Chapter 18
The BOXPLOT Procedure

Chapter Table of Contents

Chapter 18
The BOXPLOT Procedure

Overview

The BOXPLOT procedure creates side-by-side box-and-whisker plots of measurements organized in groups. A box-and-whisker 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-whisker 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

- control the style of the box-and-whisker 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
- control the layout and appearance of the plot

Getting Started

This section demonstrates how you can use the BOXPLOT procedure to produce box plots for your data.

Suppose that a petroleum company uses a turbine to heat water into steam that is pumped into the ground to make oil more 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 20 work 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
19JUL94 3795 3872 3559 3432 3322 3587 3336 3732 3451 3215
19JUL94 3594 3410 3335 3216 3336 3638 3419 3515 3399 3709
20JUL94 3850 3431 3460 3623 3516 3810 3671 3602 3480 3388
20JUL94 3365 3845 3520 3708 3202 3365 3731 3840 3182 3677
21JUL94 3711 3648 3212 3664 3281 3371 3416 3636 3701 3385
21JUL94 3769 3586 3540 3703 3320 3323 3480 3750 3490 3395
22JUL94 3596 3436 3757 3288 3417 3331 3475 3600 3690 3534
22JUL94 3306 3077 3357 3528 3530 3327 3113 3812 3711 3599
25JUL94 3428 3760 3641 3393 3182 3381 3425 3467 3451 3189
25JUL94 3588 3484 3759 3292 3063 3442 3712 3061 3815 3339
26JUL94 3746 3426 3320 3819 3584 3877 3779 3506 3787 3676
26JUL94 3727 3366 3288 3684 3500 3501 3427 3508 3392 3814
27JUL94 3676 3475 3595 3122 3429 3474 3125 3307 3467 3832
27JUL94 3383 3114 3431 3693 3363 3486 3928 3753 3552 3524
28JUL94 3349 3422 3674 3501 3639 3682 3354 3595 3407 3400
28JUL94 3401 3359 3167 3524 3561 3801 3496 3476 3480 3570
29JUL94 3618 3324 3475 3621 3376 3540 3585 3320 3256 3443
29JUL94 3415 3445 3561 3494 3140 3090 3561 3800 3056 3536
01AUG94 3421 3787 3454 3699 3307 3917 3292 3310 3283 3536
01AUG94 3756 3145 3571 3331 3725 3605 3547 3421 3257 3574
;
run;
```

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 rational groups, it is referred to as the *group variable*. The variable kwatts contains the output measurements and is referred to as the *analysis variable*.

You can create a box plot to examine the distribution of power output for each day. The following statements create the box plot shown in Figure 18.1.

```
symbol color = salmon h = .8;
goptions ftext=swiss;
axis1 minor=none color=black label=(angle=90 rotate=0);
title ’Box Plot for Power Output’;

proc boxplot data=Turbine;
   plot kwatts*day/ cframe   = vligb
                    cboxes   = dagr
                    cboxfill = ywh
                    vaxis    = axis1;
run;
```

The input data set Turbine is specified with the DATA= option in the PROC BOXPLOT statement. The PLOT statement requests a box-and-whisker 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).

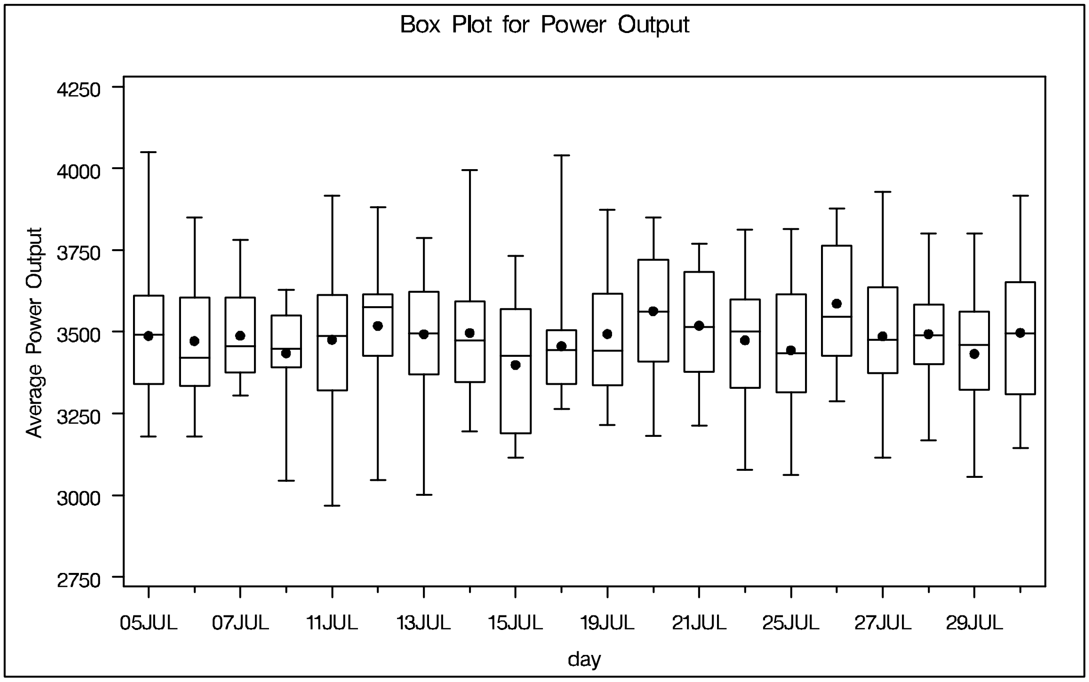


Figure 18.1. Box Plot for Power Output Data

The box plot displayed in Figure 18.1 represents summary statistics for the analysis variable kwatts; each of the 20 box-and-whisker 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 the 75th percentiles)

- the dot in the box interior represents the mean
- the horizontal line in the box interior represents the median
- the vertical lines issuing from the box extend to the minimum and maximum values of the analysis variable

Syntax

The syntax for the BOXPLOT procedure is as follows:

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

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 starts the BOXPLOT procedure. The following options can appear in the PROC BOXPLOT statement.

ANNOTATE=*SAS-data-set*

specifies an ANNOTATE= type data set, as described in *SAS/GRAPH Software: Reference*, that enhances all box plots requested in subsequent PLOT statements.

DATA=*SAS-data-set*

names the SAS data set containing the data to be analyzed. If you omit the DATA= option, the procedure uses the most recently created SAS data set.

GOUT=<*libref.*>*output catalog*

specifies the SAS catalog in which to save the 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.

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 SYMBOLn statements (refer to *SAS/GRAPH Software: 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 for each option follow.

Table 18.1 lists all options in the PLOT statement by function.

Table 18.1. PLOT Statement Options

Option	Description
Options for Controlling Box Appearance	
BOXCONNECT	connects group means in box-and-whisker plots
BOXCONNECT=	connects group means, medians, maximum values, minimum values, or quartiles in box-and-whisker plots
BOXSTYLE=	specifies style of box-and-whisker plots
BOXWIDTH=	specifies width of box-and-whisker plots

Table 18.1. (continued)

Option	Description
BOXWIDTHSCALE=	specifies that widths of box-and-whisker plots vary proportionately to group size
CBOXES=	specifies color for outlines of box-and-whisker plots
CBOXFILL=	specifies fill color for interior of box-and-whisker plots
IDCOLOR=	specifies outlier symbol color in schematic box-and-whisker plots
IDCTEXT=	specifies outlier label color in schematic box-and-whisker plots
IDFONT=	specifies outlier label font in schematic box-and-whisker plots
IDHEIGHT=	specifies outlier label height in schematic box-and-whisker plots
IDSYMBOL=	specifies outlier symbol in schematic box-and-whisker plots
LBOXES=	specifies line types for outlines of box-and-whisker plots
NOSERIFS	eliminates serifs from the whiskers of box-and-whisker plots
NOTCHES	specifies that box-and-whisker plots are to be notched
PCTLDEF=	specifies percentile definition used for box-and-whisker plots
Options for Plotting and Labeling Points	
CCONNECT=	specifies color for line segments that connect points on plot
SYMBOLLEGEND=	specifies LEGEND statement for levels of the symbol variable
SYMBOLORDER=	specifies order in which symbols are assigned for levels of the symbol variable
Reference Line Options	
CHREF=	specifies color for lines requested by HREF= option
CVREF=	specifies color for lines requested by VREF= option
HREF=	specifies position of reference lines perpendicular to horizontal axis on box plot
HREFLABELS=	specifies labels for HREF= lines
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 is to be applied uniformly to plots created for all BY groups
VREF=	specifies position of reference lines perpendicular to vertical axis on box plot
VREFLABELS=	specifies labels for VREF= lines
VREFLABPOS=	specifies position of VREFLABELS= labels
Block Variable Legend Options	
BLOCKLABELPOS=	specifies position of label for the block variable legend
BLOCKLABTYPE=	specifies text size of the block variable legend
BLOCKPOS=	specifies vertical position of the block variable legend

Table 18.1. (continued)

Option	Description
BLOCKREP	repeats identical consecutive labels in the block variable legend
CBLOCKLAB=	specifies color for filling background in the block variable legend
CBLOCKVAR=	specifies one or more variables whose values are colors for filling background of the block variable legend
Axis and Axis Label Options	
CAXIS=	specifies color for axis lines and tick marks
CFRAME=	specifies fill colors for frame for plot area
CONTINUOUS	produces horizontal axis for continuous group variable values
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 label for horizontal axis
NOTICKREP	specifies that only the first occurrence of repeated, adjacent group values is to 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 axis labels that are strung out vertically
VAXIS=	specifies major tick mark values for vertical axis of box plot
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 missing values between identical character group values signify the start of a new group
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 the BOXWIDTHSCALE= option
DESCRIPTION=	specifies string that appears in the description field of the PROC GREPLAY master menu for box plot
FONT=	specifies software font for labels and legends on plots
NAME=	specifies name that appears in the name field of the PROC GREPLAY master menu for box plot

Table 18.1. (continued)

Option	Description
NLEGEND	requests a legend displaying group sample sizes
PAGENUM=	specifies the form of the label used in pagination
PAGENUMPOS=	specifies the position of the page number requested with the PAGENUM= option
Grid Options	
ENDGRID	adds grid after last box-and-whisker plot
GRID	adds grid to box plot
LENDGRID=	specifies line type for grid requested with the ENDGRID option
LGRID=	specifies line type for grid requested with the 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 a numeric group variable
MAXPANELS=	specifies maximum number of pages or screens for plot
NOFRAME	suppresses frame for plot area
NPANELPOS=	specifies number of group positions per panel on each plot
REPEAT	repeats last group position on panel as first group position of next panel
TOTPANELS=	specifies number of pages or screens to be used to display plot

Following are explanations of the options that you can specify in the PLOT statement after a slash (/).

ANNOTATE=*SAS-data-set*

specifies an ANNOTATE= type data set, as described in *SAS/GRAPH Software: 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
BLOCKLABTYPE=*height*

specifies how lengthy block variable values are to be treated when there is insufficient space to display them in the block legend. If you specify the BLOCKLABTYPE=SCALED option, the values are uniformly reduced in height so that they fit. If you specify the BLOCKLABTYPE=TRUNCATED option, 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 422.

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 are to be displayed. By default, only the first block variable value in any block is displayed, and repeated block variable values are not displayed.

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

specifies that the points representing group means, medians, maximum values, minimum values, first quartiles, or third quartiles are to 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.

BOXSTYLE=*keyword*

specifies the style of the box-and-whisker 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-whisker 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 $1.5 \times \mathrm{IQR}$ below the 25th percentile; the upper fence is located at $1.5 \times \mathrm{IQR}$ above the 75th percentile. Similarly, the lower far fence is located at $3 \times \mathrm{IQR}$ below the 25th percentile; the upper far fence is located at $3 \times \mathrm{IQR}$ above the 75th percentile.

If you specify BOXSTYLE=SCHEMATIC, a whisker is drawn from the upper edge of the box to the largest value within the upper fence and from the lower edge of the box to the smallest value within the lower fence. Serifs are added to the whiskers by default. Observations outside the fences are identified with a special symbol; 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-whisker plot described in Chapter 2 of Tukey (1977). See Figure 18.4 and the discussion in the section "Styles of Box Plots" on page 418 for more information.

If you specify BOXSTYLE=SCHEMATICID, a schematic box-and-whisker plot is displayed in which the value of the first variable listed in the ID statement is used to label the symbol marking each observation outside the upper and lower fences.

If you specify BOXSTYLE=SCHEMATICIDFAR, a schematic box-and-whisker plot is displayed in which the value of the first variable listed in the ID statement 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 18.2 illustrates the elements of a skeletal box-and-whisker plot.

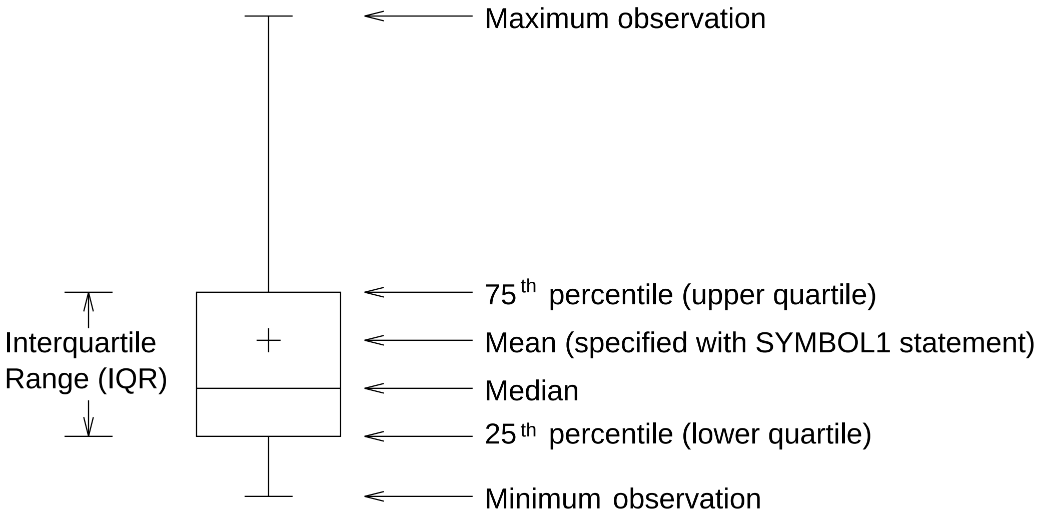

Maximum observation
75th percentile (upper quartile)
Mean (specified with SYMBOL1 statement)
Interquartile Range (IQR)
Median
25th percentile (lower quartile)
Minimum observation

Figure 18.2. Skeletal Box-and-Whisker Plot

The skeletal style of the box-and-whisker plot shown in Figure 18.2 is the default.

BOXWIDTH=*value*

specifies the width (in horizontal percent screen units) of the box-and-whisker plots.

BOXWIDTHSCALE=*value*

specifies that the box-and-whisker plot width is to vary proportionately to a particular function of the group size n. The function is determined by the *value*.

If you specify a positive value, the widths are proportional to n^{value}. In particular, if you specify BOXWIDTHSCALE=1, the widths are proportional to the group size. If you specify BOXWIDTHSCALE=0.5, the widths are proportional to $\sqrt{n}$, as described by McGill, Tukey, and Larsen (1978). If you specify BOXWIDTHSCALE=0, the widths are proportional to $\log(n)$. See Example 18.4 on page 432 for an illustration of the BOXWIDTHSCALE= option.

You can specify the BWSLEGEND option to display a legend identifying the function of n used to determine the box-and-whisker plot widths.

By default, the box widths are constant.

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. The default value is the first color in the device color list.

CBLOCKLAB=*color*

specifies a fill color for the frame that encloses the block variable label in a block legend. By default, this area is not filled.

CBLOCKVAR=*variable | (variable-list)*

specifies variables whose values are colors for filling the background of the legend associated with block variables. 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 (refer to *SAS/GRAPH Software: 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*
CBOXES=*(variable)*

specifies the colors for the outlines of the box-and-whisker 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-whisker plots.
- You can specify CBOXES=*(variable)* to provide a distinct outline color for each box-and-whisker plot as the value of the variable. The variable must be a character variable of length 8 or less in the input data set, and its values must be valid SAS/GRAPH color names (refer to *SAS/GRAPH Software: 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.

The default color is the second color in the device color list.

CBOXFILL=*color*
CBOXFILL=*(variable)*

specifies the interior fill colors for the box-and-whisker 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-whisker plots.
- You can specify CBOXFILL=*(variable)* to provide a distinct color for each box-and-whisker plot as the value of the variable. The variable must be a character variable of length 8 or less 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). Refer to *SAS/GRAPH Software: 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.

CCONNECT=*color*

specifies the color for the 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.

CFRAME=*color*
CFRAME=*(color-list)*

specifies the colors 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. You can specify a single color to fill the entire area.

CHREF=*color*

specifies the color for the lines requested by the HREF= option. The default value is the first color in the device color list.

CONTINUOUS

specifies that numeric group variable values are to be treated as continuous values. By default, the values of a numeric group variable are considered discrete values unless the HAXIS= option is specified. For more information, see the discussion in the section “Continuous Group Variables” on page 420.

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. The default value is the first color in the device color list.

DESCRIPTION=*'string'*
DES=*'string'*

specifies a description of the box plot, not longer than 40 characters, that appears in the PROC GREPLAY master menu. The default string is the 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 box-and-whisker plot. You can use the HAXIS= option to force space to be added to the horizontal axis.

FONT=*font*

specifies a software 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. Hardware characters are used by default. Refer to *SAS/GRAPH Software: Reference* for more information on the GOPTIONS statement.

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=*values*
HAXIS=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. Optionally, you can specify an axis name defined in a previous AXIS statement. Refer to *SAS/GRAPH Software: Reference* for more information on the AXIS statement.

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 420. Numeric values can be given in an explicit or implicit list. If the group variable is character, values must be quoted strings of length 16 or less. 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, and 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 may 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 software fonts specified with the FONT= option or with the FTEXT= option in the GOPTIONS statement. Refer to *SAS/GRAPH Software: Reference* for complete information on the GOPTIONS statement.

HMINOR=*n*
HM=*n*

specifies the number of minor tick marks between each major tick mark 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.

HREF=*values*
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 length 16 or less. 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='*label*$_1$' ... '*label*$_n$'
HREFLABEL='*label*$_1$' ... '*label*$_n$'
HREFLAB='*label*$_1$' ... '*label*$_n$'

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 HREFLABEL= 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

IDCOLOR=*color*

specifies the color of the symbol marker used to identify outliers in schematic box-and-whisker plots (that is, when you also specify one of the following options: BOXSTYLE=SCHEMATIC, BOXSTYLE=SCHEMATICID, and BOXSTYLE=SCHEMATICIDFAR). The default color is the color specified with the CBOXES= option; otherwise, the second color in the device color list is used.

IDCTEXT=*color*

specifies the color for the text used to label outliers when you specify one of the keywords 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 one of the keywords 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 one of the keywords SCHEMATICID or SCHEMATICIDFAR with the BOXSTYLE= option. The default value is the height specified with the HTEXT= option in the GOPTIONS statement. Refer to *SAS/GRAPH Software: Reference* for complete information on the GOPTIONS statement.

IDSYMBOL=*symbol*

specifies the symbol marker used to identify outliers in schematic box plots when you also specify one of the following options: BOXSTYLE=SCHEMATIC, BOXSTYLE=SCHEMATICID, and BOXSTYLE=SCHEMATICIDFAR. The default symbol is SQUARE.

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 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 (screen or page). 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=20 by default, the plot is displayed with two panels (screens or pages).

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 50 hours, matching the scale of the data, and the plot is displayed with two 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 420.

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

specifies the line types for the outlines of the box-and-whisker 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-whisker plots.
- You can specify LBOXES=*(variable)* to provide a distinct line type for each box-and-whisker 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. Refer to the description of the SYMBOL statement in *SAS/GRAPH Software: Reference* for more information on valid linetypes.

LENDGRID=*n*

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

LGRID=*n*

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

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. Refer to the description of the SYMBOL statement in *SAS/GRAPH Software: Reference* for more information on valid linetypes.

MAXPANELS=*n*

specifies the maximum number of pages or screens for a 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 for the box plot, not more than 8 characters, that appears in the PROC GREPLAY master menu.

NLEGEND

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

NOBYREF

specifies that the reference line information in an HREF= or VREF= data set is to 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.

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.

NOSERIFS

eliminates serifs from the whiskers of box-and-whisker plots.

NOTCHES

specifies that box-and-whisker plots are to be notched. The endpoints of the notches are located at the median plus and minus $1.58(\mathrm{IQR}/\sqrt{n})$, where IQR is the interquartile range and n is the group sample size. The medians (central lines) of two box-and-whisker plots are significantly different at approximately the 0.05 level if the corresponding notches do not overlap. Refer to McGill, Tukey, and Larsen (1978) for more information. Figure 18.3 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 sample size is small.

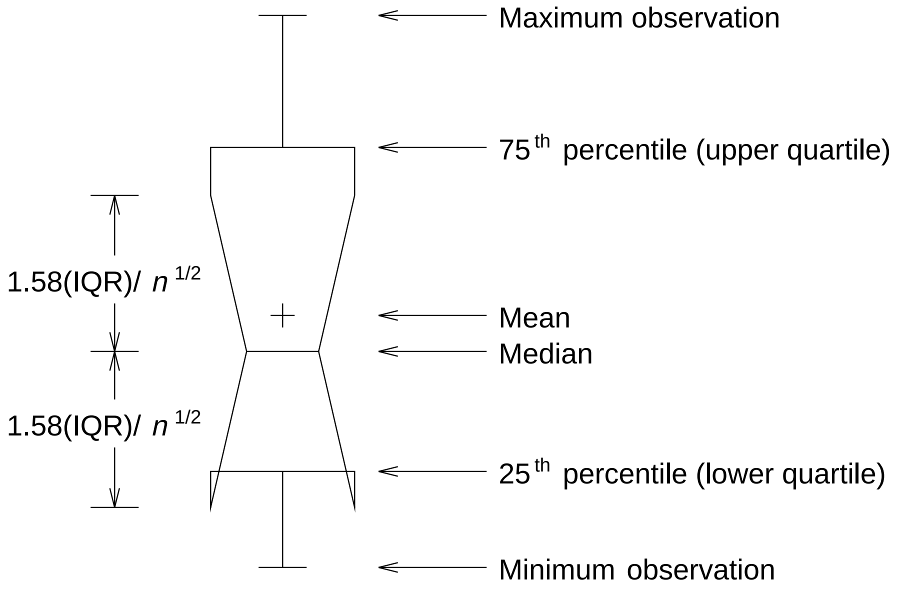


Figure 18.3. Box Plot: the NOTCHES Option

NOTICKREP

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

NOVANGLE

requests vertical axis labels that are oriented vertically. By default, the labels are drawn at an angle of 90 degrees if a software font is used.

NPANELPOS=*n*
NPANEL=*n*

specifies the number of group positions per panel. A panel is defined as a screen or page. You typically specify the NPANELPOS= option to display more box-and-whisker plots on a panel than the default number, which is $n = 20$.

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 $|n|$ 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.

PAGENUM=*'string'*

specifies the form of the label used for pagination. The string must be no longer than 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 keyword is BR.

PCTLDEF=*index*

specifies one of five definitions used to calculate percentiles in the construction of box-and-whisker 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 419. The default index is 5.

REPEAT
REP

specifies that the horizontal axis of a plot that spans multiple pages is to be arranged so that the last group position on a page is repeated as the first group position on the next page. 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*
SYMBOLLEGEND=NONE

controls the legend for the levels of a symbol variable (see Example 18.1). 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. Refer to *SAS/GRAPH Software: Reference* for more information on 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. 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 NPANEL= 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. You should specify a software font (using the FONT= option) in conjunction with the TURNHLABELS option. Otherwise, the labels may be displayed with a mixture of hardware and software fonts.

Note that arranging the labels vertically may leave insufficient room on the screen or page 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 on 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.

VMINOR=*n*
VM=*n*

specifies the number of minor tick marks between each major tick mark 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*
VREF=*SAS-data-set*

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

- Specify the values for the lines with a VREF= list. Examples of the VREF= option follow:

  ```
  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 length 16 or less. 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'*

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 VREFLABEL= 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*.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC BOXPLOT to obtain separate box plots for each group defined by the levels of the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables.

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

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED or DESCENDING in the BY statement for the 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 using the DATASETS procedure.

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

ID Statement

ID *variables* **;**

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

If you specify either the BOXSTYLE=SCHEMATICID option or the BOXSTYLE=SCHEMATICIDFAR option, the value of the first variable listed in the ID statement is used to label each extreme observation. For an example illustrating the use of the ID statement, see Example 18.2 or Example 18.3.

Details

Summary Statistics Represented by Box Plots

Table 18.2 lists the summary statistics represented in each box-and-whisker plot.

Table 18.2. Summary Statistics Represented by Box Plots

Group Summary Statistic	Feature of Box-and-Whisker 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," which follows, and as illustrated in Example 18.2 on page 427.

Input Data Set

You can read data (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

- *block-variables*
- *symbol-variable*
- BY variables
- ID variables

Styles of Box Plots

A box-and-whisker plot is displayed for the measurements in each group on the box plot. The skeletal style of the box-and-whisker plot shown in Figure 18.2 is the default. Figure 18.4 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 on page 403 for complete details.

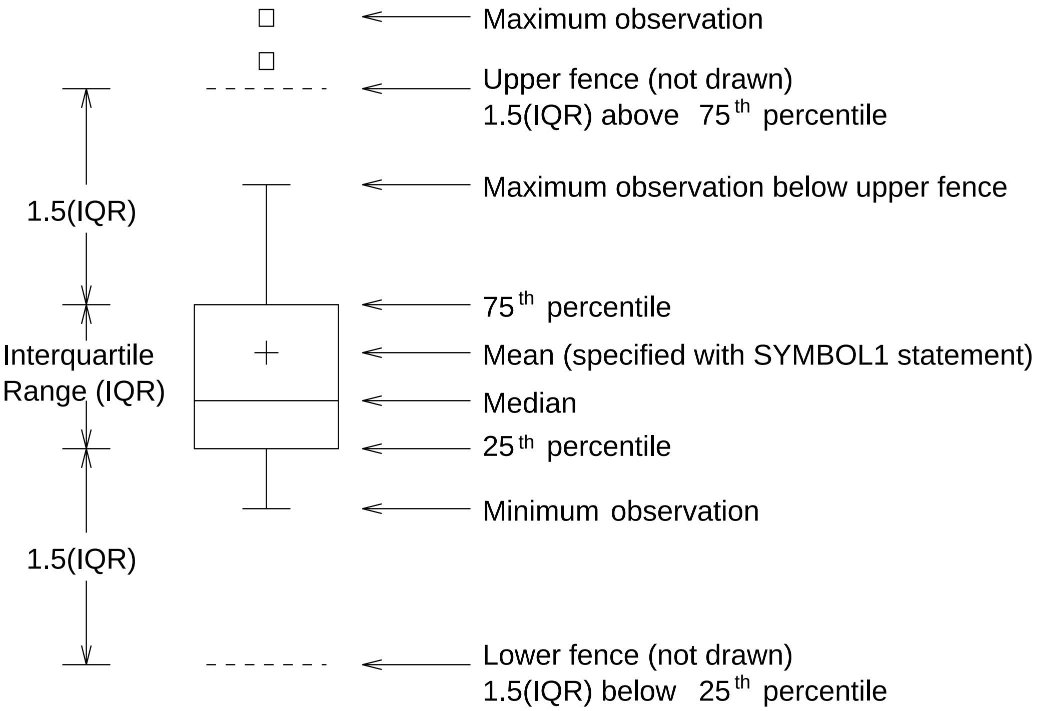


Figure 18.4. BOXSTYLE= SCHEMATIC

You can draw connecting lines between adjacent box-and-whisker plots 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 equals the number of nonmissing values for a variable and that $x_1, x_2, \ldots, x_n$ represents the ordered values of the analysis variable. For the *t*th percentile, set $p = t/100$.

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

$$np = j + g$$

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

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

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

PCTLDEF=3 empirical distribution function

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

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

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

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

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

PCTLDEF=5 empirical distribution function with averaging

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

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

Missing Values

An observation read from a DATA= 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.

Missing values of analysis variables generally lead to unequal group sizes.

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 plot shown in Figure 18.5:

```
symbol v=dot c=salmon;
goptions ftext=swiss;
axis1 minor=none color=black label=(angle=90 rotate=0);
title 'Box Plot for Power Output';

proc boxplot data=turbine;
   plot kwatts*day / turnhlabel
                     cframe   = vligb
                     cboxes   = dagr
                     cboxfill = ywh
                     vaxis    = axis1;
run;
```

The labels on the horizontal axis in Figure 18.5 do not represent 20 consecutive days, but the box-and-whisker plots are evenly spaced (note that the TURNHLABEL option orients the horizontal axis labels vertically so there is room to display them all).

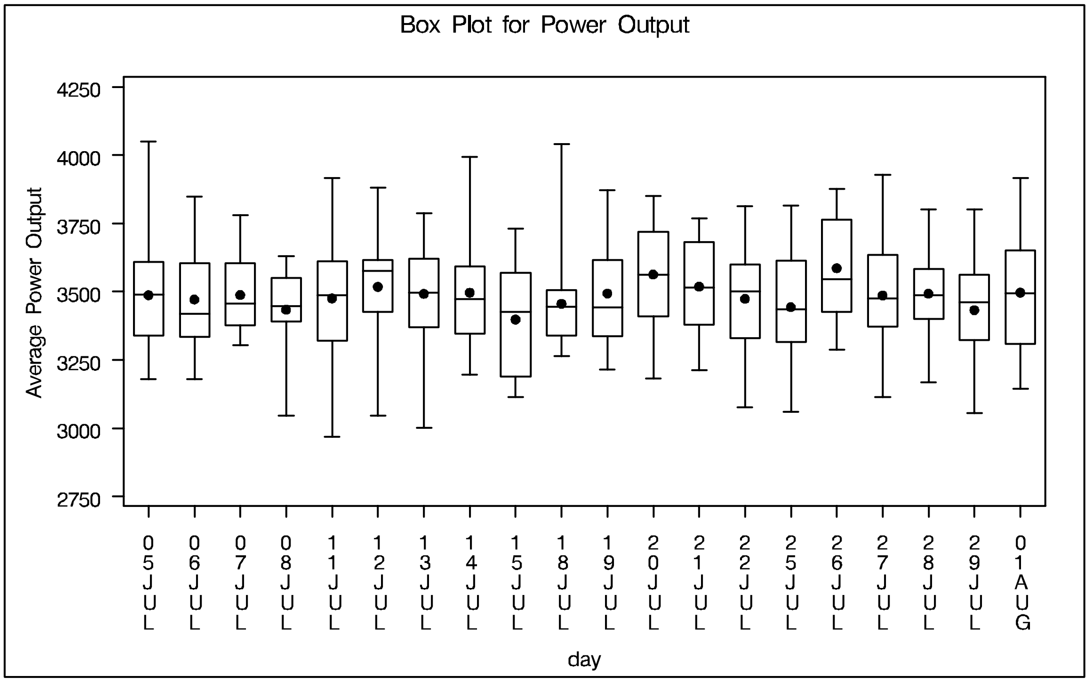


Figure 18.5. Box Plot with Discrete Group Variable

In order to treat the group variable as *continuous*, you can specify the CONTINUOUS or HAXIS= option. Either option produces a box plot with a horizontal axis scaled for continuous group variable values.

The following statements produce the plot shown in Figure 18.6. Note that the values on the horizontal axis represent consecutive days. Box-and-whisker plots are not produced for days when no turbine data was collected. The TOTPANELS= option is specified to display the entire box plot on one panel.

```
symbol v=dot c=salmon;
title 'Box Plot for Power Output';
axis1 minor=none color=black label=(angle=90 rotate=0);

proc boxplot data=turbine;
   plot kwatts*day / turnhlabel
                     cframe    = vligb
                     cboxes    = dagr
                     cboxfill  = ywh
                     totpanels = 1
                     vaxis     = axis1
                     continuous;

run;
```

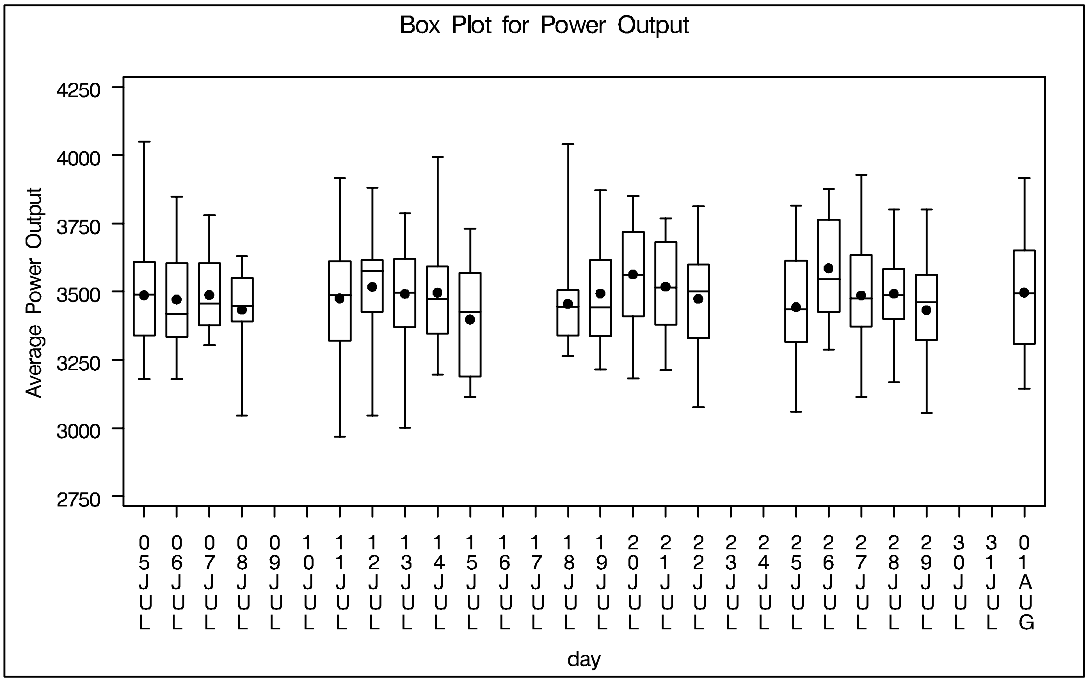


Figure 18.6. Box Plot with Continuous Group Variable

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 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 data in the Parts data set grouped into blocks by the *block-variable* Machine. The plot is shown in Figure 18.7.

```
symbol v=dot c=salmon;
goptions ftext=swiss;
axis1 minor=none color=black label=(angle=90 rotate=0);
title ’Box Plot for Diameter Grouped By Machine’;

proc boxplot data=Parts;
   plot Diam*Sample (Machine) / cframe = vligb vaxis = axis1;
```

```
      label Sample  = 'Sample Number'
            Machine = 'Machine'
            Diam    = 'Diameter' ;
   run;
```

The unique consecutive values of Machine (A386, A455, and C334) are displayed in a legend above the plot. Note the LABEL statement used to provide labels for the axes and for the block legend.

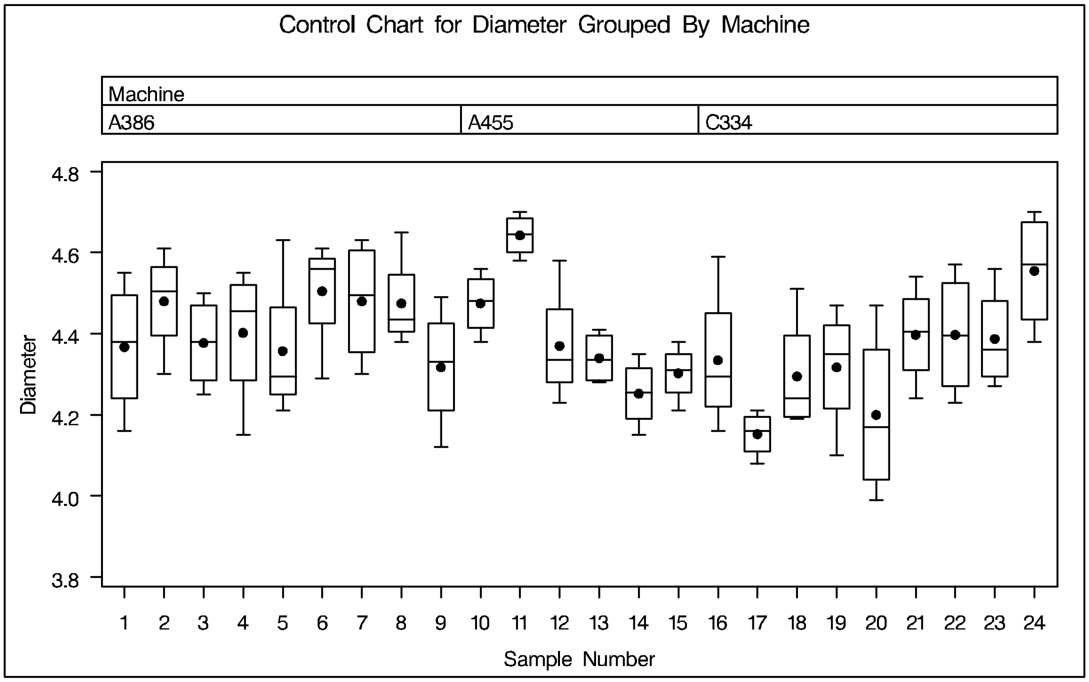


Figure 18.7. Box Plot Using a Block Variable

By default, the block legend is placed above the plot, as in Figure 18.7. You can control the position of the legend with the BLOCKPOS=*n* option; see the BLOCKPOS= option on page 403.

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. Choose BLOCKLABTYPE=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=*keyword* option. The valid keywords are ABOVE (the default, as shown in Figure 18.7) and LEFT.

Examples

This section provides advanced examples of the PLOT statement.

Example 18.1. Using Box Plots to Compare Groups

In the following example, a box plot is used to compare the delay times for 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;

data Times;
   merge Times cancel;
   by day;
run;
```

The following statements create a data set named Weather that contains information about possible causes for delays. This data set is merged 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;
```

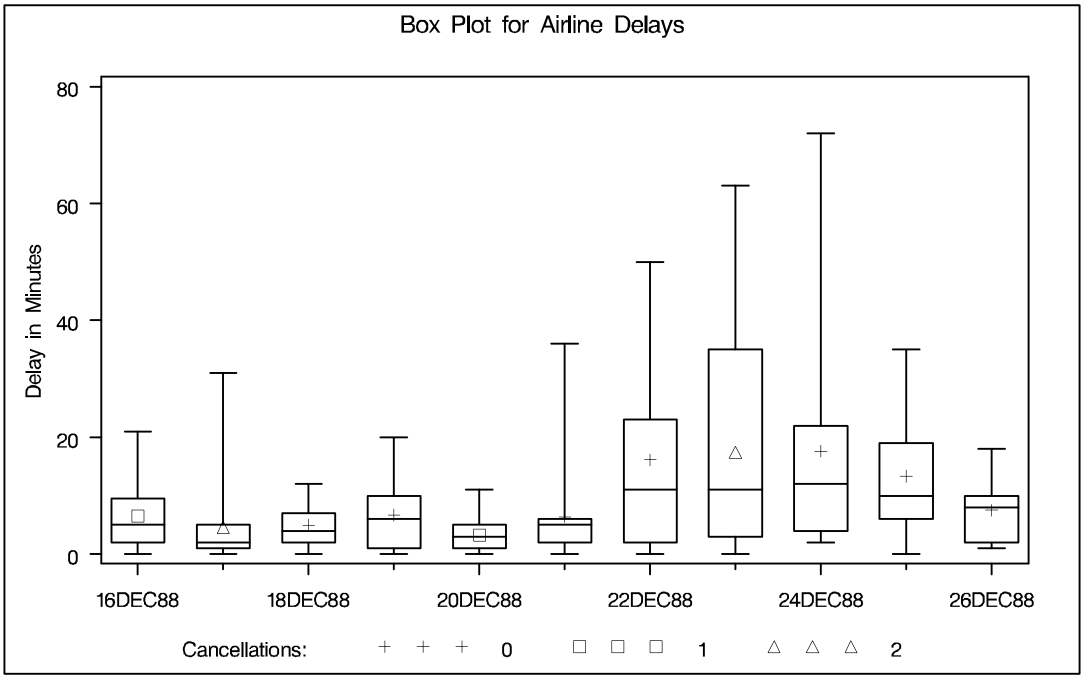
The following statements create a box plot for the complete set of data.

```
symbol1 v=plus     c=salmon;
symbol2 v=square   c=vigb;
symbol3 v=triangle c=vig;
goptions ftext=swiss;
axis1 minor=none color=black label=(angle=90 rotate=0);
title ’Box Plot for Airline Delays’;

proc boxplot data=times;
   plot delay * day = ncancel /
                      nohlabel
                      symbollegend = legend1
                      cboxes       = dagr
                      cboxfill     = ywh
                      cframe       = vligb
                      vaxis        = axis1;
   legend1 label=(’Cancellations:’)
           cborder=black cframe=ligr;
   label delay = ’Delay in Minutes’;
run;
```

The box plot is shown in Output 18.1.1. 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 label for the horizontal axis.

Output 18.1.1. 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 18.2. Creating Various Styles of Box-and-Whisker Plots

The following example uses the flight delay data of the preceding example to illustrate how you can create box plots with various styles of box-and-whisker plots. The following statements create a plot, shown in Output 18.2.1, that displays skeletal box-and-whisker plots:

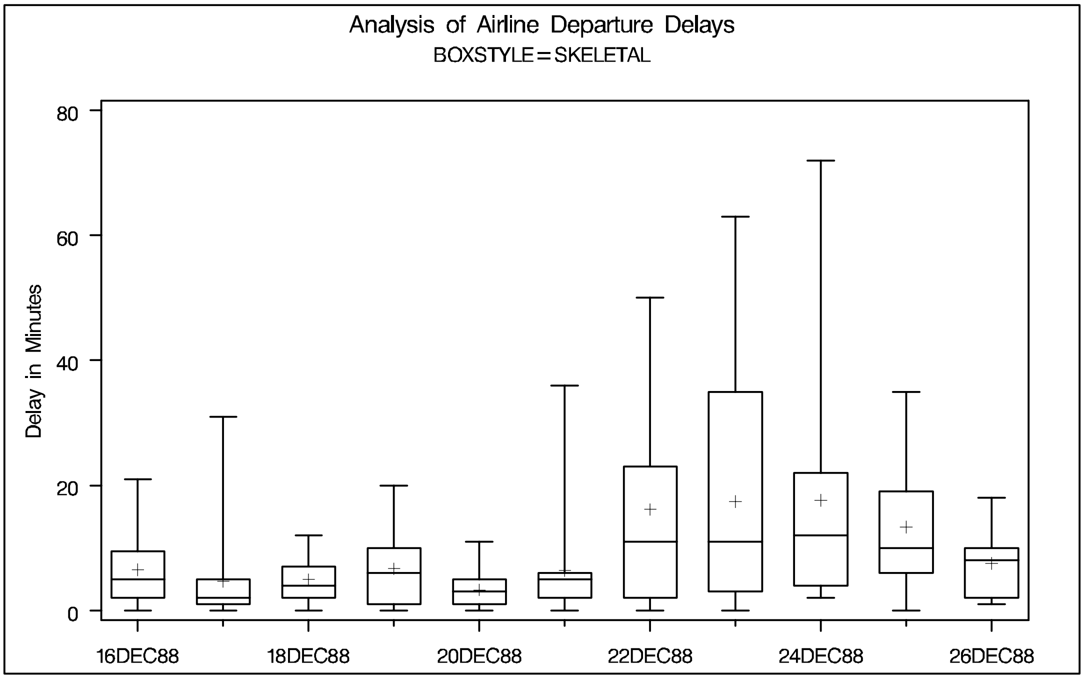
```
symbol v=plus c=salmon;
axis1 minor=none color=black label=(angle=90 rotate=0);
title ’Analysis of Airline Departure Delays’;
title2 ’BOXSTYLE=SKELETAL’;

proc boxplot data=times;
```

```
      plot delay * day /
                   boxstyle = skeletal
                   nohlabel
                   cframe   = vligb
                   cboxes   = dagr
                   cboxfill = ywh
                   vaxis    = axis1;
         label delay = 'Delay in Minutes';
   run;
```

In a skeletal box-and-whisker plot, the whiskers are drawn from the quartiles to the extreme values of the group. The skeletal box-and-whisker plot is the default style; consequently, you can also request this style by omitting the BOXSTYLE= option.

Output 18.2.1. BOXSTYLE=SKELETAL

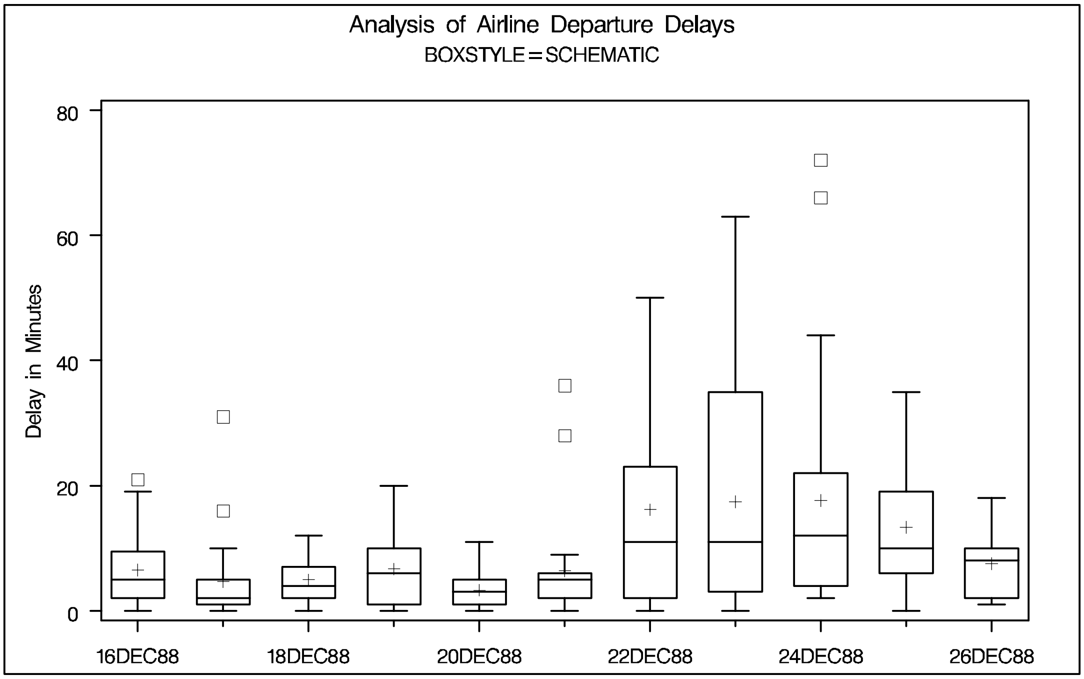


The following statements request a box plot with schematic box-and-whisker plots:

```
   title2 'BOXSTYLE=SCHEMATIC';
   proc boxplot data=times;
      plot delay * day /
                   boxstyle  = schematic
                   nohlabel
                   cframe   = vligb
                   cboxes   = dagr
                   cboxfill = ywh
                   idcolor  = salmon
                   vaxis    = axis1;
      label delay = 'Delay in Minutes';
   run;
```

The plot is shown in Output 18.2.2. When BOXSTYLE=SCHEMATIC is specified, the whiskers are drawn to the most extreme points in the group that lie within 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 on page 403.

Output 18.2.2. BOXSTYLE=SCHEMATIC

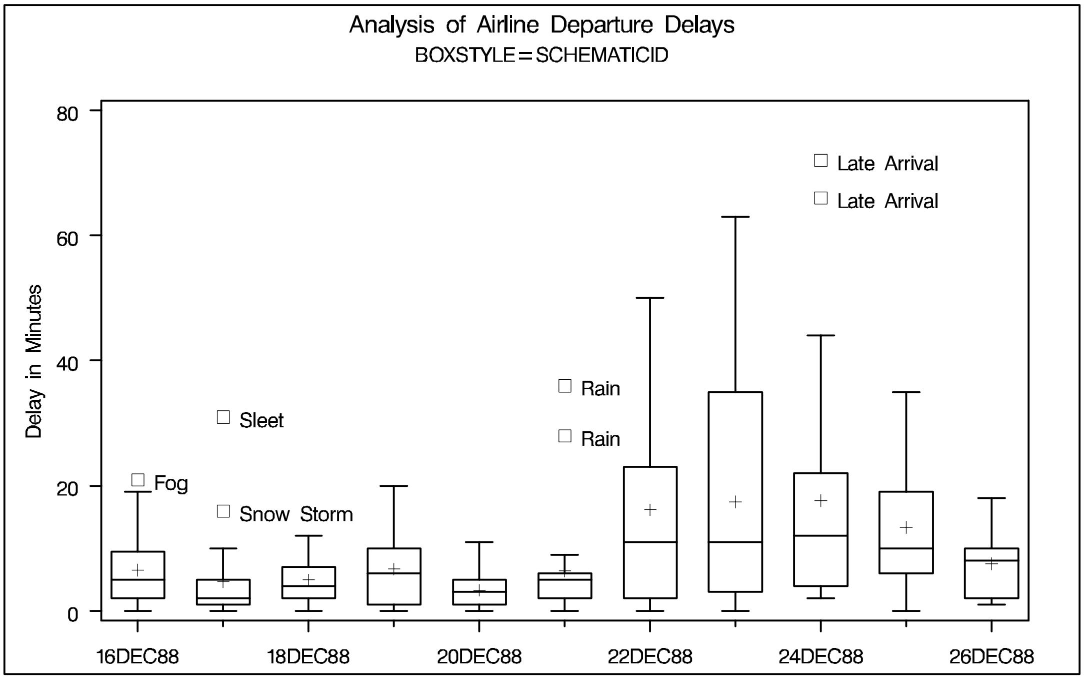


The following statements create a box plot with schematic box-and-whisker plots in which the observations outside the fences are labeled:

```
title2 'BOXSTYLE=SCHEMATICID';
proc boxplot data=times;
   plot delay * day /
                boxstyle = schematicid
                nohlabel
                cboxes   = dagr
                cboxfill = ywh
                cframe   = vligb
                idcolor  = salmon
                vaxis    = axis1;
   id reason;
   label delay = 'Delay in Minutes';
run;
```

The plot is shown in Output 18.2.3. If you specify BOXSTYLE=SCHEMATICID, schematic box-and-whisker plots are displayed in which the value of the first ID variable (in this case, reason) is used to label each observation outside the fences.

Output 18.2.3. BOXSTYLE=SCHEMATICID

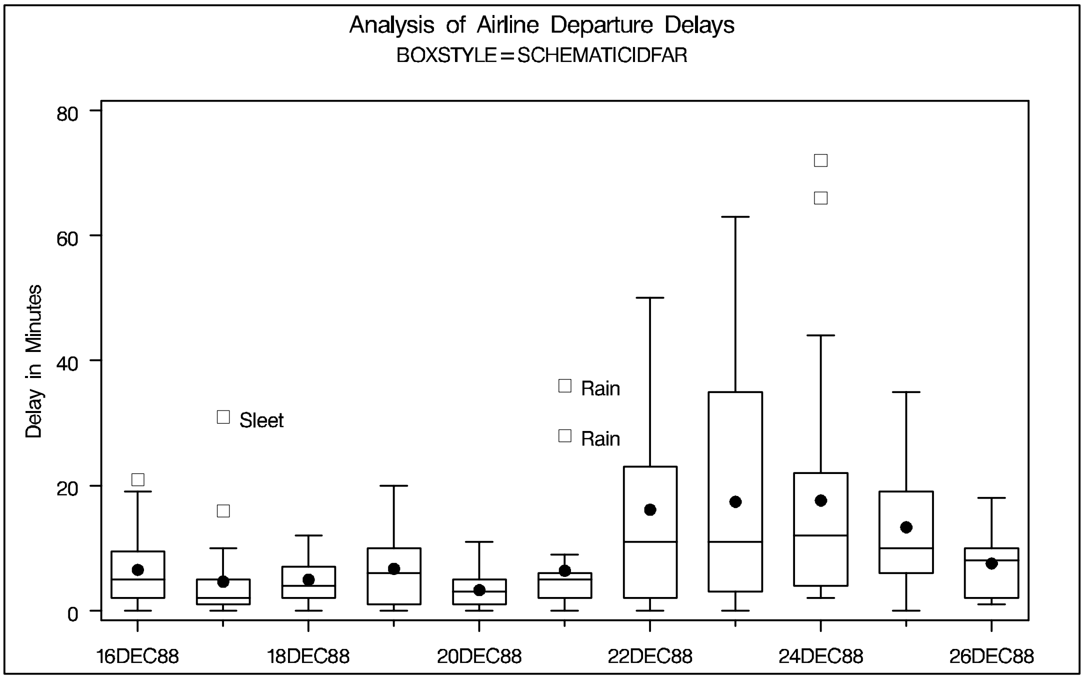


The following statements create a box plot with schematic box-and-whisker plots in which only the extreme observations outside the fences are labeled:

```
title2 ’BOXSTYLE=SCHEMATICIDFAR’;
proc boxplot data=times;
   plot delay * day /
                boxstyle = schematicidfar
                nohlabel
                cframe   = vligb
                cboxes   = dagr
                cboxfill = ywh
                idcolor  = salmon
                vaxis    = axis1;
   id reason;
   label delay   = ’Delay in Minutes’;
run;
```

The plot is shown in Output 18.2.4. If you specify BOXSTYLE=SCHEMATICIDFAR, schematic plots are displayed in which 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 above the 75th percentile, respectively. Observations between the fences and the far fences are identified with a symbol but are not labeled.

Output 18.2.4. BOXSTYLE=SCHEMATICIDFAR



Other options for controlling the display of box-and-whisker plots include the BOXWIDTH=, BOXWIDTHSCALE=, CBOXES=, CBOXFILL=, and LBOXES= options.

Example 18.3. Creating Notched Box-and-Whisker Plots

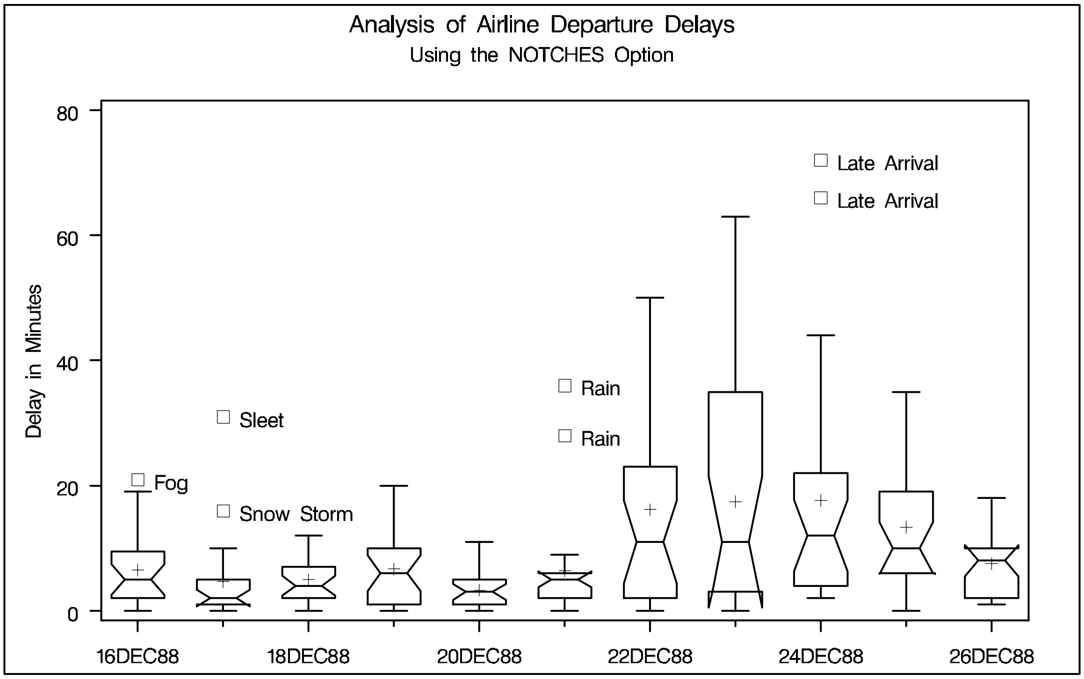
The following statements use the flight delay data of Example 18.1 to illustrate how to create side-by-side box-and-whisker plots with notches:

```
title 'Analysis of Airline Departure Delays';
title2 'Using the NOTCHES Option';
proc boxplot data=times;
   plot delay * day /
                boxstyle = schematicid
                cboxfill = ywh
                nohlabel
                notches
                cboxes   = dagr
                cframe   = vligb
                idcolor  = salmon
                vaxis    = axis1;
   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 0.05 level if the corresponding notches do not overlap.

For example, in Output 18.3.1, the median for December 20 is significantly different from the median for December 24.

Output 18.3.1. Notched Side-by-Side Box-and-Whisker Plots



Example 18.4. Creating Box-and-Whisker Plots with Varying Widths

The following example shows how to create a box plot with box-and-whisker 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
;
```

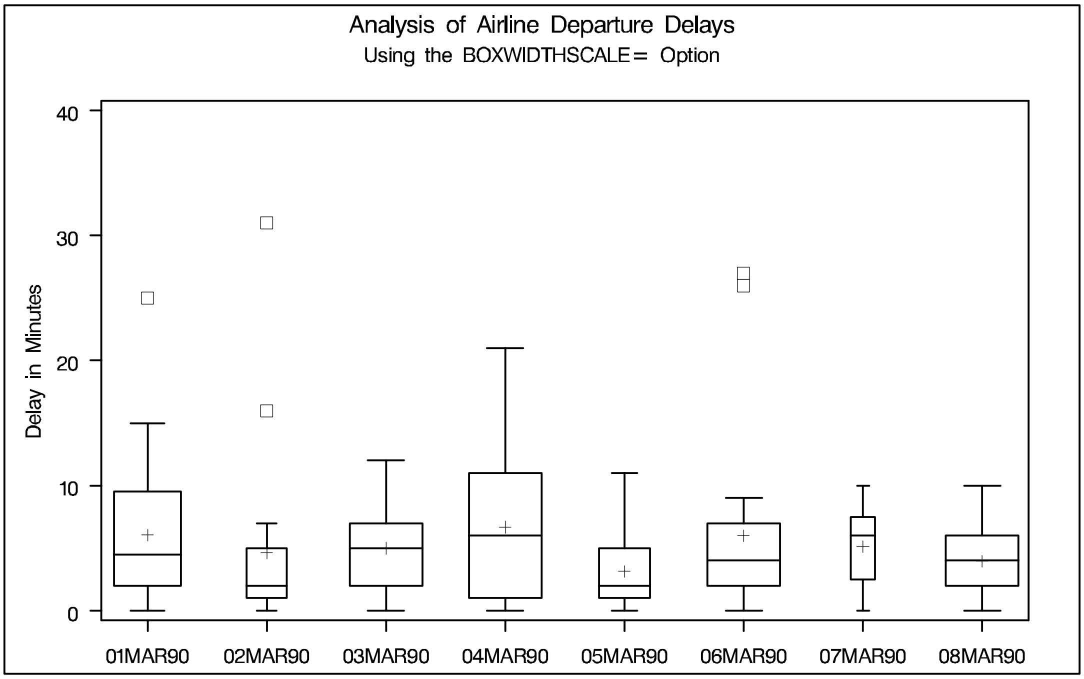
The following statements create the box plot shown in Output 18.4.1:

```
goptions ftext=swiss;
title ’Analysis of Airline Departure Delays’;
title2 ’Using the BOXWIDTHSCALE= Option’;

proc boxplot data=Times2;
   plot delay * day /
                boxwidthscale = 1
                boxstyle      = schematic
                nohlabel
                cframe        = vligb
                cboxes        = dagr
                cboxfill      = ywh
                idcolor       = salmon
                vaxis         = axis1;
run;
```

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

Output 18.4.1. Box Plot with Box-and-Whisker Plots of Varying Widths



References

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

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

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

SAS/STAT® User's Guide, Version 8

ISBN 1–58025–494–2

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

1st printing, October 1999

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