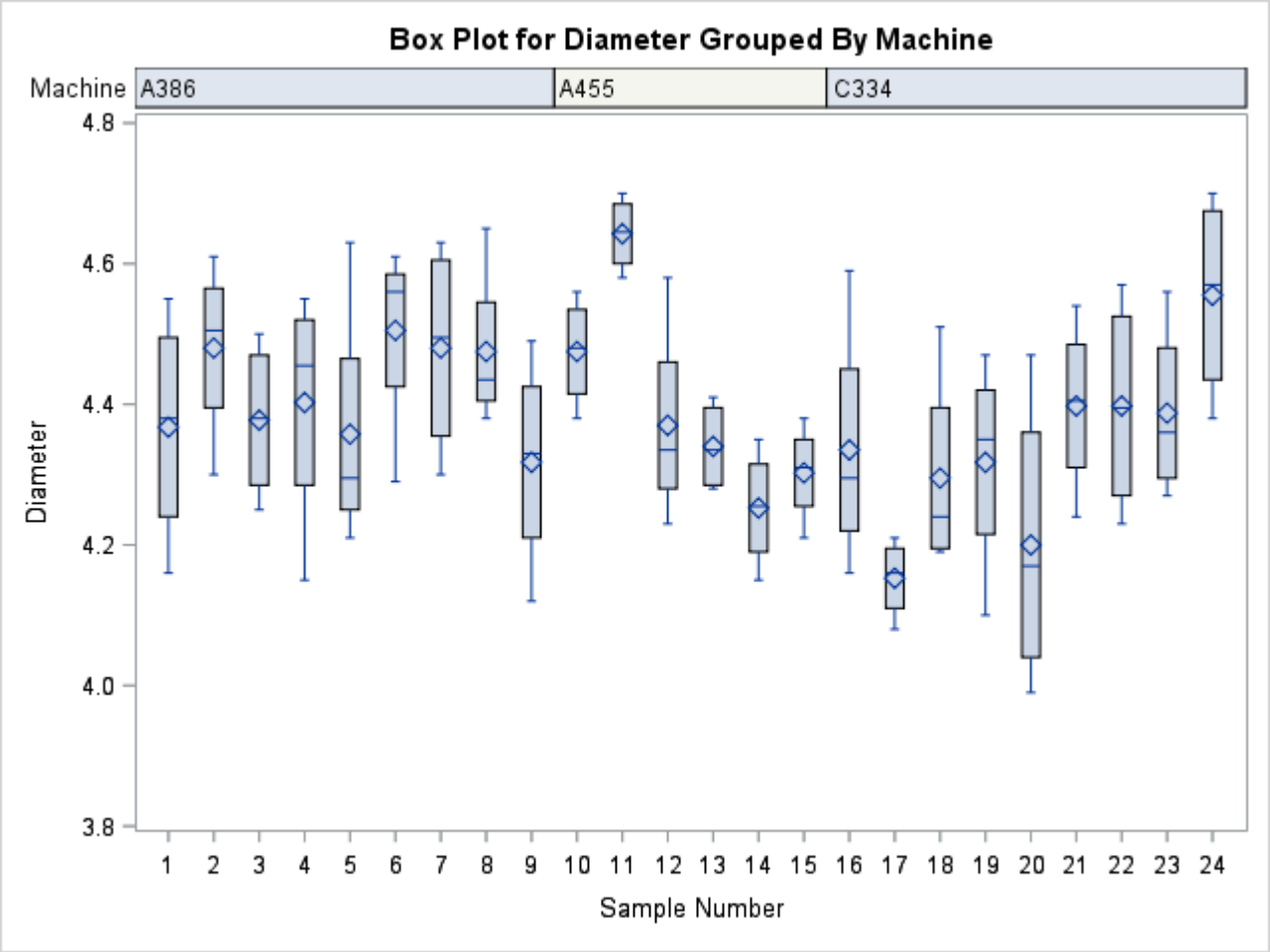
```
data Parts;
   length Machine $ 4;
   input Sample Machine $ @;
   do i= 1 to 4;
      input Diam @;
      output;
   end;
   drop i;
   datalines;
 1  A386  4.32 4.55 4.16 4.44
 2  A386  4.49 4.30 4.52 4.61
 3  A386  4.44 4.32 4.25 4.50
 4  A386  4.55 4.15 4.42 4.49
 5  A386  4.21 4.30 4.29 4.63
 6  A386  4.56 4.61 4.29 4.56
 7  A386  4.63 4.30 4.41 4.58
 8  A386  4.38 4.65 4.43 4.44
 9  A386  4.12 4.49 4.30 4.36
10  A455  4.45 4.56 4.38 4.51
11  A455  4.62 4.67 4.70 4.58
12  A455  4.33 4.23 4.34 4.58
13  A455  4.29 4.38 4.28 4.41
14  A455  4.15 4.35 4.28 4.23
15  A455  4.21 4.30 4.32 4.38
16  C334  4.16 4.28 4.31 4.59
17  C334  4.14 4.18 4.08 4.21
18  C334  4.51 4.20 4.28 4.19
19  C334  4.10 4.33 4.37 4.47
20  C334  3.99 4.09 4.47 4.25
21  C334  4.24 4.54 4.43 4.38
22  C334  4.23 4.48 4.31 4.57
23  C334  4.27 4.40 4.32 4.56
24  C334  4.70 4.65 4.49 4.38
;
```

The following statements create a box plot for the measurements in the Parts data set grouped into blocks by the block variable Machine:

```
ods graphics on;
title 'Box Plot for Diameter Grouped By Machine';
proc boxplot data=Parts;
   plot Diam*Sample (Machine) / odstitle = title;
   label Sample  = 'Sample Number'
         Machine = 'Machine'
         Diam    = 'Diameter';
run;
```

The ODSTITLE= option uses the title specified in the SAS title as the graph title. Note the LABEL statement used to provide labels for the axes and for the block legend. The plot is shown in Figure 28.15.

Figure 28.15 Box Plot Using a Block Variable



The unique consecutive values of Machine (A386, A455, and C334) are displayed in a legend above the plot. That is the default location of the block legend. You can control the position of the block legend with the BLOCKPOS= option. See the BLOCKPOS= option for details.

By default, block variable values that are too long to fit into the available space in a block legend are not displayed. You can specify the BLOCKLABTYPE= option to display lengthy labels. Specify BLOCKLABTYPE=SCALED to scale down the text size of the values so they all fit. Use BLOCKLAB-

TYPE=TRUNCATED to truncate lengthy values. You can also use BLOCKLABTYPE=*height* to specify a text height in vertical percent screen units for the values.

You can control the position of legend labels with the BLOCKLABELPOS= option. Valid BLOCKLABELPOS= values are ABOVE (the default, as shown in Figure 28.15) and LEFT.

Clipping Extreme Values

By default a box plot's vertical axis is scaled to accommodate all the values in all groups. If the variation between groups is large with respect to the variation within groups, or if some groups contain extreme outlier values, the vertical axis scale can become so large that the box-and-whiskers plots are compressed. In such cases, you can clip the extreme values to produce a more readable plot, as illustrated in the following example.

A company produces copper tubing. The diameter measurements (in millimeters) for 15 batches of five tubes each are provided in the data set Newtubes:

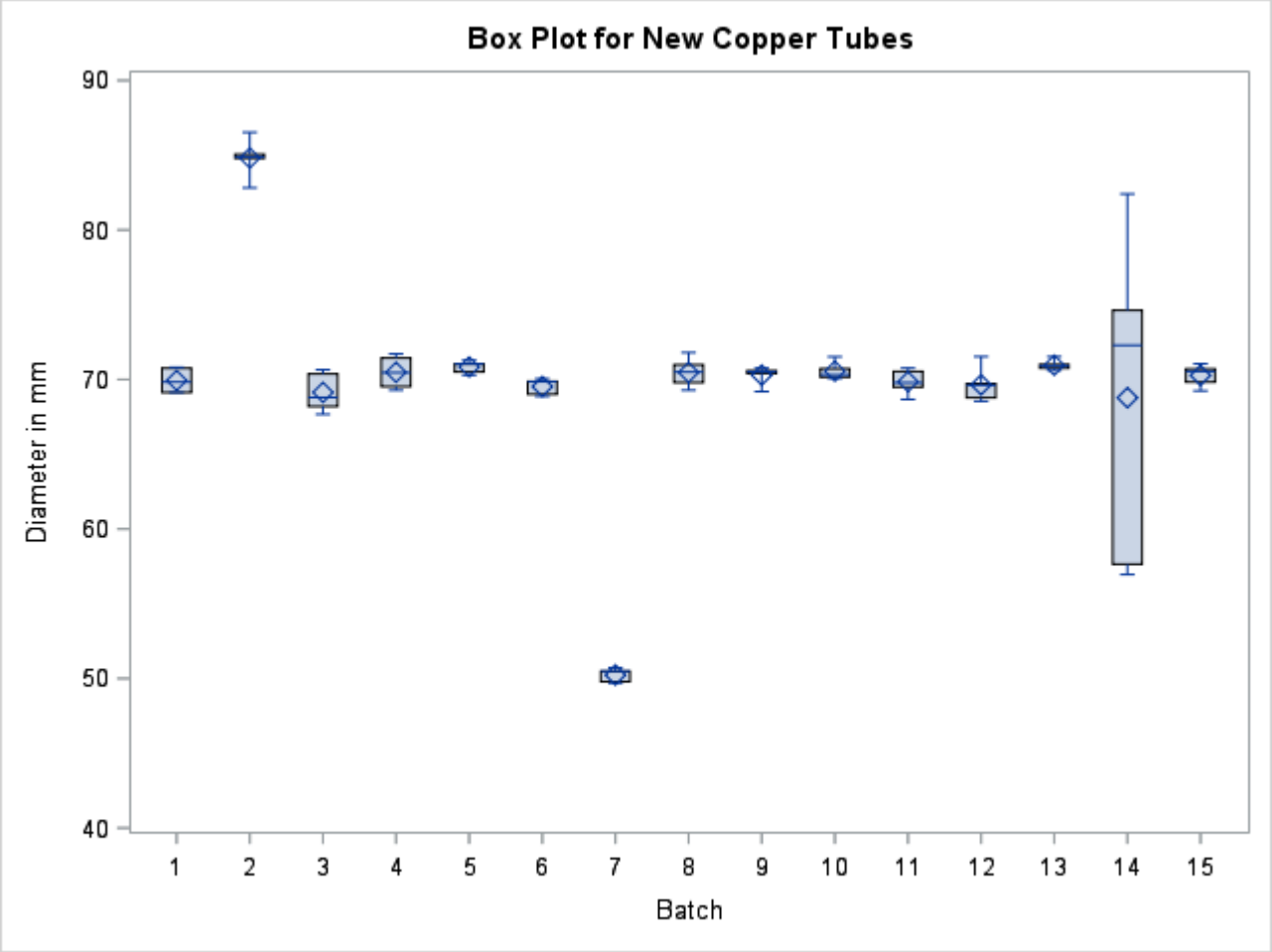
```
data Newtubes;
   label Diameter='Diameter in mm';
   do Batch = 1 to 15;
      do  i = 1 to 5;
         input Diameter @@;
         output;
      end;
   end;
   datalines;
69.13  69.83  70.76  69.13  70.81
85.06  82.82  84.79  84.89  86.53
67.67  70.37  68.80  70.65  68.20
71.71  70.46  71.43  69.53  69.28
71.04  71.04  70.29  70.51  71.29
69.01  68.87  69.87  70.05  69.85
50.72  50.49  49.78  50.49  49.69
69.28  71.80  69.80  70.99  70.50
70.76  69.19  70.51  70.59  70.40
70.16  70.07  71.52  70.72  70.31
68.67  70.54  69.50  69.79  70.76
68.78  68.55  69.72  69.62  71.53
70.61  70.75  70.90  71.01  71.53
74.62  56.95  72.29  82.41  57.64
70.54  69.82  70.71  71.05  69.24
;
```

The following statements create a box plot of the tube diameters:

```
ods graphics on;
title 'Box Plot for New Copper Tubes' ;
proc boxplot data=Newtubes;
   plot Diameter*Batch / odstitle = title;
run;
```

The box plot is shown in Figure 28.16.

Figure 28.16 Compressed Box Plots



Note that the diameters in batch 2 are significantly larger, and those in batch 7 significantly smaller, than those in most of the other batches. The default vertical axis scaling causes the box-and-whiskers plots to be compressed.

You can produce a more useful box plot by specifying the CLIPFACTOR=*factor* option, where *factor* is a value greater than one. Clipping is applied as follows:

1. The mean of the first quartile values ($\overline{Q1}$) and the mean of the third quartile values ($\overline{Q3}$) are computed across all groups.
2. The following values define the clipping range:

   y_{max} D $\overline{Q1}$ C .$\overline{Q3}$ $\overline{Q1}$/ $factor$

   and

   y_{min} D $\overline{Q3}$.$\overline{Q3}$ $\overline{Q1}$/ $factor$

   Any statistic greater than y_{max} or less than y_{min} is ignored during vertical axis scaling.

NOTE:

Clipping is applied only to the plotted statistics and not to the statistics saved in an output data set.

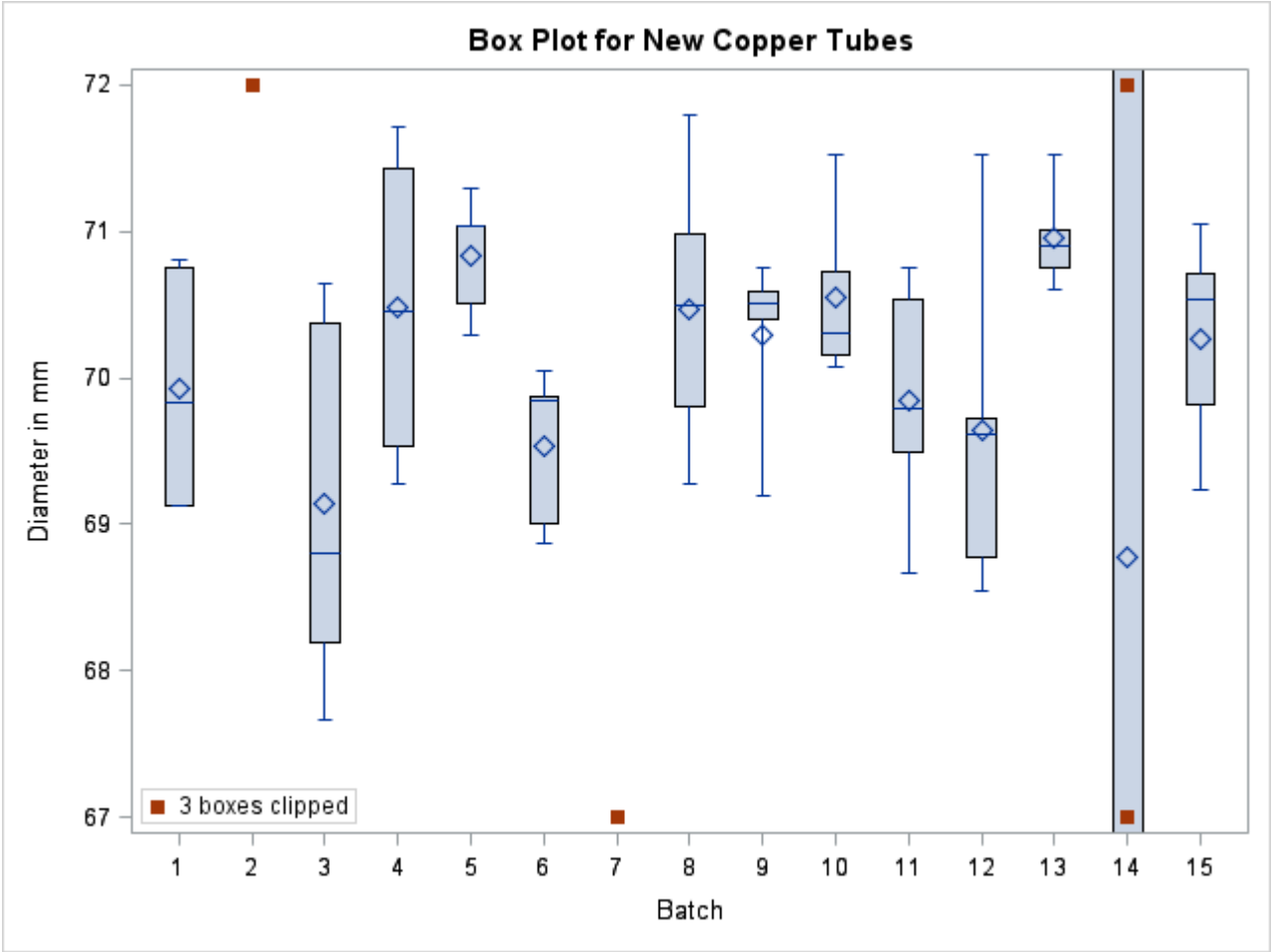
A special symbol is used for clipped points (the default symbol is a square), and a legend is added to the chart indicating the number of boxes that were clipped.

The following statements use a clipping factor of 1.5 to create a box plot of the same data plotted in Figure 28.16:

```
title 'Box Plot for New Copper Tubes' ;
proc boxplot data=Newtubes;
   plot Diameter*Batch /
      odstitle   = title
      clipfactor = 1.5;
run;
```

The clipped box plot is shown in Figure 28.17.

Figure 28.17 Box Plot with Clip Factor of 1.5



In Figure 28.17 the extreme values are clipped, making the box plot more readable. The box-and-whiskers plots for batches 2 and 7 are clipped completely, while the plot for batch 14 is clipped at both the top and bottom. Clipped points are marked with a square, and a clipping legend is added at the lower right of the display.

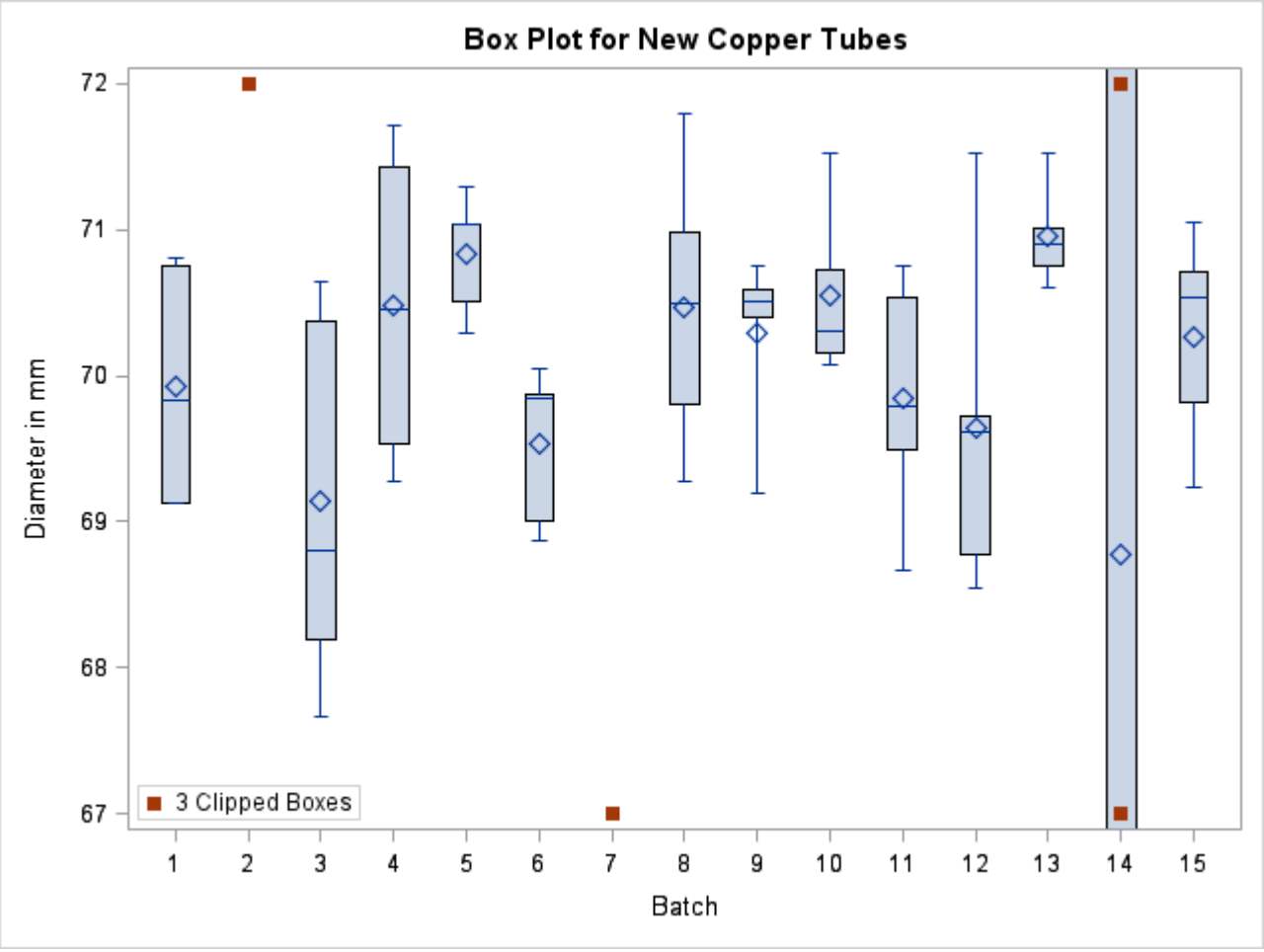
Other clipping options are available, as illustrated by the following statements:

```
title 'Box Plot for New Copper Tubes' ;
proc boxplot data=Newtubes;
   plot Diameter*Batch /
      odstitle    = title
      clipfactor  = 1.5
      cliplegend  = '# Clipped Boxes'
      clipsubchar = '#';
run;
```

The CLIPLEGEND= option requests a user-specified legend for the number of clipped boxes. Each occurrence in the legend of the character specified in the CLIPSUBCHAR= option is replaced by the number of clipped boxes.

Figure 28.18 shows the box plot with the modified clipping legend.

Figure 28.18 Box Plot with Clipping Options



For more information about clipping options, see the appropriate entries in the section "PLOT Statement Options" on page 1139.

ODS Graphics

Statistical procedures use ODS Graphics to create graphs as part of their output. ODS Graphics is described in detail in Chapter 21, "Statistical Graphics Using ODS."

Before you create graphs, ODS Graphics must be enabled (for example, by specifying the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" on page 607 in Chapter 21, "Statistical Graphics Using ODS."

The overall appearance of graphs is controlled by ODS styles. Styles and other aspects of using ODS Graphics are discussed in the section "A Primer on ODS Statistical Graphics" on page 606 in Chapter 21, "Statistical Graphics Using ODS."

The appearance of a box plot produced using ODS Graphics is determined by the style associated with the ODS destination where the graph is produced. PLOT statement options used to control the appearance of traditional graphs are ignored for ODS Graphics output.

When producing ODS graphical displays, the PLOT statement assigns a name to each graph it creates. You can use this name to reference the graph when using ODS. The name is listed in Table 28.10.

Table 28.10 Graphs Produced by PROC BOXPLOT

ODS Graph Name	Plot Description
Boxplot	box-and-whiskers plots for groups

Examples: BOXPLOT Procedure

This section provides advanced examples of the PLOT statement.

Example 28.1: Displaying Summary Statistics in a Box Plot

This example demonstrates how you can use the INSET and INSETGROUP statements to include tables of summary statistics in your box plots. The following statements produce a box plot of the Turbine data set from the section "Getting Started: BOXPLOT Procedure" on page 1123, augmented with insets containing summary statistics:

```
ods graphics off;
title 'Box Plot for Power Output';
proc boxplot data=Turbine;
   plot KWatts*Day;
   inset min mean max stddev /
      header = 'Overall Statistics'
```

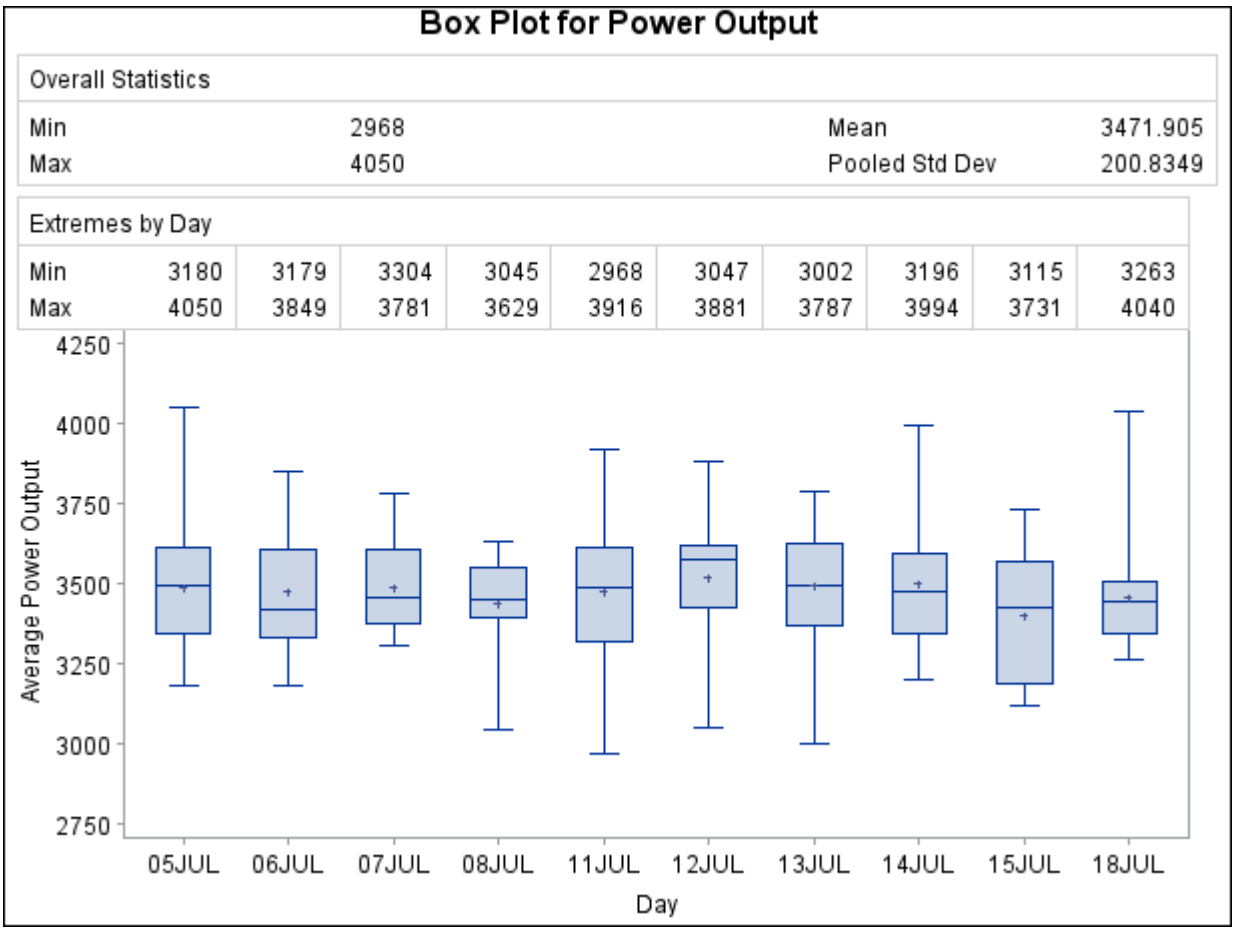
```
      pos    = tm;
   insetgroup min max /
      header = 'Extremes by Day';
run;
```

The INSET statement produces an inset of overall summary statistics. The keywords listed before the slash (/) request the minimum, mean, maximum, and standard deviation computed over all days. The POS=TM option places the inset in the top margin of the plot.

The INSETGROUP statement produces an inset containing statistics calculated for each group separately. The MIN and MAX keywords request the minimum and maximum observations from each day, respectively.

The resulting plot is shown in Output 28.1.1.

Output 28.1.1 Box Plot with Insets



Example 28.2: Using Box Plots to Compare Groups

In this example a box plot is used to compare the delay times of airline flights during the Christmas holidays with the delay times prior to the holiday period. The following statements create a data set named Times with the delay times in minutes for 25 flights each day. When a flight is canceled, the delay is recorded as a missing value.

```
data Times;
   informat Day date7. ;
   format   Day date7. ;
   input Day @ ;
   do Flight=1 to 25;
      input Delay @ ;
      output;
   end;
   datalines;
16DEC88   4  12   2   2  18   5   6  21   0   0
          0  14   3   .   2   3   5   0   6  19
          7   4   9   5  10
17DEC88   1  10   3   3   0   1   5   0   .   .
          1   5   7   1   7   2   2  16   2   1
          3   1  31   5   0
18DEC88   7   8   4   2   3   2   7   6  11   3
          2   7   0   1  10   2   3  12   8   6
          2   7   2   4   5
19DEC88  15   6   9   0  15   7   1   1   0   2
          5   6   5  14   7  20   8   1  14   3
         10   0   1  11   7
20DEC88   2   1   0   4   4   6   2   2   1   4
          1  11   .   1   0   6   5   5   4   2
          2   6   6   4   0
21DEC88   2   6   6   2   7   7   5   2   5   0
          9   2   4   2   5   1   4   7   5   6
          5   0   4  36  28
22DEC88   3   7  22   1  11  11  39  46   7  33
         19  21   1   3  43  23   9   0  17  35
         50   0   2   1   0
23DEC88   6  11   8  35  36  19  21   .   .   4
          6  63  35   3  12  34   9   0  46   0
          0  36   3   0  14
24DEC88  13   2  10   4   5  22  21  44  66  13
          8   3   4  27   2  12  17  22  19  36
          9  72   2   4   4
25DEC88   4  33  35   0  11  11  10  28  34   3
         24   6  17   0   8   5   7  19   9   7
         21  17  17   2   6
26DEC88   3   8   8   2   7   7   8   2   5   9
          2   8   2  10  16   9   5  14  15   1
         12   2   2  14  18
;
```

In the following statements, the MEANS procedure is used to count the number of canceled flights for each day. This information is then added to the data set Times.

```
proc means data=Times noprint;
   var Delay;
   by Day;
   output out=Cancel nmiss=ncancel;
run;
```

```
data Times;
   merge Times Cancel;
   by Day;
run;
```

The following statements create a data set named Weather containing information about possible causes for delays, and then merge this data set with the data set Times:

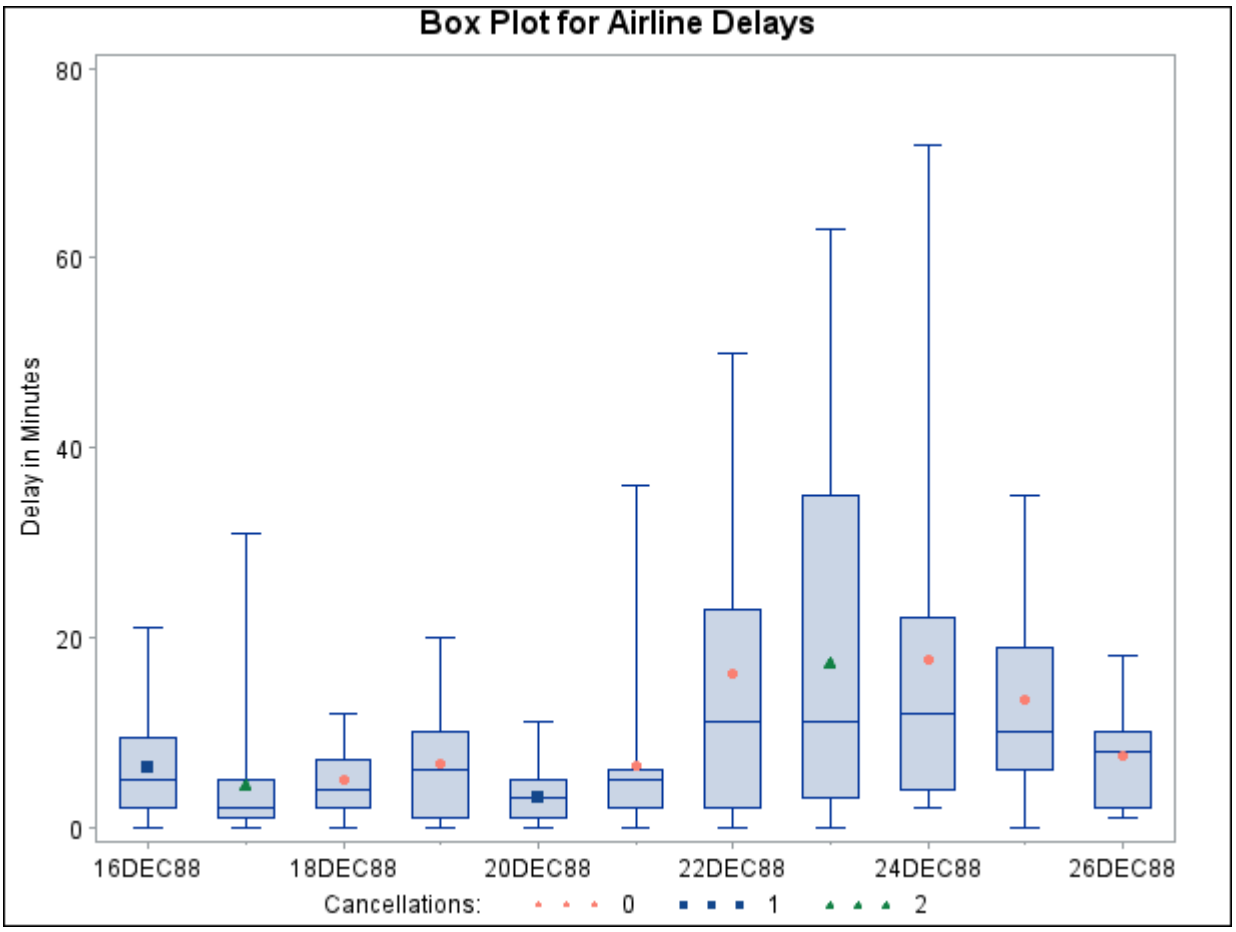
```
data Weather;
   informat Day date7. ;
   format   Day date7. ;
   length Reason $ 16 ;
   input Day Flight Reason & ;
   datalines;
16DEC88  8   Fog
17DEC88  18  Snow Storm
17DEC88  23  Sleet
21DEC88  24  Rain
21DEC88  25  Rain
22DEC88  7   Mechanical
22DEC88  15  Late Arrival
24DEC88  9   Late Arrival
24DEC88  22  Late Arrival
;

data Times;
   merge Times Weather;
   by Day Flight;
run;
```

The following statements create a box plot for the complete set of data:

```
ods graphics off;
symbol1 value=dot            c=salmon h=2.0 pct;
symbol2 value=squarefilled   c=vigb   h=2.0 pct;
symbol3 value=trianglefilled c=vig    h=2.0 pct;
title 'Box Plot for Airline Delays';
proc boxplot data=Times;
   plot Delay*Day = ncancel /
       nohlabel
       symbollegend = legend1;
   legend1 label = ('Cancellations:');
   label Delay = 'Delay in Minutes';
run;
goptions reset=symbol;
```

The level of the *symbol variable* ncancel determines the symbol marker for each group mean, and the SYMBOLLEGEND= option controls the appearance of the legend for the symbols. The NOHLABEL option suppresses the horizontal axis label. The resulting box plot is shown in Output 28.2.1.

Output 28.2.1 Box Plot for Airline Data



The delay distributions from December 22 through December 25 are drastically different from the delay distributions during the pre-holiday period. Both the mean delay and the variability of the delays are much greater during the holiday period.

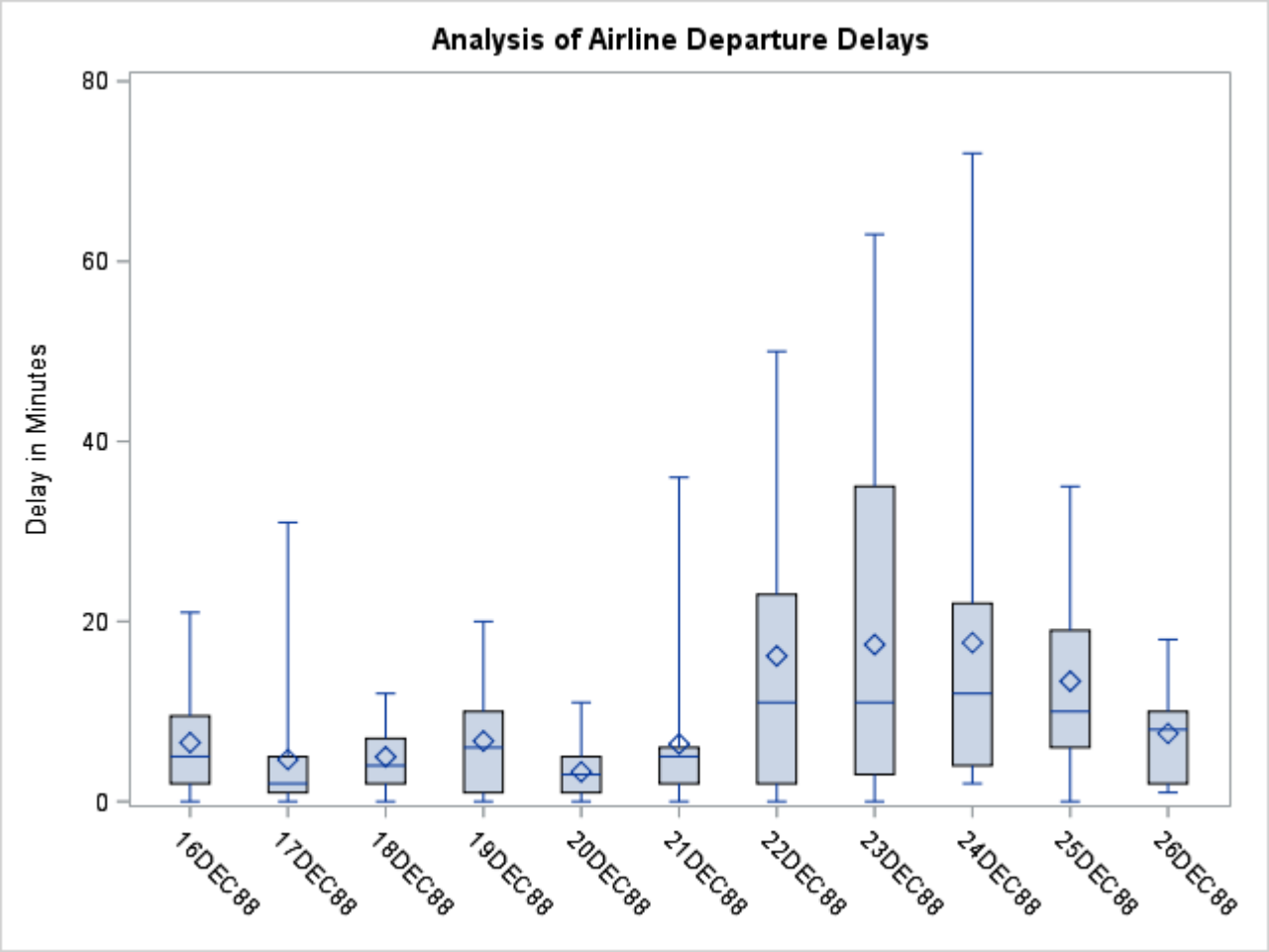
Example 28.3: Creating Various Styles of Box-and-Whiskers Plots

This example uses the flight delay data of the preceding example to illustrate how you can create box plots with various styles of box-and-whiskers plots. The following statements create a plot that displays skeletal box-and-whiskers plots:

```
ods graphics on;
title 'Analysis of Airline Departure Delays';
title2 'BOXSTYLE=SKELETAL';
proc boxplot data=Times;
   plot Delay*Day /
      boxstyle = skeletal
      odstitle = title
      nohlabel;
   label Delay = 'Delay in Minutes';
run;
```

In a skeletal box-and-whiskers plot, the whiskers are drawn from the quartiles to the extreme values of the group. The skeletal box plot is the default style, so you can also produce a skeletal box plot by omitting the BOXSTYLE= option. Output 28.3.1 shows the skeletal box plot.

Output 28.3.1 BOXSTYLE=SKELETAL



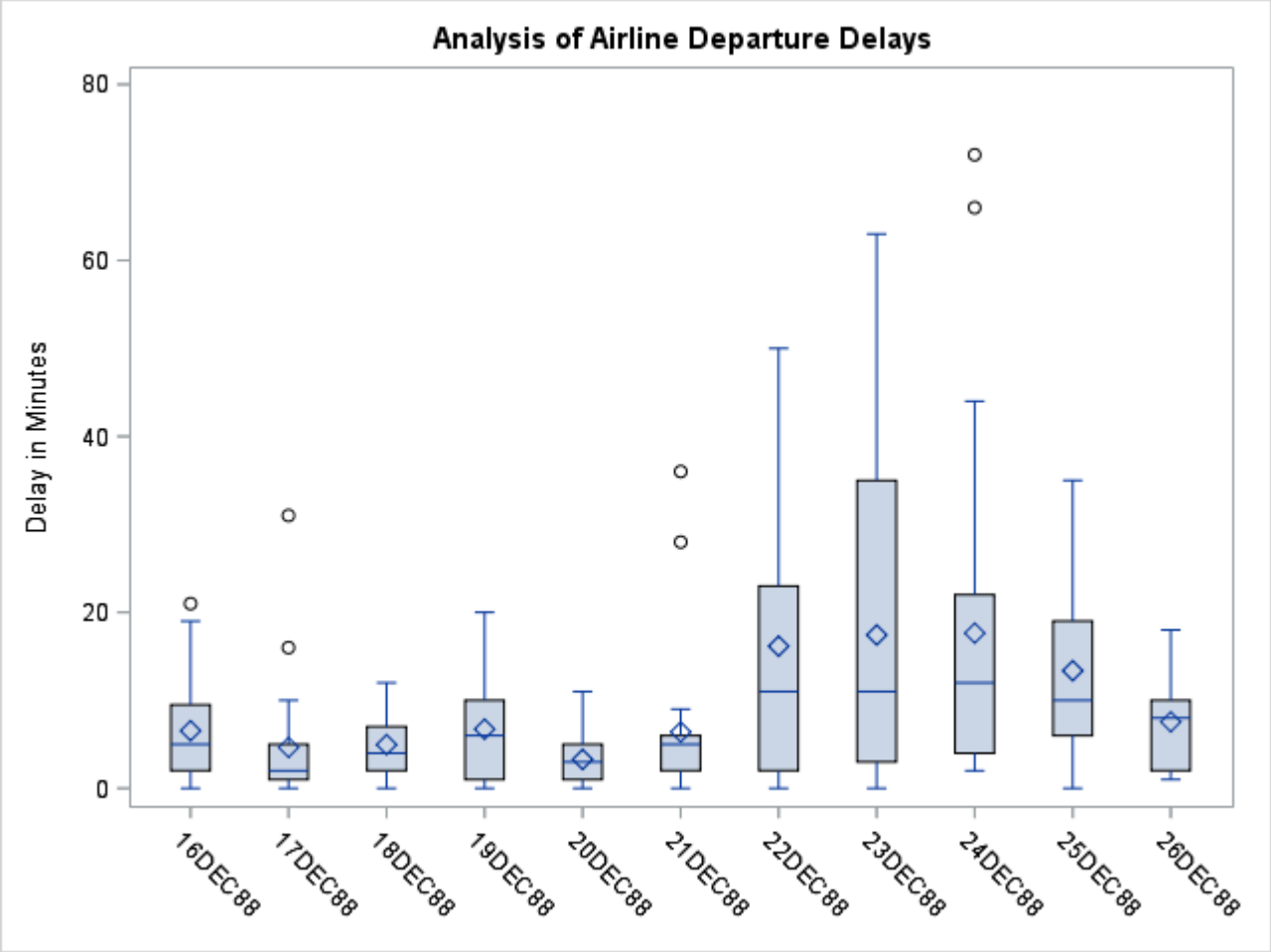
The following statements request a schematic box:

```
title 'Analysis of Airline Departure Delays';
title2 'BOXSTYLE=SCHEMATIC';
proc boxplot data=Times;
   plot Delay*Day /
      boxstyle = schematic
      odstitle = title
      nohlabel;
   label Delay = 'Delay in Minutes';
run;
```

When you specify BOXSTYLE=SCHEMATIC, the whiskers are drawn to the most extreme points in the group that lie within or equal to the *fences.* The *upper fence* is defined as the third quartile (represented by the upper edge of the box) plus 1.5 times the interquartile range (IQR). The *lower fence* is defined as the first quartile (represented by the lower edge of the box) minus 1.5 times the interquartile range. Observations

outside the fences are identified with a special symbol. The default symbol is a square, and you can specify the shape and color for this symbol with the IDSYMBOL= and IDCOLOR= options. Serifs are added to the whiskers by default. For further details, see the entry for the BOXSTYLE= option. The plot is shown in Output 28.3.2.

Output 28.3.2 BOXSTYLE=SCHEMATIC

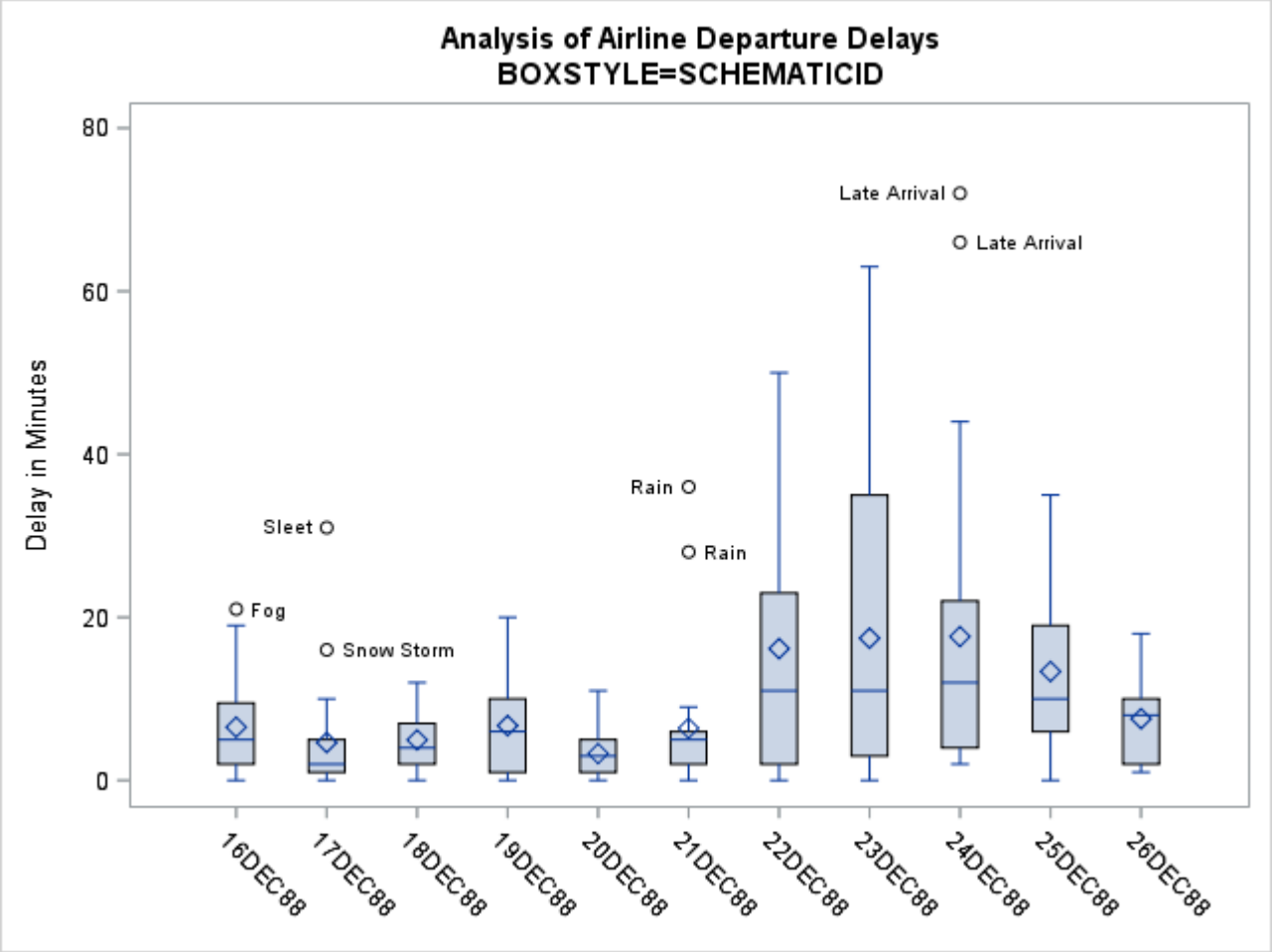


The following statements create a schematic box plot in which the observations outside the fences are labeled:

```
title 'Analysis of Airline Departure Delays';
title2 'BOXSTYLE=SCHEMATICID';
proc boxplot data=Times;
   plot Delay*Day /
      boxstyle  = schematicid
      odstitle  = title
      odstitle2 = title2
      nohlabel;
   id Reason;
   label Delay = 'Delay in Minutes';
run;
```

If you specify BOXSTYLE=SCHEMATICID, schematic box-and-whiskers plots are created and the value of the first ID variable (in this case, Reason) is used to label each observation outside the fences. The box plot is shown in Output 28.3.3.

Output 28.3.3 BOXSTYLE=SCHEMATICID



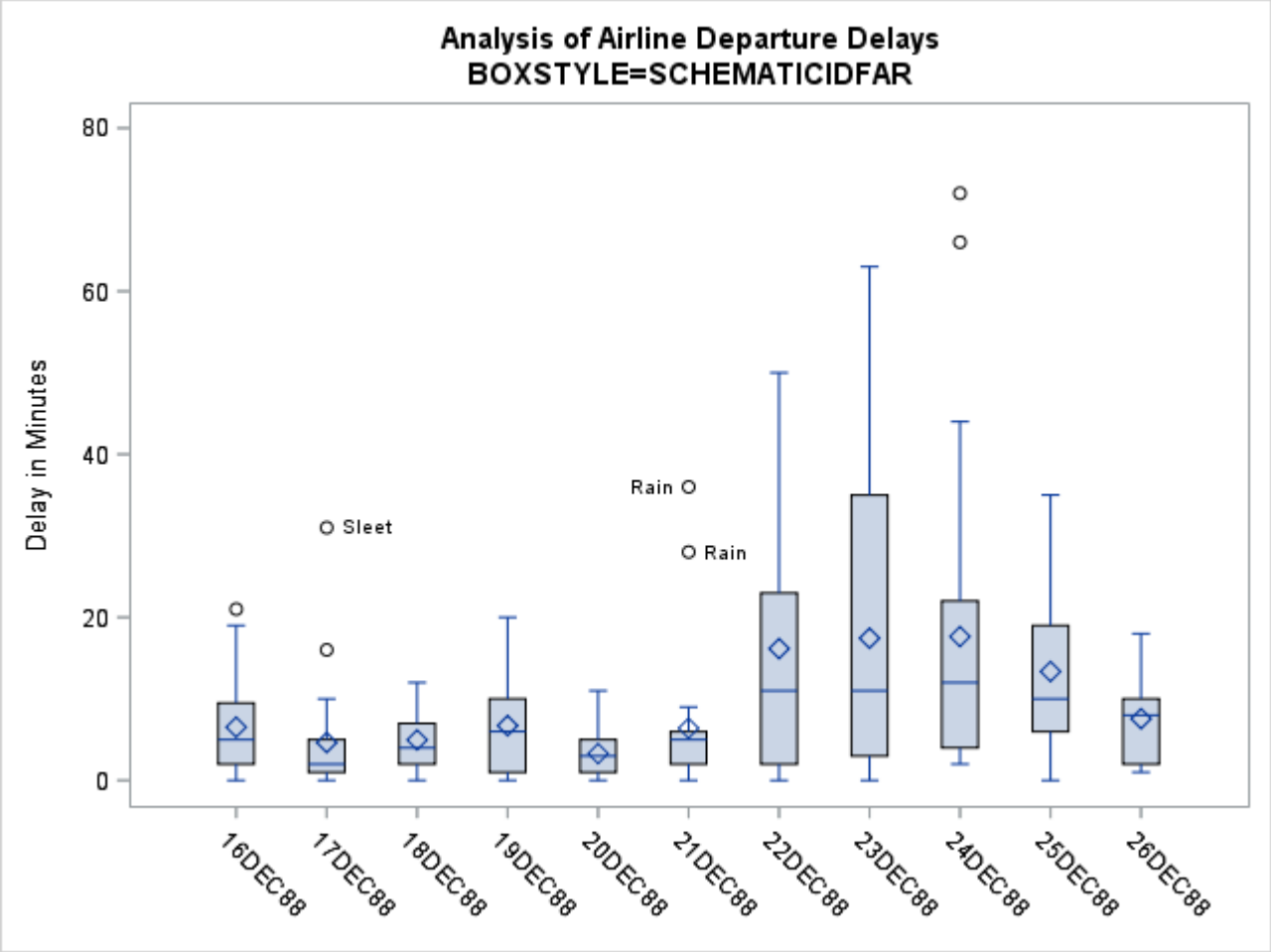
The following statements create a box plot with schematic box-and-whiskers plots in which only the extreme observations outside the fences are labeled:

```
title 'Analysis of Airline Departure Delays';
title2 'BOXSTYLE=SCHEMATICIDFAR';
proc boxplot data=Times;
   plot Delay*Day /
      boxstyle = schematicidfar
      odstitle  = title
      odstitle2 = title2
      nohlabel;
   id Reason;
   label Delay = 'Delay in Minutes';
run;
```

If you specify BOXSTYLE=SCHEMATICIDFAR, the value of the first ID variable is used to label each observation outside the lower and upper *far fences*. The lower and upper far fences are located 3 IQR below

the 25th percentile and 3 IQR above the 75th percentile, respectively. Observations between the fences and the far fences are identified with a symbol but are not labeled. The box plot is shown in Output 28.3.4.

Output 28.3.4 BOXSTYLE=SCHEMATICIDFAR



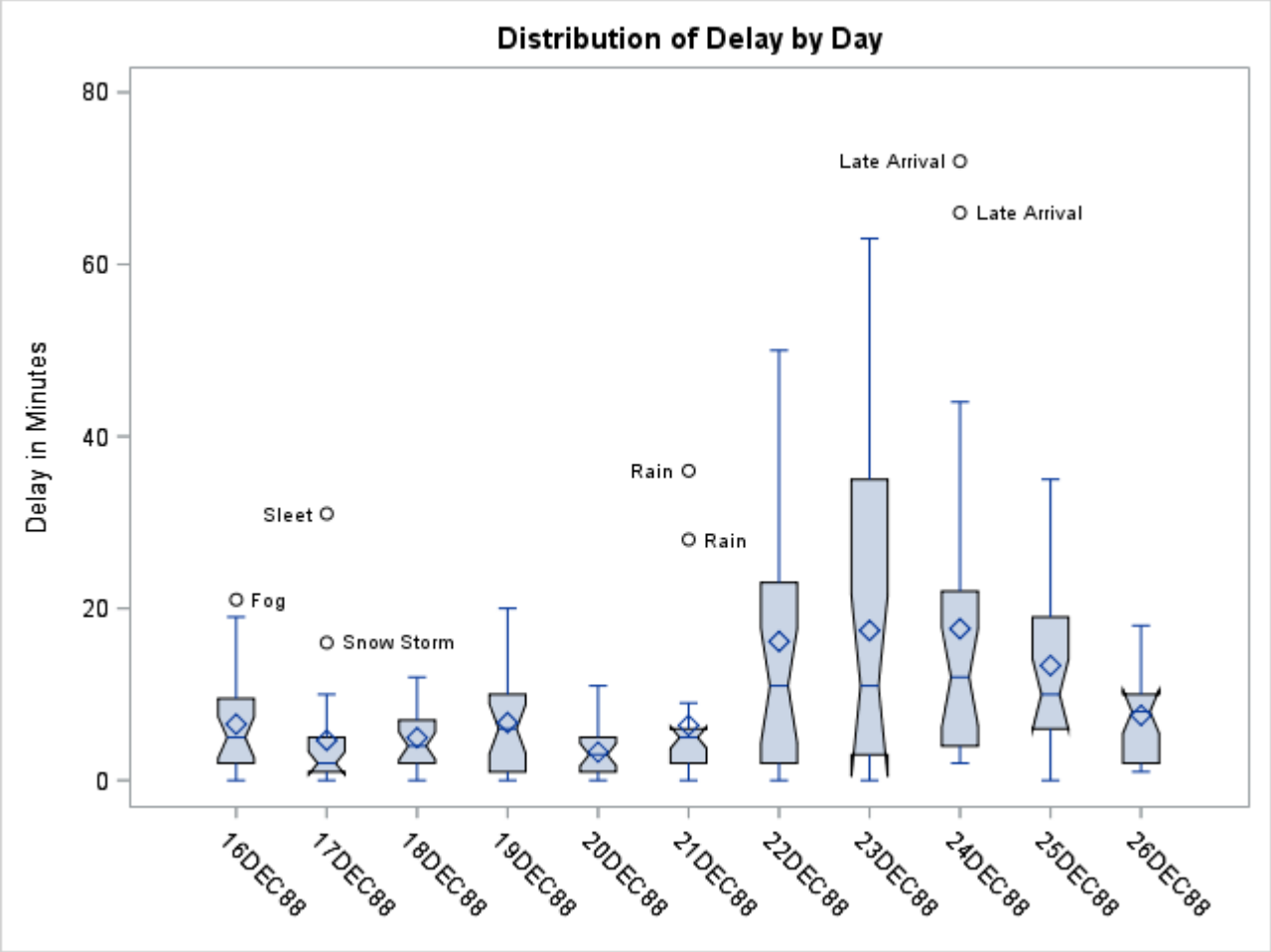
Example 28.4: Creating Notched Box-and-Whiskers Plots

The following statements use the flight delay data of Example 28.2 to create box-and-whiskers plots with notches:

```
proc boxplot data=Times;
   plot Delay*Day /
      boxstyle  = schematicid
      odstitle  = title
      odstitle2 = title2
      nohlabel
      notches;
   id Reason;
   label Delay = 'Delay in Minutes';
run;
```

The notches, requested with the NOTCHES option, measure the significance of the difference between two medians. The medians of two box plots are significantly different at approximately the 0.95 confidence level if the corresponding notches do not overlap. For example, in Output 28.4.1, the median for December 20 is significantly different from the median for December 24.

Output 28.4.1 Notched Side-by-Side Box-and-Whiskers Plots



Example 28.5: Creating Box-and-Whiskers Plots with Varying Widths

This example shows how to create a box plot with box-and-whiskers plots whose widths vary proportionately with the group size. The following statements create a SAS data set named Times2 that contains flight departure delays (in minutes) recorded daily for eight consecutive days:

```
data Times2;
   label Delay = 'Delay in Minutes';
   informat Day date7. ;
   format   Day date7. ;
   input Day @ ;
   do Flight=1 to 25;
      input Delay @ ;
      output;
```

```
   end;
   datalines;
01MAR90   12  4   2   2  15   8   0  11   0   0
          0  12   3   .   2   3   5   0   6  25
          7   4   9   5  10
02MAR90   1   .   3   .   0   1   5   0   .   .
          1   5   7   .   7   2   2  16   2   1
          3   1  31   .   0
03MAR90   6   8   4   2   3   2   7   6  11   3
          2   7   0   1  10   2   5  12   8   6
          2   7   2   4   5
04MAR90  12   6   9   0  15   7   1   1   0   2
          5   6   5  14   7  21   8   1  14   3
         11   0   1  11   7
05MAR90   2   1   0   4   .   6   2   2   1   4
          1  11   .   1   0   .   5   5   .   2
          3   6   6   4   0
06MAR90   8   6   5   2   9   7   4   2   5   1
          2   2   4   2   5   1   3   9   7   8
          1   0   4  26  27
07MAR90   9   6   6   2   7   8   .   .  10   8
          0   2   4   3   .   .   .   7   .   6
          4   0   .   .   .
08MAR90   1   6   6   2   8   8   5   3   5   0
          8   2   4   2   5   1   6   4   5  10
          2   0   4   1   1
;
```

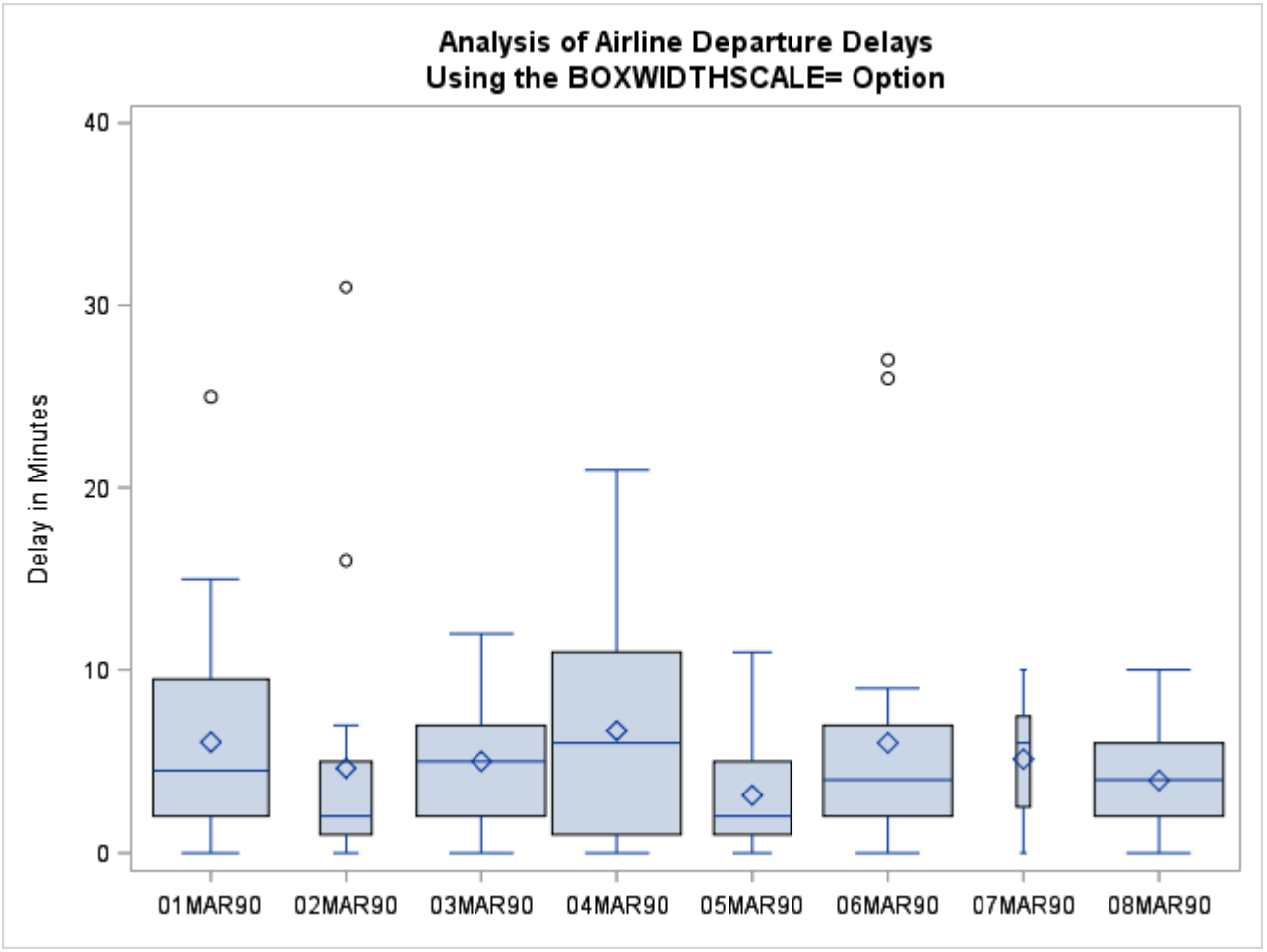
The following statements create a box plot with varying box widths:

```
title 'Analysis of Airline Departure Delays';
title2 'Using the BOXWIDTHSCALE= Option';
proc boxplot data=Times2;
   plot Delay*Day /
      boxstyle      = schematic
      odstitle      = title
      odstitle2     = title2
      boxwidthscale = 1
      nohlabel
      bwslegend;
run;
```

The BOXWIDTHSCALE=*value* option specifies that the widths of the box-and-whiskers plots vary in proportion to a particular function of the group size *n*. The function is determined by *value* and is identified on the box plot with a legend if the BWSLEGEND option is specified. The BOXWIDTHSCALE= option is useful in situations where the group sizes vary widely.

Output 28.5.1 shows the resulting box plot.

Output 28.5.1 Box Plot with Box-and-Whiskers Plots of Varying Widths



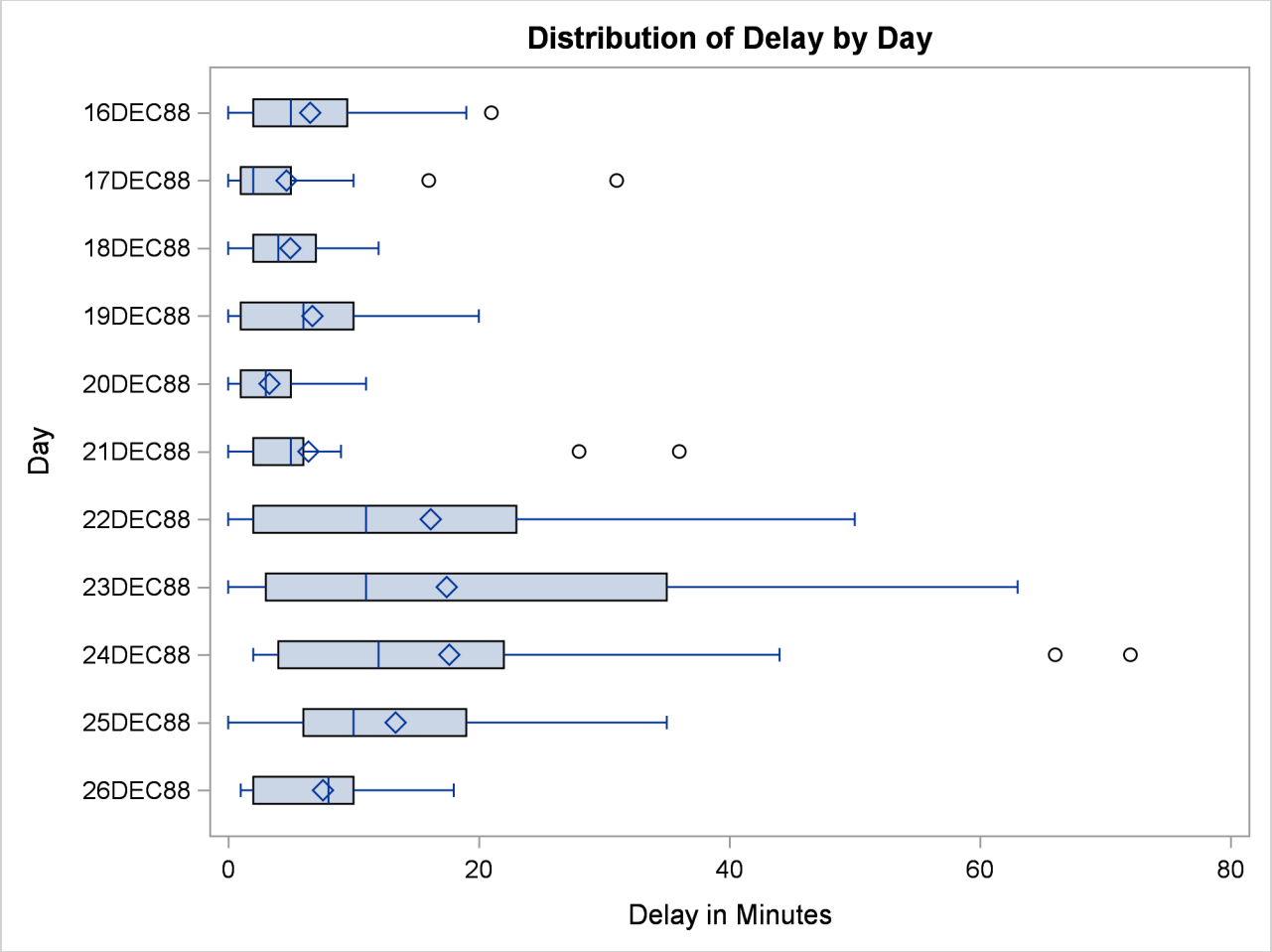
Example 28.6: Creating Horizontal Box-and-Whiskers Plots

The following statements use the HORIZONTAL option, which is supported only for ODS Graphics output, to produce a horizontal box plot:

```
proc boxplot data=Times;
   plot Delay*Day /
      boxstyle = schematic
      horizontal;
   label Delay = 'Delay in Minutes';
run;
```

The horizontal box plot is shown in Output 28.6.1.

Output 28.6.1 Horizontal Box Plot



References

McGill, R., Tukey, J. W., and Larsen, W. A. (1978). "Variations of Box Plots." *American Statistician* 32:12–16.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Subject Index

Syntax Index