

Chapter 29

BOXCHART Statement

Chapter Table of Contents

OVERVIEW	1109
GETTING STARTED	1110
Creating Box Charts from Raw Data	1110
Creating Box Charts from Subgroup Summary Data	1113
Saving Summary Statistics	1116
Saving Control Limits	1118
Reading Preestablished Control Limits	1121
SYNTAX	1122
Summary of Options	1123
DETAILS	1133
Constructing Box Charts	1133
Output Data Sets	1136
ODS Tables	1140
Input Data Sets	1140
Methods for Estimating the Standard Deviation	1144
Percentile Definitions	1144
Axis Labels	1145
Missing Values	1145
EXAMPLES	1147
Example 29.1 Using Box Charts to Compare Subgroups	1147
Example 29.2 Creating Various Styles of Box-and-Whisker Plots	1149
Example 29.4 Creating Notched Box-and-Whisker Plots	1153
Example 29.4 Creating Box-and-Whisker Plots with Varying Widths	1154
Example 29.5 Creating Box-and-Whisker Plots with Different Line Styles and Colors	1156
Example 29.6 Computing the Control Limits for Subgroup Maximums	1157
Example 29.7 Constructing Multi-Vari Charts	1160

Chapter 29

BOXCHART Statement

Overview

The BOXCHART statement creates an \bar{X} chart for subgroup means superimposed with box-and-whisker plots of the measurements in each subgroup. Throughout this chapter, a chart of this type is referred to as a *box chart*. This chart is recommended for large subgroup sample sizes (typically greater than ten). You can also use the BOXCHART statement to create standard side-by-side box-and-whisker plots (see Example 29.2 on page 1149 and Example 29.3.1 on page 1154).

You can use options in the BOXCHART statement to

- specify control limits for subgroup means or medians
- compute control limits from the data based on a multiple of the standard error of the means (or medians) or as probability limits
- tabulate subgroup summary statistics and control limits
- save control limits in an output data set
- save subgroup summary statistics in an output data set
- read preestablished control limits from a data set
- apply tests for special causes (also known as runs tests and Western Electric rules)
- specify one of several methods for estimating the process standard deviation
- specify whether subgroup standard deviations or subgroup ranges are used to estimate the process standard deviation
- specify a known (standard) process mean and standard deviation for computing control limits
- create a secondary chart that displays a time trend removed from the data (see “Displaying Trends in Process Data” on page 1838)
- specify one of several methods for calculating quantile statistics (percentiles)
- control the style of the box-and-whisker plots
- display distinct sets of control limits for data from successive time phases
- add block legends and symbol markers to reveal stratification in process data
- clip extreme points to make the chart more readable
- display vertical and horizontal reference lines
- control axis values and labels
- control layout and appearance of the chart

Getting Started

This section introduces the BOXCHART statement with simple examples that illustrate commonly used options. Complete syntax for the BOXCHART statement is presented in the “Syntax” section on page 1122, and advanced examples are given in the “Examples” section on page 1147.

Creating Box Charts from Raw Data

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 that contains the power output measurements for 20 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;
  cards;
04JUL94 3196 3507 4050 3215 3583 3617 3789 3180 3505 3454
04JUL94 3417 3199 3613 3384 3475 3316 3556 3607 3364 3721
05JUL94 3390 3562 3413 3193 3635 3179 3348 3199 3413 3562
05JUL94 3428 3320 3745 3426 3849 3256 3841 3575 3752 3347
06JUL94 3478 3465 3445 3383 3684 3304 3398 3578 3348 3369
06JUL94 3670 3614 3307 3595 3448 3304 3385 3499 3781 3711
.
.
.
23JUL94 3421 3787 3454 3699 3307 3917 3292 3310 3283 3536
23JUL94 3756 3145 3571 3331 3725 3605 3547 3421 3257 3574
;

```

A partial listing of TURBINE is shown in Figure 29.1. This data set is said to be in “strung-out” form since each observation contains the day and 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 subgroups, it is referred to as the *subgroup-variable*. The variable KWATTS contains the output measurements and is referred to as the *process variable* (or *process* for short).

Kilowatt Power Output Data		
Obs	day	kwatts
1	04JUL	3196
2	04JUL	3507
3	04JUL	4050
4	04JUL	3215
5	04JUL	3583
.	.	.
.	.	.
.	.	.
396	23JUL	3605
397	23JUL	3547
398	23JUL	3421
399	23JUL	3257
400	23JUL	3574

Figure 29.1. Partial Listing of the Data Set TURBINE

You can use a box chart to examine the distribution of power output for each day and to determine whether the mean level of the heating process is in control. The following statements create the box chart shown in Figure 29.2:

```

title 'Box Chart for Power Output';
symbol v=dot;
proc shewhart data=turbine;
  boxchart kwatts*day;
run;

```

This example illustrates the basic form of the BOXCHART statement. After the keyword BOXCHART, you specify the *process* to analyze (in this case, KWATTS), followed by an asterisk and the *subgroup-variable* (DAY).

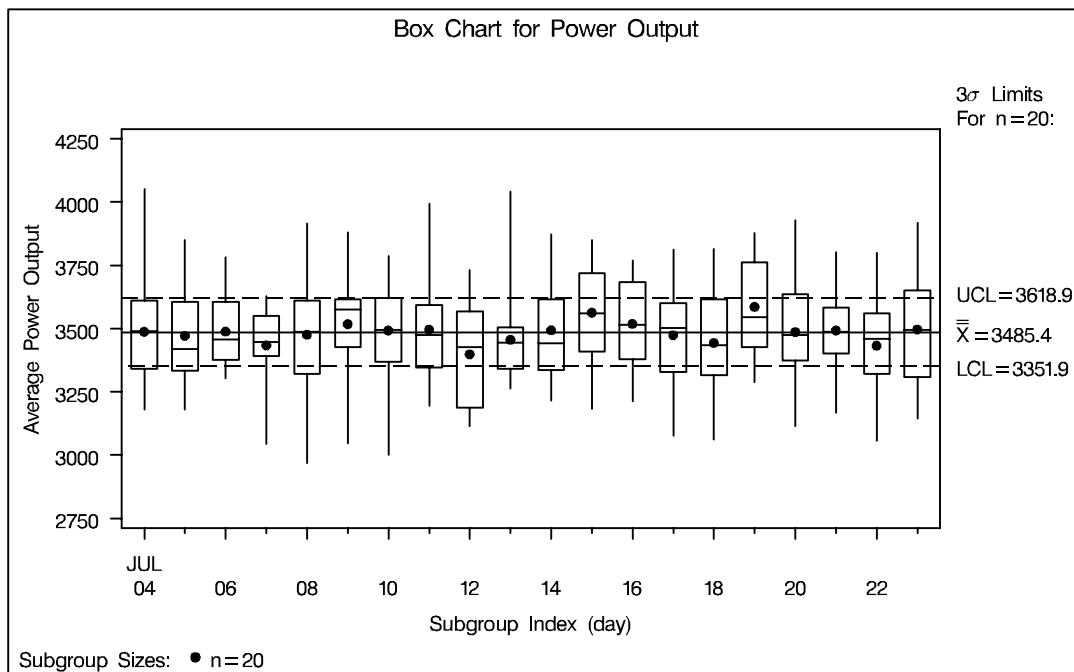


Figure 29.2. Box Chart for Power Output Data

The input data set is specified with the DATA= option in the PROC SHEWHART statement.

By default, the BOXCHART statement requests an \bar{X} chart superimposed with box-and-whisker plots for each subgroup. Table 29.1 lists the summary statistics represented by each plot. For details on the computation of percentiles, see “Percentile Definitions” on page 1144.

Table 29.1. Summary Statistics Represented by Box-and-Whisker Plots

Subgroup Summary Statistic	Feature of Box-and-Whisker Plot
Maximum	Endpoint of upper whisker
Third quartile (75 th percentile)	Upper edge of box
Median (50 th percentile)	Line inside box
Mean	Symbol marker (in this example, a dot)
First quartile (25 th percentile)	Lower edge of box
Minimum	Endpoint of lower whisker

The within-subgroup variation in power output is stable, as indicated in Figure 29.2 by the edges of the boxes and the endpoints of the whiskers. Since the subgroup means, indicated by the dots, lie within the control limits, you can conclude that the heating process is in statistical control.

The skeletal style of the box-and-whisker plots shown in Figure 29.2 is the default. You can request different styles, as illustrated in Example 29.2 on page 1149. By default, the control limits shown are 3σ limits estimated from the data; the formulas for the limits are given in Table 29.22 on page 1134 and Table 29.23 on page 1134.

You can also create box charts in which the control limits apply to the subgroup medians. For example, the following statements create the chart shown in Figure 29.3:

```

title 'Box Chart for Power Output';
symbol v=dot;
proc shewhart data=turbine;
    boxchart kwatts*day / controlstat=median;
run;

```

The CONTROLSTAT=MEDIAN option requests control limits that apply to the medians. Alternatively, you can specify the NOLIMITS option to suppress the display of control limits and create ordinary side-by-side box-and-whisker plots. See Example 29.2 on page 1149.

Options such as CONTROLSTAT= and NOLIMITS are specified after the slash (/) in the BOXCHART statement. A complete list of options is presented in the “Syntax” section on page 1122.

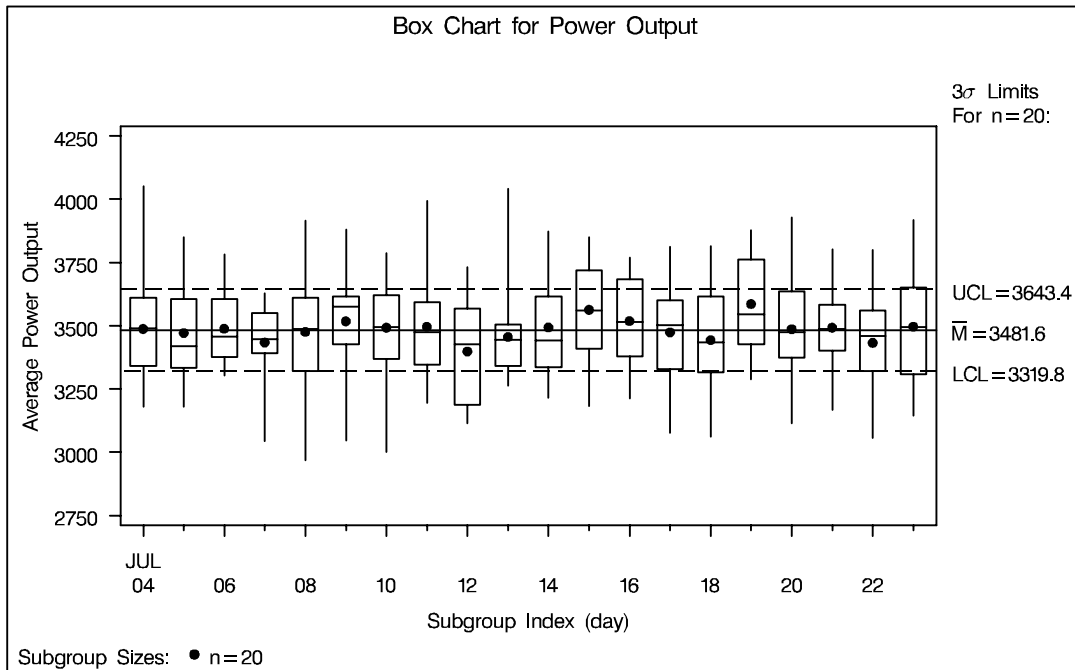


Figure 29.3. Box Chart for Power Output Data

Creating Box Charts from Subgroup Summary Data

The previous example illustrates how you can create box charts using raw data (process measurements). However, in many applications the data are provided as subgroup summary statistics. This example illustrates how you can use the BOXCHART statement with data of this type.

See SHWBOXA
in the SAS/QC
Sample Library

The following data set (OILSUM) provides the data from the preceding example in summarized form. There is exactly one observation for each subgroup (note that the subgroups are still indexed by DAY).

```
data oilsum;
  input day kwatts1 kwatts1 kwattsx kwattsm
        kwatts3 kwattsh kwattsr kwattsn;
  informat day date7. ;
  format day date5. ;
  label day      = 'Date of Measurement'
        kwatts1 = 'Minimum Power Output'
        kwatts1 = '25th Percentile'
        kwattsx = 'Average Power Output'
        kwattsm = 'Median Power Output'
        kwatts3 = '75th Percentile'
        kwattsh = 'Maximum Power Output'
        kwattsr = 'Range of Power Output'
        kwattsn = 'Subgroup Sample Size';
  cards;
04JUL94 3180 3340.0 3487.40 3490.0 3610.0 4050 870 20
```

```

05JUL94 3179 3333.5 3471.65 3419.5 3605.0 3849 670 20
06JUL94 3304 3376.0 3488.30 3456.5 3604.5 3781 477 20
07JUL94 3045 3390.5 3434.20 3447.0 3550.0 3629 584 20
08JUL94 2968 3321.0 3475.80 3487.0 3611.5 3916 948 20
09JUL94 3047 3425.5 3518.10 3576.0 3615.0 3881 834 20
10JUL94 3002 3368.5 3492.65 3495.5 3621.5 3787 785 20
11JUL94 3196 3346.0 3496.40 3473.5 3592.5 3994 798 20
12JUL94 3115 3188.5 3398.50 3426.0 3568.5 3731 616 20
13JUL94 3263 3340.0 3456.05 3444.0 3505.5 4040 777 20
14JUL94 3215 3336.0 3493.60 3441.5 3616.0 3872 657 20
15JUL94 3182 3409.5 3563.30 3561.0 3719.5 3850 668 20
16JUL94 3212 3378.0 3519.05 3515.0 3682.5 3769 557 20
17JUL94 3077 3329.0 3474.20 3501.5 3599.5 3812 735 20
18JUL94 3061 3315.5 3443.60 3435.0 3614.5 3815 754 20
19JUL94 3288 3426.5 3586.35 3546.0 3762.5 3877 589 20
20JUL94 3114 3373.0 3486.45 3474.5 3635.5 3928 814 20
21JUL94 3167 3400.5 3492.90 3488.0 3582.5 3801 634 20
22JUL94 3056 3322.0 3432.80 3460.0 3561.0 3800 744 20
23JUL94 3145 3308.5 3496.90 3495.0 3652.0 3917 772 20
;

```

A partial listing of OILSUM is shown in Figure 29.4.

Summary Data Set for Power Outputs								
day	kwattsl	kwatts1	kwattsx	kwattsm	kwatts3	kwattsh	kwattsr	kwattsn
04JUL	3180	3340.0	3487.40	3490.0	3610.0	4050	870	20
05JUL	3179	3333.5	3471.65	3419.5	3605.0	3849	670	20
06JUL	3304	3376.0	3488.30	3456.5	3604.5	3781	477	20
07JUL	3045	3390.5	3434.20	3447.0	3550.0	3629	584	20
08JUL	2968	3321.0	3475.80	3487.0	3611.5	3916	948	20
.
.
.

Figure 29.4. The Summary Data Set OILSUM

There are eight summary variables in OILSUM.

- KWATTSL contains the subgroup minimums (low values).
- KWATTS1 contains the 25th percentile (first quartile) for each subgroup.
- KWATTSX contains the subgroup means.
- KWATTSM contains the subgroup medians.
- KWATTS3 contains the 75th percentile (third quartile) for each subgroup.
- KWATTSH contains the subgroup maximums (high values).
- KWATTSR contains the subgroup ranges.
- KWATTSN contains the subgroup sample sizes.

You can read this data set by specifying it as a HISTORY= data set in the PROC SHEWHART statement, as illustrated by the following statements, which create the box chart shown in Figure 29.5:

```

title 'Box Chart for Power Output';
symbol v=dot;

```

```
proc shewhart history=oilsum;
  boxchart kwatts*day;
run;
```

Note that the *process* KWATTS is *not* the name of a SAS variable in the data set but is, instead, the common prefix for the names of the eight summary variables. The suffix characters L, 1, X, M, 3, H, R, and N indicate the contents of the variable. For example, the suffix characters 1 and 3 indicate first and third quartiles. The name DAY specified after the asterisk is the name of the *subgroup-variable*.

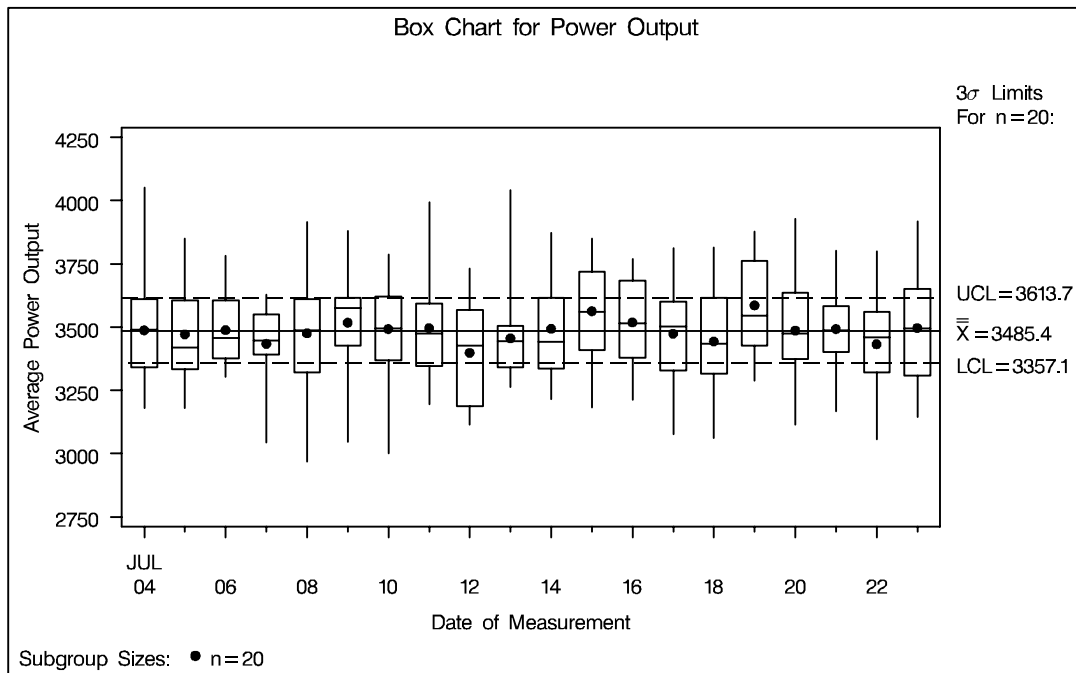


Figure 29.5. Box Chart for Power Output Data

In general, a HISTORY= input data set used with the BOXCHART statement must contain the following variables:

- subgroup variable
- subgroup minimum variable
- subgroup first quartile variable
- subgroup mean variable
- subgroup median variable
- subgroup third quartile variable
- subgroup maximum variable
- subgroup sample size variable
- either a subgroup range variable or a subgroup standard deviation variable

Furthermore, the names of the summary variables must begin with the *process* name specified in the BOXCHART statement and end with the appropriate suffix character. If the names do not follow this convention, you can use the RENAME option in the PROC SHEWHART statement to rename the variables for the duration of the SHEWHART procedure step (see page 1616).

If you specify the STDDEVIATIONS option in the BOXCHART statement, the HISTORY= data set must contain a subgroup standard deviation variable; otherwise, the HISTORY= data set must contain a subgroup range variable. The STDDEVIATIONS option specifies that the estimate of the process standard deviation σ is to be calculated from subgroup standard deviations rather than subgroup ranges. For example, in the following statements, the data set OILSUM2 must contain a subgroup standard deviation variable named KWATTSS:

```
title 'Box Chart for Power Output';
symbol v=dot;
proc shewhart history=oilsum2;
    boxchart kwatts*day / stddeviations;
run;
```

In summary, the interpretation of *process* depends on the input data set.

- If raw data are read using the DATA= option (as in the previous example), *process* is the name of the SAS variable containing the process measurements.
- If summary data are read using the HISTORY= option (as in this example), *process* is the common prefix for the names of the variables containing the summary statistics.

For more information, see “HISTORY= Data Set” on page 1141.

Saving Summary Statistics

In this example, the BOXCHART statement is used to create a summary data set that can be read later by the SHEWHART procedure (as in the preceding example). The following statements read measurements from the data set TURBINE and create a summary data set named TURBHIST:

```
title 'Summary Data Set for Power Output';
proc shewhart data=turbine;
    boxchart kwatts*day / outhistory = turbhist
                    nochart;
run;
```

The OUTHISTORY= option names the output data set, and the NOCHART option suppresses the display of the chart, which would be identical to the chart in Figure 29.2.

Figure 29.6 contains a partial listing of TURBHIST.

See SHWBOXA
in the SAS/QC
Sample Library

Summary Data Set for Power Output								
day	kwattsL	kwatts1	kwattsX	kwattsM	kwatts3	kwattsH	kwattsR	kwattsN
04JUL	3180	3340.0	3487.40	3490.0	3610.0	4050	870	20
05JUL	3179	3333.5	3471.65	3419.5	3605.0	3849	670	20
06JUL	3304	3376.0	3488.30	3456.5	3604.5	3781	477	20
07JUL	3045	3390.5	3434.20	3447.0	3550.0	3629	584	20
08JUL	2968	3321.0	3475.80	3487.0	3611.5	3916	948	20
.
.
.

Figure 29.6. The Summary Data Set TURBHIST

There are nine variables in the data set TURBHIST.

- DAY is the subgroup variable.
- KWATTSL contains the subgroup minimums.
- KWATTS1 contains the first quartiles for each subgroup.
- KWATTSX contains the subgroup means.
- KWATTSM contains the subgroup medians.
- KWATTS3 contains the third quartiles for each subgroup.
- KWATTSH contains the subgroup maximums.
- KWATTSR contains the subgroup ranges.
- KWATTSN contains the subgroup sample sizes.

Note that the summary statistic variables are named by adding the suffix characters *L*, *I*, *X*, *M*, *3*, *H*, *R*, and *N* to the *process* KWATTS specified in the BOXCHART statement. In other words, the variable naming convention for OUTHISTORY= data sets is the same as that for HISTORY= data sets.

If you specify the STDDEVIATIONS option, the OUTHISTORY= data set includes a subgroup standard deviation variable, rather than a subgroup range variable, as demonstrated by the following statements:

```

title 'Summary Data Set for Power Output';
proc shewhart data=turbine;
    boxchart kwatts*day / outhistory = turbhst2
                    stddeviations
                    nochart;
run;

```

Figure 29.7 contains a partial listing of TURBHST2. The variable KWATTSS contains the subgroup standard deviations.

The STDDEVIATIONS option is recommended when the subgroup sample sizes are greater than 10, and it is also recommended when you use the NOLIMITS option to create standard side-by-side box-and-whisker plots.

For more information, see “OUTHISTORY= Data Set” on page 1137.

Summary Data Set for Power Output								
day	kwattsL	kwatts1	kwattsX	kwattsM	kwatts3	kwattsH	kwattsS	kwattsN
04JUL	3180	3340.0	3487.40	3490.0	3610.0	4050	220.260	20
05JUL	3179	3333.5	3471.65	3419.5	3605.0	3849	210.427	20
06JUL	3304	3376.0	3488.30	3456.5	3604.5	3781	147.025	20
07JUL	3045	3390.5	3434.20	3447.0	3550.0	3629	157.637	20
08JUL	2968	3321.0	3475.80	3487.0	3611.5	3916	258.949	20
.
.
.

Figure 29.7. The Summary Data Set TURBHST2

Saving Control Limits

You can save the control limits for a box chart in a SAS data set; this enables you to apply the control limits to future data (see “Reading Prestablished Control Limits” on page 1121) or modify the limits with a DATA step program.

The following statements read measurements from the data set TURBINE (see page 1110) and save the control limits displayed in Figure 29.2 in a data set named TURBLIM:

```

title 'Control Limits for Power Output Data';
proc shewhart data=turbine;
  boxchart kwatts*day / outlimits=turblim
  nochart;
run;

```

The OUTLIMITS= option names the data set containing the control limits, and the NOCHART option suppresses the display of the chart. The data set TURBLIM is listed in Figure 29.8.

Control Limits for Power Output Data						
VAR	_SUBGRP_	_TYPE_	_LIMITN_	_ALPHA_	_SIGMAS_	_LCLX_
kwatts	day	ESTIMATE	20	.002699796	3	3357.14
MEAN	_UCLX_	_LCLR_	_R_	_UCLR_	_STDDEV_	
3485.41	3613.68	296.159	714.15	1132.14	191.207	

Figure 29.8. The Data Set TURBLIM Containing Control Limit Information

The data set TURBLIM contains one observation with the limits for *process* KWATTS. The variables `_LCLX_` and `_UCLX_` contain the lower and upper control limits for the means, and the variable `_MEAN_` contains the central line. The value of `_MEAN_` is an estimate of the process mean, and the value of `_STDDEV_` is an estimate of the process standard deviation σ . The value of `_LIMITN_` is the nominal sample size associated with the control limits, and the value of `_SIGMAS_` is the multiple of σ associated with the control limits. The variables `_VAR_` and `_SUBGRP_` are bookkeeping variables that save the *process* and *subgroup-variable*. The variable `_TYPE_` is a bookkeeping variable that indicates whether the values of `_MEAN_` and `_STDDEV_` are estimates or standard values.

The variables `_LCLR_`, `_R_`, and `_UCLR_` are not used to create box charts, but

See SHWBOXA
in the SAS/QC
Sample Library

they are included so that the data set TURBLIM can be used to create an R chart; see Chapter 40, “XRCHART Statement.”. If you specify the STDDEVIATIONS option in the BOXCHART statement, the variables _LCLS_, _S_, and _UCLS_, rather than the variables _LCLR_, _R_, and _UCLR_, are included in the OUTLIMITS= data set. These variables can be used to create an s chart; see Chapter 41, “XSCHART Statement.”.

If you specify CONTROLSTAT=MEDIAN to request control limits for medians, the variables _LCLM_ and _UCLM_, rather than the variables _LCLX_ and _UCLX_, are included in the OUTLIMITS= data set as demonstrated by the following statements:

```

title 'Control Limits for Power Output Data';
proc shewhart data=turbine;
  boxchart kwatts*day / outlimits =turblim2
                controlstat=median
                stddeviations
                nochart;
run;

```

Since the STDDEVIATIONS option is specified, the variables _LCLS_, _S_, and _UCLS_ are included in TURBLIM2, which is listed in Figure 29.9.

For more information, see “OUTLIMITS= Data Set” on page 1136.

Control Limits for Power Output Data						
VAR	_SUBGRP_	_TYPE_	_LIMITN_	_ALPHA_	_SIGMAS_	_LCLM_
kwatts	day	ESTIMATE	20	.002776264	3	3319.85
MEAN	_UCLM_	_LCLS_	_S_	_UCLS_	_STDDEV_	
3481.63	3643.40	100.207	196.396	292.584	198.996	

Figure 29.9. The Data Set TURBLIM2 Containing Control Limit Information

You can create an output data set containing both control limits and summary statistics with the OUTTABLE= option, as illustrated by the following statements:

```

title 'Summary Statistics and Control Limit Information';
proc shewhart data=turbine;
  boxchart kwatts*day / outtable=turbtab
                nochart;
run;

```

The data set TURBTAB is partially listed in Figure 29.10.

Summary Statistics and Control Limit Information							
VAR	day	_SIGMAS_	_LIMITN_	_SUBN_	_LCLX_	_SUBX_	_MEAN_
kwatts	04JUL	3	20	20	3357.14	3487.40	3485.41
kwatts	05JUL	3	20	20	3357.14	3471.65	3485.41
kwatts	06JUL	3	20	20	3357.14	3488.30	3485.41
kwatts	07JUL	3	20	20	3357.14	3434.20	3485.41
.
.
.
UCLX	_EXLIM_	_SUBMIN_	_SUBQ1_	_SUBMED_	_SUBQ3_	_SUBMAX_	
3613.68		3180	3340.0	3490.0	3610.0	4050	
3613.68		3179	3333.5	3419.5	3605.0	3849	
3613.68		3304	3376.0	3456.5	3604.5	3781	
3613.68		3045	3390.5	3447.0	3550.0	3629	
.
.
.

Figure 29.10. The OUTTABLE= Data Set TURBTAB

This data set contains one observation for each subgroup sample. The variable `_SUBMIN_` contains the subgroup minimums, and the variable `_SUBQ1_` contains the first quartile for each subgroup. The variable `_SUBX_` contains the subgroup means, and the variable `_SUBMED_` contains the subgroup medians. The variable `_SUBQ3_` contains the third quartiles, and the variable `_SUBMAX_` contains the subgroup maximums. The variable `_SUBN_` contains the subgroup sample sizes. The variables `_LCLX_` and `_UCLX_` contain the lower and upper control limits for the means. The variable `_MEAN_` contains the central line. The variables `_VAR_` and `DAY` contain the *process* name and values of the *subgroup-variable*, respectively. For more information, see “OUTTABLE= Data Set” on page 1138.

An OUTTABLE= data set can be read later as a TABLE= data set. For example, the following statements read TURBTAB and display a box chart (not shown here) identical to the chart in Figure 29.2:

```

title 'Box Chart for Power Output';
symbol v=dot;
proc shewhart table=turbtab;
    boxchart kwatts*day;
    label _SUBX_ = 'Average Power Output';
run;

```

Because the SHEWHART procedure simply displays the information in a TABLE= data set, you can use TABLE= data sets to create specialized control charts (see Chapter 46, “Specialized Control Charts,”).

For more information, see “TABLE= Data Set” on page 1143.

Reading Prestablished Control Limits

In the previous example, the OUTLIMITS= data set TURBLIM saved control limits computed from the measurements in TURBINE. This example shows how these limits can be applied to new data. The following statements create the box chart in Figure 29.11 using new measurements in a data set named TURBINE2 (not listed here) and the control limits in TURBLIM:

```
title 'Box Chart for Power Output';
symbol v=dot;
proc shewhart data=turbine2 limits=turblim;
  boxchart kwatts*day;
run;
```

The LIMITS= option in the PROC SHEWHART statement specifies the data set containing the control limits. By default,* this information is read from the first observation in the LIMITS= data set for which

- the value of `_VAR_` matches the *process* name KWATTS
- the value of `_SUBGRP_` matches the *subgroup-variable* name DAY

The chart reveals an increase in variability beginning on August 1.

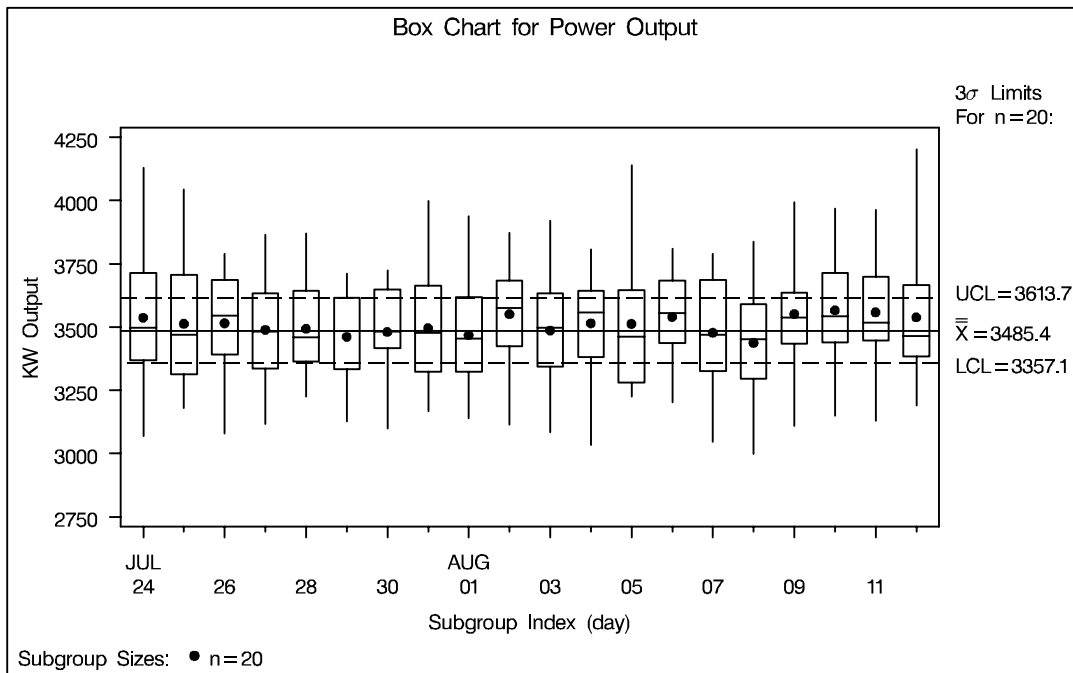


Figure 29.11. Box Chart for Second Set of Power Outputs

In this example, the LIMITS= data set was created in a previous run of the SHEWHART procedure. You can also create a LIMITS= data set with the DATA step. See “LIMITS= Data Set” on page 1141 for details concerning the variables that you must provide.

*In Release 6.09 and in earlier releases, it is also necessary to specify the READLIMITS option to read control limits from a LIMITS= data set.

Syntax

The basic syntax for the BOXCHART statement is as follows:

```
BOXCHART process*subgroup-variable ;
```

The general form of this syntax is as follows:

```
BOXCHART (processes)*subgroup-variable <(block-variables) >  
          <=symbol-variable | ='character' > <| options >;
```

You can use any number of BOXCHART statements in the SHEWHART procedure. The components of the BOXCHART statement are described as follows.

process

processes

identify one or more processes to be analyzed. The specification of *process* depends on the input data set specified in the PROC SHEWHART statement.

- If raw data are read from a DATA= data set, *process* must be the name of the variable containing the raw measurements. For an example, see “Creating Box Charts from Raw Data” on page 1110.
- If summary data are read from a HISTORY= data set, *process* must be the common prefix of the summary variables in the HISTORY= data set. For an example, see “Creating Box Charts from Subgroup Summary Data” on page 1113.
- If summary data and control limits are read from a TABLE= data set, *process* must be the value of the variable _VAR_ in the TABLE= data set. For an example, see “Saving Control Limits” on page 1118.

A *process* is required. If you specify more than one *process*, enclose the list in parentheses. For example, the following statements request distinct box charts for WEIGHT, LENGTH, and WIDTH:

```
proc shewhart data=summary;  
  boxchart (weight length width)*day;  
run;
```

subgroup-variable

is the variable that identifies subgroups in the data. The *subgroup-variable* is required. In the preceding BOXCHART statement, DAY is the subgroup variable. For details, see “Subgroup Variables” on page 1646.

block-variables

are optional variables that group the data into blocks of consecutive subgroups. These blocks are labeled in a legend, and each *block-variable* provides one level of labels in the legend. See “Displaying Stratification in Blocks of Observations” on page 1809 for an example.

symbol-variable

is an optional variable whose levels (unique values) determine the symbol marker or character used to plot the means.

- If you produce a chart on a line printer, an ‘A’ is displayed for the points corresponding to the first level of the *symbol-variable*, a ‘B’ is displayed for the points corresponding to the second level, and so on.
- If you produce a chart on a graphics device, 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 “Displaying Stratification in Levels of a Classification Variable” on page 1807 for an example.

character

specifies a plotting character for charts produced on line printers. For example, the following statements create a box chart using an asterisk (*) to plot the means:

```
proc shewhart data=values;
    boxchart weight*day='*';
run;
```

options

enhance the appearance of the box chart, request additional analyses, save results in data sets, and so on. The “Summary of Options” section, which follows, lists all options by function. Chapter 43, “Dictionary of Options,” describes each option in detail.

Summary of Options

The following tables list the BOXCHART statement options by function. For complete descriptions, see Chapter 43, “Dictionary of Options.”

Table 29.2. Options for Controlling Box Appearance

BOXCONNECT	connects subgroup means in box-and-whisker plots
BOXCONNECT= <i>keyword</i>	connects subgroup means, medians, maximum values, minimum values, or quartiles in box-and-whisker plots
BOXSTYLE= <i>keyword</i>	specifies style of box-and-whisker plots
BOXWIDTH= <i>value</i>	specifies width of box-and-whisker plots
BOXWIDTHSCALE= <i>value</i>	specifies that widths of box-and-whisker plots vary proportionately to subgroup sample size
CBOXES= <i>color</i> (<i>variable</i>)	specifies color for outlines of box-and-whisker plots
CBOXFILL= <i>color</i> (<i>variable</i>)	specifies fill color for interior of box-and-whisker plots
IDCOLOR= <i>color</i>	specifies outlier symbol color in schematic box-and-whisker plots
IDCTEXT= <i>color</i>	specifies text color to label outliers or process variable values
IDFONT= <i>font</i>	specifies text font to label outliers or process variable values
IDHEIGHT= <i>value</i>	specifies text height to label outliers or process variable values
IDSYMBOL= <i>symbol</i>	specifies outlier symbol in schematic box-and-whisker plots
LBOXES= <i>linetype</i> (<i>variable</i>)	specifies line types for outlines of box-and-whisker plots
NOTCHES	specifies that box-and-whisker plots are to be notched
PCTLDEF= <i>n</i>	specifies percentile definition used for box-and-whisker plots
SERIFS	adds serifs to the whiskers of skeletal box-and-whisker plots

Table 29.3. Tabulation Options

TABLE	creates a basic table of subgroup values, subgroup sample sizes, subgroup summary statistics, and control limits
TABLEALL	is equivalent to the options TABLE, TABLEBOX, TABLECENTRAL, TABLEID, TABLELEGEND, TABLEOUT, and TABLETEST
TABLEBOX	augments basic table with columns for minimum, 2 ⁵ th percentile, median, 75 th percentile, and maximum of observations in a subgroup
TABLECENTRAL	augments basic table with values of central lines
TABLEID	augments basic table with columns for ID variables
TABLELEGEND	augments basic table with legend for tests for special causes
TABLEOUTLIM	augments basic table with columns indicating control limits exceeded
TABLETESTS	augments basic table with a column indicating which tests for special causes are positive

Note that specifying (EXCEPTIONS) after a tabulation option creates a table for exceptional points only.

Table 29.4. Options for Specifying Tests for Special Causes

TESTS= <i>value-list</i> <i>customized-pattern-list</i>	specifies tests for special causes for the box chart
TEST2RUN= <i>n</i>	specifies length of pattern for Test 2
TEST3RUN= <i>n</i>	specifies length of pattern for Test 3
TESTACROSS	applies tests across <i>phase</i> boundaries
TESTLABEL= <i>'label'</i> <i>(variable)</i> <i>keyword</i>	provides labels for points where test is positive
TESTLABEL <i>n</i> = <i>'label'</i>	specifies label for <i>n</i> th test for special causes
TESTNMETHOD= STANDARDIZE	applies tests to standardized chart statistics
TESTOVERLAP	performs tests on overlapping patterns of points
ZONES	adds lines to box chart delineating zones A, B, and C
ZONEVALPOS= <i>n</i>	specifies position of ZONEVALUES labels
ZONEVALUES	labels zone lines with their values

Table 29.5. Line Printer Options for Displaying Tests for Special Causes

TESTCHAR= <i>'character'</i>	specifies character for line segments that connect any sequence of points for which a test for special causes is positive
ZONECHAR= <i>'character'</i>	specifies character for lines that delineate zones for tests for special causes

Table 29.6. Graphical Options for Displaying Tests for Special Causes

CTESTS= <i>color</i> <i>test-color-list</i>	specifies color for labels indicating points where test is positive
CZONES= <i>color</i>	specifies color for lines and labels delineating zones A, B, and C
LABELFONT= <i>font</i>	specifies software font for labels at points where test is positive (alias for the TESTFONT= option)
LABELHEIGHT= <i>value</i>	specifies height of labels at points where test is positive (alias for the TESTHEIGHT= option)
LTESTS= <i>linetype</i>	specifies type of line connecting points where test is positive
LZONES= <i>linetype</i>	specifies line type for lines delineating zones A, B, and C
TESTFONT= <i>font</i>	specifies software font for labels at points where test is positive
TESTHEIGHT= <i>value</i>	specifies height of labels at points where test is positive

Table 29.7. Clipping Options

CCLIP= <i>color</i>	specifies color for plot symbol for clipped points
CLIPCHAR= <i>'character'</i>	specifies plot character for clipped points
CLIPFACTOR= <i>value</i>	determines extent to which extreme points are clipped
CLIPLEGEND= <i>'string'</i>	specifies text for clipping legend
CLIPLEGPOS= <i>keyword</i>	specifies position of clipping legend
CLIPSUBCHAR= <i>'character'</i>	specifies substitution character for CLIPLEGEND= text
CLIPSYMBOL= <i>symbol</i>	specifies plot symbol for clipped points
LIPSYMBOLHT= <i>value</i>	specifies symbol marker height for clipped points

Table 29.8. Options for Plotting and Labeling Points

ALLLABEL=VALUE (variable)	labels every point on box chart
ALLLABEL2=VALUE (variable)	labels every point on trend chart
CCONNECT= <i>color</i>	specifies color for line segments that connect points on chart
CFRAMELAB= <i>color</i>	specifies fill color for frame around labeled points
CONNECTCHAR= <i>'character'</i>	specifies character used to form line segments that connect points on chart
COUT= <i>color</i>	specifies color for portions of line segments that connect points outside control limits
COUTFILL= <i>color</i>	specifies color for shading areas between the connected points and control limits outside the limits
NOCONNECT	suppresses line segments that connect points on chart
NOTRENDCONNECT	suppresses line segments that connect points on trend chart
OUTLABEL=VALUE (variable)	labels points outside control limits on box chart
SYMBOLCHARS= <i>'characters'</i>	specifies characters indicating <i>symbol-variable</i>
SYMBOLLEGEND= NONE <i>name</i>	specifies LEGEND statement for levels of <i>symbol-variable</i>
SYMBOLORDER= <i>keyword</i>	specifies order in which symbols are assigned for levels of <i>symbol-variable</i>

Table 29.9. Reference Line Options

CHREF= <i>color</i>	specifies color for lines requested by HREF= and HREF2= options
CVREF= <i>color</i>	specifies color for lines requested by VREF= and VREF2= options
HREF= <i>values</i> <i>SAS-data-set</i>	specifies position of reference lines perpendicular to horizontal axis on box chart
HREF2= <i>values</i> <i>SAS-data-set</i>	specifies position of reference lines perpendicular to horizontal axis on trend chart
HREFCHAR= <i>'character'</i>	specifies line character for HREF= and HREF2= lines
HREFLABELS= <i>('label1'...'labeln')</i>	specifies labels for HREF= lines
HREF2LABELS= <i>('label1'...'labeln')</i>	specifies labels for HREF2= lines
HREFLABPOS= <i>n</i>	specifies position of HREFLABELS= and HREF2LABELS= labels
LHREF= <i>linetype</i>	specifies line type for HREF= and HREF2= lines
LVREF= <i>linetype</i>	specifies line type for VREF= and VREF2= lines
NOBYREF	specifies that reference line information in a data set is to be applied uniformly to charts created for all BY groups
VREF= <i>values</i> <i>SAS-data-set</i>	specifies position of reference lines perpendicular to vertical axis on box chart
VREF2= <i>values</i> <i>SAS-data-set</i>	specifies position of reference lines perpendicular to vertical axis on trend chart
VREFCHAR= <i>'character'</i>	specifies line character for VREF= and VREF2= lines
VREFLABELS= <i>'label1'...'labeln'</i>	specifies labels for VREF= lines
VREF2LABELS= <i>'label1'...'labeln'</i>	specifies labels for VREF2= lines
VREFLABPOS= <i>n</i>	specifies position of VREFLABELS= and VREF2LABELS= labels

Table 29.10. Block Variable Legend Options

BLOCKLABELPOS= <i>keyword</i>	specifies position of label for <i>block-variable</i> legend
BLOCKLABTYPE= <i>n</i> <i>keyword</i>	specifies text size of <i>block-variable</i> legend
BLOCKPOS= <i>n</i>	specifies vertical position of <i>block-variable</i> legend
BLOCKREP	repeats identical consecutive labels in <i>block-variable</i> legend
CBLOCKLAB= <i>color</i>	specifies color for filling background in <i>block-variable</i> legend
CBLOCKVAR= <i>variable</i> <i>(variables)</i>	specifies one or more variables whose values are colors for filling background of <i>block-variable</i> legend

Table 29.11. Options for Specifying Control Limits

ALPHA= <i>value</i>	requests probability limits for control charts
CONTROLSTAT= <i>keyword</i>	specifies whether control limits are computed for subgroup means or subgroup medians
LIMITN= <i>n</i> VARYING	specifies either nominal sample size for fixed control limits or varying limits
NOREADLIMITS	computes control limits for each <i>process</i> from the data rather than from a LIMITS= data set (Release 6.10 and later releases)
READALPHA	reads _ALPHA_ instead of _SIGMAS_ from a LIMITS= data set
READINDEXES=ALL <i>'label1' ... 'labeln'</i>	reads multiple sets of control limits for each <i>process</i> from a LIMITS= data set
READLIMITS	reads single set of control limits for each <i>process</i> from a LIMITS= data set (Release 6.09 and earlier releases)
SIGMAS= <i>k</i>	specifies width of control limits in terms of multiple <i>k</i> of standard error of plotted statistic

Table 29.12. Options for Displaying Control Limits

CINFILL= <i>color</i>	specifies color for area inside control limits
CLIMITS= <i>color</i>	specifies color of control limits, central line, and related labels
LCLLABEL= <i>'label'</i>	specifies label for lower control limit on box chart
LIMLABSUBCHAR= <i>'character'</i>	specifies a substitution character for labels provided as quoted strings; the character is replaced with the value of the control limit
LLIMITS= <i>linetype</i>	specifies line type for control limits
NDECIMAL= <i>n</i>	specifies number of digits to right of decimal place in default labels for control limits and central line in box chart
NOCTL	suppresses display of central line in box chart
NOLCL	suppresses display of lower control limit in box chart
NOLIMITLABEL	suppresses labels for control limits and central line
NOLIMITS	suppresses display of control limits
NOLIMITSFRAME	suppresses default frame around control limit information when multiple sets of control limits are read from LIMITS= data set
NOLIMITSLEGEND	suppresses legend for control limits
NOUCL	suppresses display of upper control limit in box chart
UCLLABEL= <i>'string'</i>	specifies label for upper control limit in box chart
WLIMITS= <i>n</i>	specifies width for control limits and central line
XSYMBOL= <i>'string'</i> <i>keyword</i>	specifies label for central line in box chart

Table 29.13. Axis and Axis Label Options

CAXIS= <i>color</i>	specifies color for axis lines and tick marks
CFRAME= <i>color</i> (<i>color-list</i>)	specifies fill colors for frame for plot area
CTEXT= <i>color</i>	specifies color for tick mark values and axis labels
HAXIS= <i>values</i> <i>AXISn</i>	specifies major tick mark values for horizontal axis
HEIGHT= <i>value</i>	specifies height of axis label and axis legend text
HMINOR= <i>n</i>	specifies number of minor tick marks between major tick marks on horizontal axis
HOFFSET= <i>value</i>	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 subgroup values is to be labeled on horizontal axis
NOVANGLE	requests vertical axis labels that are strung out vertically
SKIPHLABELS= <i>n</i>	specifies thinning factor for tick mark labels on horizontal axis
SPLIT='character'	specifies splitting character for axis labels
TURNHLABELS	requests horizontal axis labels that are strung out vertically
VAXIS= <i>values</i> <i>AXISn</i>	specifies major tick mark values for vertical axis of box chart
VAXIS2= <i>values</i> <i>AXISn</i>	specifies major tick mark values for vertical axis of trend chart
VMINOR= <i>n</i>	specifies number of minor tick marks between major tick marks on vertical axis
VOFFSET= <i>value</i>	specifies length of offset at both ends of vertical axis
VZERO	forces origin to be included in vertical axis for primary chart
VZERO2	forces origin to be included in vertical axis for secondary chart
WAXIS= <i>n</i>	specifies width of axis lines

Table 29.14. Phase Options

CPHASEBOX= <i>color</i>	specifies color for box enclosing all plotted points for a phase
CPHASEBOXCONNECT= <i>color</i>	specifies color for line segments connecting adjacent enclosing boxes
CPHASEBOXFILL= <i>color</i>	specifies fill color for box enclosing all plotted points for a phase
CPHASELEG= <i>color</i>	specifies text color for <i>phase</i> legend
CPHASEMEANCONNECT= <i>color</i>	specifies color for line segments connecting average value points within a phase
NOPHASEFRAME	suppresses default frame for <i>phase</i> legend
OUTPHASE='string'	specifies value of <code>_PHASE_</code> in the <code>OUTHISTORY=</code> data set
PHASEBREAK	disconnects last point in a <i>phase</i> from first point in next <i>phase</i>
PHASELABTYPE= <i>value</i> <i>keyword</i>	specifies text size of <i>phase</i> legend
PHASELEGEND	displays <i>phase</i> labels in a legend across top of chart
PHASEMEANSYMBOL= <i>symbol</i>	specifies symbol marker for average of values within a phase
PHASEREF	delineates <i>phases</i> with vertical reference lines
READPHASES= ALL 'label1' ... 'labeln'	specifies <i>phases</i> to be read from an input data set

Table 29.15. Specification Limit Options

CIALPHA= <i>value</i>	specifies α value for computing capability index confidence limits
CITYPE= <i>keyword</i>	specifies capability index confidence limits type
LSL= <i>value-list</i>	specifies list of lower specification limits
TARGET= <i>value-list</i>	specifies list of target values
USL= <i>value-list</i>	specifies list of upper specification limits

Table 29.16. Process Mean and Standard Deviation Options

MEDCENTRAL= <i>keyword</i>	specifies method for estimating process mean μ
MU0= <i>value</i>	specifies known value of μ_0 for process mean μ
RANGES	specifies that estimate of process standard deviation σ is to be calculated from subgroup ranges
SIGMA0= <i>value</i>	specifies known value σ_0 for process standard deviation σ
SMETHOD= <i>keyword</i>	specifies method for estimating process standard deviation σ
TYPE= <i>keyword</i>	identifies whether parameters are estimates or standard values and specifies value of <code>_TYPE_</code> in the <code>OUTLIMITS=</code> data set

Table 29.17. Input Data Set Options

MISSBREAK	specifies that observations with missing values are not to be processed
-----------	---

Table 29.18. Output Data Set Options

IMAGEMAP= <i>SAS-data-set</i>	creates OUTTABLE= data set with additional graph coordinate data
OUTHISTORY= <i>SAS-data-set</i>	creates output data set containing subgroup summary statistics
OUTINDEX= <i>'string'</i>	specifies value of <code>_INDEX_</code> in the OUTLIMITS= data set
OUTLIMITS= <i>SAS-data-set</i>	creates output data set containing control limits
OUTTABLE= <i>SAS-data-set</i>	creates output data set containing subgroup summary statistics and control limits

Table 29.19. Graphical Enhancement Options

ANNOTATE= <i>SAS-data-set</i>	specifies annotate data set that adds features to box chart
ANNOTATE2= <i>SAS-data-set</i>	specifies annotate data set that adds features to trend chart
DESCRIPTION= <i>'string'</i>	specifies string that appears in the description field of the PROC GREPLAY master menu for box chart
FONT= <i>font</i>	specifies software font for labels and legends on charts
HTML=(<i>variable</i>)	specifies a variable whose values are URLs to be associated with subgroups
NAME= <i>'string'</i>	specifies name that appears in the name field of the PROC GREPLAY master menu for box chart
PAGENUM= <i>'string'</i>	specifies the form of the label used in pagination
PAGENUMPOS= <i>keyword</i>	specifies the position of the page number requested with the PAGENUM= option
WTREND= <i>n</i>	specifies width of line segments connecting points on trend chart

Table 29.20. Grid Options

ENDGRID	adds grid after last plotted point
GRID	adds grid to control chart
LENDGRID= <i>linetype</i>	specifies line type for grid requested with the ENDGRID option
LGRID= <i>linetype</i>	specifies line type for grid requested with the GRID option
WGRID= <i>n</i>	specifies width of grid lines

Table 29.21. Plot Layout Options

ALLN	plots summary statistics for all subgroups
BILEVEL	creates control charts using half-screens and half-pages
EXCHART	creates control charts for a process variable only when exceptions occur
INTERVAL= <i>keyword</i>	specifies natural time interval between consecutive subgroup positions when time, date, or datetime format is associated with a numeric subgroup variable
MAXPANELS= <i>n</i>	specifies maximum number of pages or screens for chart
NMARKERS	requests special markers for points corresponding to sample sizes not equal to nominal sample size for fixed control limits
NOCHART	suppresses creation of box chart
NOFRAME	suppresses frame for plot area
NOLEGEND	suppresses legend for subgroup sample sizes
NPANELPOS= <i>n</i>	specifies number of subgroup positions per panel on each chart
REPEAT	repeats last subgroup position on panel as first subgroup position of next panel
TOTPANELS= <i>n</i>	specifies number of pages or screens to be used to display chart
TRENDVAR= <i>variable</i> (<i>variable-list</i>)	specifies list of trend variables
YPCT1= <i>value</i>	specifies length of vertical axis on box chart as a percentage of sum of lengths of vertical axes for box and trend charts
ZEROSTD	displays box chart regardless of whether $\hat{\sigma} = 0$

Details

Constructing Box Charts

The following notation is used in this section:

μ	process mean (expected value of the population of measurements)
σ	process standard deviation (standard deviation of the population of measurements)
\bar{X}_i	mean of measurements in i^{th} subgroup
n_i	sample size of i^{th} subgroup
N	the number of subgroups
x_{ij}	j^{th} measurement in the i^{th} subgroup, $j = 1, 2, 3, \dots, n_i$
$x_{i(j)}$	j^{th} largest measurement in the i^{th} subgroup: $x_{i(1)} \leq x_{i(2)} \leq \dots \leq x_{i(n_i)}$
$\bar{\bar{X}}$	weighted average of subgroup means
M_i	median of the measurements in the i^{th} subgroup: $M_i = \begin{cases} x_{i((n_i+1)/2)} & \text{if } n_i \text{ is odd} \\ (x_{i(n_i/2)} + x_{i((n_i/2)+1)})/2 & \text{if } n_i \text{ is even} \end{cases}$
\bar{M}	average of the subgroup medians: $\bar{M} = (n_1 M_1 + \dots + n_N M_N) / (n_1 + \dots + n_N)$
\tilde{M}	median of the subgroup medians. Denote the j^{th} largest median by $M_{(j)}$ so that $M_{(1)} \leq M_{(2)} \leq \dots \leq M_{(N)}$. $\tilde{M} = \begin{cases} M_{((N+1)/2)} & \text{if } N \text{ is odd} \\ (M_{(N/2)} + M_{(N/2)+1})/2 & \text{if } N \text{ is even} \end{cases}$
$e_M(n)$	standard error of the median of n independent, normally distributed variables with unit standard deviation (the value of $e_M(n)$ can be calculated with the STDMED function in a DATA step)
$Q_p(n)$	100 p^{th} percentile ($0 < p < 1$) of the distribution of the median of n independent observations from a normal population with unit standard deviation
z_p	100 p^{th} percentile of the standard normal distribution
$D_p(n)$	100 p^{th} percentile of the distribution of the range of n independent observations from a normal population with unit standard deviation

Elements of Box-and-Whisker Plots

A box-and-whisker plot is displayed for the measurements in each subgroup on the box chart. Figure 29.12 illustrates the elements of each plot.

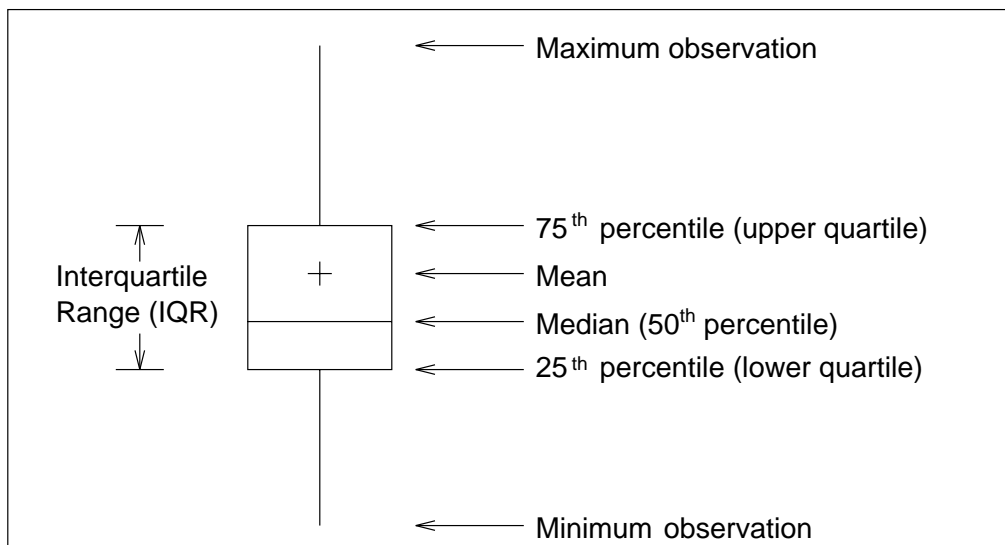


Figure 29.12. Box-and-Whisker Plot

The skeletal style of the box-and-whisker plot shown in Figure 29.12 is the default. You can specify alternative styles with the `BOXSTYLE=` option; see Example 29.2 on page 1149 or the entry for the `BOXSTYLE=` option on page 1739.

Control Limits and Central Line

You can compute the limits in the following ways:

- as a specified multiple (k) of the standard error of \bar{X}_i (or M_i) above and below the central line. The default limits are computed with $k = 3$ (these are referred to as 3σ limits).
- as probability limits defined in terms of α , a specified probability that \bar{X}_i (or M_i) exceeds the limits

The `CONTROLSTAT=` option specifies whether control limits are computed for subgroup means (the default) or subgroup medians. The following tables provide the formulas for the limits:

Table 29.22. Control Limits and Central Line for Box Charts

CONTROLSTAT=MEAN	CONTROLSTAT=MEDIAN
LCLX = lower limit = $\bar{\bar{X}} - k\hat{\sigma}/\sqrt{n_i}$	LCLM = lower limit = $\bar{M} - k\hat{\sigma}e_M(n_i)$
Central Line = $\bar{\bar{X}}$	Central Line = \bar{M}
UCLX = upper limit = $\bar{\bar{X}} + k\hat{\sigma}/\sqrt{n_i}$	UCLM = upper limit = $\bar{M} + k\hat{\sigma}e_M(n_i)$

Table 29.23. Probability Limits and Central Line for Box Charts

CONTROLSTAT=MEAN	CONTROLSTAT=MEDIAN
LCLX = lower limit = $\bar{\bar{X}} - z_{\alpha/2}(\hat{\sigma}/\sqrt{n_i})$	LCLM = lower limit = $\bar{M} - Q_{\alpha/2}(n_i)\hat{\sigma}$
Central Line = $\bar{\bar{X}}$	Central Line = \bar{M}
UCLX = upper limit = $\bar{\bar{X}} + z_{\alpha/2}(\hat{\sigma}/\sqrt{n_i})$	UCLM = upper limit = $\bar{M} + Q_{1-\alpha/2}(n_i)\hat{\sigma}$

In the preceding tables, replace \bar{M} with $\overline{\bar{X}}$ if you specify MEDCENTRAL=AVGMEAN in addition to CONTROLSTAT=MEDIAN. Likewise, replace \bar{M} with \tilde{M} if you specify MEDCENTRAL=MEDMED in addition to CONTROLSTAT=MEDIAN. If standard values μ_0 and σ_0 are available for μ and σ , replace $\overline{\bar{X}}$ with μ_0 and $\hat{\sigma}$ with σ_0 in Table 29.22 and Table 29.23.

Note that the limits vary with n_i . The formulas for median limits assume that the data are normally distributed.

You can specify parameters for the limits as follows:

- Specify k with the SIGMAS= option or with the variable `_SIGMAS_` in a LIMITS= data set.
- Specify α with the ALPHA= option or with the variable `_ALPHA_` in a LIMITS= data set.
- Specify a constant nominal sample size $n_i \equiv n$ for the control limits with the LIMITN= option or with the variable `_LIMITN_` in a LIMITS= data set.
- Specify μ_0 with the MU0= option or with the variable `_MEAN_` in a LIMITS= data set.
- Specify σ_0 with the SIGMA0= option or with the variable `_STDDEV_` in a LIMITS= data set.

Note: You can suppress the display of the control limits with the NOLIMITS option. This is useful for creating standard side-by-side box-and-whisker plots (in this case, the STDDEVIATIONS option is also recommended).

Output Data Sets

OUTLIMITS= Data Set

The OUTLIMITS= data set saves control limits and control limit parameters. The following variables can be saved:

Table 29.24. OUTLIMITS= Data Set

Variable	Description
ALPHA	probability (α) of exceeding limits
CP	capability index C_p
CPK	capability index C_{pk}
CPL	capability index C_{PL}
CPM	capability index C_{pm}
CPU	capability index C_{PU}
INDEX	optional identifier for the control limits specified with the OUTINDEX= option
LCLM	lower control limit for subgroup median
LCLR	lower control limit for subgroup range
LCLS	lower control limit for subgroup standard deviation
LCLX	lower control limit for subgroup mean
LIMITN	nominal sample size associated with the control limits
LSL	lower specification limit
MEAN	process mean (value of central line on box chart)
R	value of central line on R chart
S	value of central line on s chart
SIGMAS	multiple (k) of standard error of \bar{X}_i or M_i
STDDEV	process standard deviation ($\hat{\sigma}$ or σ_0)
SUBGRP	<i>subgroup-variable</i> specified in the BOXCHART statement
TARGET	target value
TYPE	type (estimate or standard value) of _MEAN_ and _STDDEV_
UCLM	upper control limit for subgroup median
UCLR	upper control limit for subgroup range
UCLS	upper control limit for subgroup standard deviation
UCLX	upper control limit for subgroup mean
USL	upper specification limit
VAR	<i>process</i> specified in the BOXCHART statement

Notes:

1. The variables _LCLM_ and _UCLM_ are included if you specify CONTROLSTAT=MEDIAN; otherwise, the variables _LCLX_ and _UCLX_ are included.
2. The variables _LCLS_, _S_, and _UCLS_ are included if you specify the STDDEVIATIONS option; otherwise, the variables _LCLR_, _R_, and _UCLR_ are included. These variables are not used to create box charts, but

they allow the OUTLIMITS= data set to be used as a LIMITS= data set with the XRCHART, XSCHART, MRCHART, SCHART, and RCHART statements.

3. If the control limits vary with subgroup sample size, the special missing value V is assigned to the variables `_LIMITN_`, `_LCLX_`, `_UCLX_`, `_LCLM_`, `_UCLM_`, `_LCLR_`, `_R_`, `_UCLR_`, `_LCLS_`, `_S_`, and `_UCLS_`.
4. If the limits are defined in terms of a multiple k of the standard error of \bar{X}_i , the value of `_ALPHA_` is computed as $\alpha = 2(1 - \Phi(k))$, where $\Phi(\cdot)$ is the standard normal distribution function. If the limits are defined in terms of a multiple k of the standard error of M_i , the value of `_ALPHA_` is computed as $\alpha = 2(1 - F_{med}(k, n))$, where $F_{med}(\cdot, n)$ is the cumulative distribution function of the median of a random sample of n standard normally distributed observations, and n is the value of `_LIMITN_`. If `_LIMITN_` has the special missing value V , this value is assigned to `_ALPHA_`.
5. If the limits for means are probability limits, the value of `_SIGMAS_` is computed as $k = \Phi^{-1}(1 - \alpha/2)$, where Φ^{-1} is the inverse standard normal distribution function. If the limits for medians are probability limits, the value of `_SIGMAS_` is computed as $k = F_{med}^{-1}(1 - \alpha/2, n)$, where $F_{med}^{-1}(\cdot, n)$ is the inverse distribution function of the median of a random sample of n standard normally distributed observations, and n is the value `_LIMITN_`. If `_LIMITN_` has the special missing value V , this value is assigned to `_SIGMAS_`.
6. The variables `_CP_`, `_CPK_`, `_CPL_`, `_CPU_`, `_LSL_`, and `_USL_` are included only if you provide specification limits with the LSL= and USL= options. The variables `_CPM_` and `_TARGET_` are included if, in addition, you provide a target value with the TARGET= option. See “Capability Indices” on page 1648 for computational details.
7. Optional BY variables are saved in the OUTLIMITS= data set.

The OUTLIMITS= data set contains one observation for each *process* specified in the BOXCHART statement. For an example, see “Saving Control Limits” on page 1118.

OUTHISTORY= Data Set

The OUTHISTORY= data set saves subgroup summary statistics. The following variables can be saved:

- the *subgroup-variable*
- a subgroup minimum variable named by the prefix *process* suffixed with L
- a subgroup first-quartile variable named by the prefix *process* suffixed with I
- a subgroup mean variable named by the prefix *process* suffixed with X
- a subgroup median variable named by the prefix *process* suffixed with M
- a subgroup third-quartile variable named by the prefix *process* suffixed with 3
- a subgroup maximum variable named by the prefix *process* suffixed with H
- a subgroup sample size variable named by the prefix *process* suffixed with N
- a subgroup range variable named by the prefix *process* suffixed with R or a subgroup standard deviation variable named by *process* suffixed with S

Part 9. The CAPABILITY Procedure

A subgroup standard deviation variable is included if you specify the STDDEVIATIONS option; otherwise, a subgroup range variable is included.

Given a *process* name that contains eight characters, the procedure first shortens the name to its first four characters and its last three characters, and then it adds the suffix. For example, the procedure shortens the *process* DIAMETER to DIAMTER before adding the suffix.

Subgroup summary variables are created for each *process* specified in the BOXCHART statement. For example, consider the following statements:

```
proc shewhart data=steel;
  boxchart (width diameter)*lot / outhistory=summary;
run;
```

The data set SUMMARY contains variables named LOT, WIDTHL, WIDTH1, WIDTHM, WIDTHX, WIDTH3, WIDTHH, WIDTHR, WIDTHN, DIAMTERL, DIAMTER1, DIAMTERM, DIAMTERX, DIAMTER3, DIAMTERH, DIAMTERR, and DIAMTERN.

The variables WIDTHR and DIAMTERR are included since the STDDEVIATIONS option is not specified. If you specified the STDDEVIATIONS option, the data set SUMMARY would contain the variables WIDTHS and DIAMTERS rather than WIDTHR and DIAMTERR.

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

- BY variables
- *block-variables*
- *symbol-variable*
- ID variables
- `_PHASE_` (if the OUTPHASE= option is specified)

For an example of an OUTHISTORY= data set, see “Saving Summary Statistics” on page 1116.

OUTTABLE= Data Set

The OUTTABLE= data set saves subgroup summary statistics, control limits, and related information. The following variables can be saved:

Variable	Description
<code>_ALPHA_</code>	probability (α) of exceeding control limits
<code>_EXLIM_</code>	control limit exceeded on box chart
<code>_LCLM_</code>	lower control limit for median
<code>_LCLX_</code>	lower control limit for mean
<code>_LIMITN_</code>	nominal sample size associated with the control limits
<code>_MEAN_</code>	process mean
<code>_SIGMAS_</code>	multiple (k) of the standard error associated with control limits
<i>subgroup</i>	values of the subgroup variable
<code>_SUBMAX_</code>	subgroup maximum
<code>_SUBMED_</code>	subgroup median
<code>_SUBMIN_</code>	subgroup minimum
<code>_SUBN_</code>	subgroup sample size
<code>_SUBQ1_</code>	subgroup first quartile (25 th percentile)
<code>_SUBQ3_</code>	subgroup third quartile (75 th percentile)
<code>_SUBX_</code>	subgroup mean
<code>_TESTS_</code>	tests for special causes signaled on box chart
<code>_UCLM_</code>	upper control limit for median
<code>_UCLX_</code>	upper control limit for mean
<code>_VAR_</code>	<i>process</i> specified in the BOXCHART statement

The variables `_LCLM_` and `_UCLM_` are included if you specify `CONTROLSTAT=MEDIAN`; otherwise, the variables `_LCLX_` and `_UCLX_` are included. In addition, the following variables, if specified, are included:

- BY variables
- *block-variables*
- *symbol-variable*
- ID variables
- `_PHASE_` (if the `READPHASES=` option is specified)
- `_TREND_` (if the `TRENDVAR=` option is specified)

Notes:

1. Either the variable `_ALPHA_` or the variable `_SIGMAS_` is saved depending on how the control limits are defined (with the `ALPHA=` or `SIGMAS=` options, respectively, or with the corresponding variables in a `LIMITS=` data set).
2. The variable `_TESTS_` is saved if you specify the `TESTS=` option. The k^{th} character of a value of `_TESTS_` is k if Test k is positive at that subgroup. For example, if you request all eight tests and Tests 2 and 8 are positive for a given subgroup, the value of `_TESTS_` has a 2 for the second character, an 8 for the eighth character, and blanks for the other six characters.
3. The variables `_VAR_`, `_EXLIM_`, and `_TESTS_` are character variables of length 8. The variable `_PHASE_` is a character variable of length 16. All other variables are numeric.

For an example, see “Saving Control Limits” on page 1118.

ODS Tables

The following table summarizes the ODS tables that you can request with the BOXCHART statement.

Table 29.25. ODS Tables Produced with the BOXCHART Statement

Table Name	Description	Options
BOXCHART	box plot summary statistics	TABLE, TABLEALL, TABLEBOX, TABLEC, TABLEID, TABLELEG, TABLEOUT, TABLETESTS
Tests	descriptions of tests for special causes requested with the TESTS= option for which at least one positive signal is found	TABLEALL, TABLELEG

Input Data Sets

DATA= Data Set

You can read raw data (process measurements) from a DATA= data set specified in the PROC SHEWHART statement. Each *process* specified in the BOXCHART statement must be a SAS variable in the data set. This variable provides measurements which must be grouped into subgroup samples indexed by the *subgroup-variable*. The *subgroup-variable*, specified in the BOXCHART 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 *process* and a value for the *subgroup-variable*. If the t^{th} subgroup contains n_i measurements, there should be n_i consecutive observations for which the value of the *subgroup-variable* is the index of the t^{th} subgroup. For example, if each subgroup contains 20 items and there are 30 subgroup samples, the DATA= data set should contain 600 observations. Other variables that can be read from a DATA= data set include

- `_PHASE_` (if READPHASES= is specified)
- *block-variables*
- *symbol-variable*
- BY variables
- ID variables

By default, the SHEWHART procedure reads all of the observations in a DATA= data set. However, if the data set includes the variable `_PHASE_`, you can read selected groups of observations (referred to as *phases*) with the READPHASES= option (for an example, see “Displaying Stratification in Phases” on page 1814).

For an example of a DATA= data set, see “Creating Box Charts from Raw Data” on page 1110.

LIMITS= Data Set

You can read preestablished control limits (or parameters from which the control limits can be calculated) from a LIMITS= data set specified in the PROC SHEWHART statement. For example, the following statements read control limit information from the data set CONLIMS:*

```
proc shewhart data=info limits=conlims;
  boxchart weight*batch;
run;
```

The LIMITS= data set can be an OUTLIMITS= data set that was created in a previous run of the SHEWHART procedure. Such data sets always contain the variables required for a LIMITS= data set; see Table 29.24 on page 1136. The LIMITS= data set can also be created directly using a DATA step. When you create a LIMITS= data set, you must provide one of the following:

- the variables `_LCLX_`, `_MEAN_`, and `_UCLX_` or (if you specify `CONTROLSTAT=MEDIAN`) the variables `_LCLM_`, `_MEAN_`, and `_UCLM_`. These variables specify the control limits directly.
- the variables `_MEAN_` and `_STDDEV_`, which are used to calculate the control limits according to the equations in Table 29.22 on page 1134 and Table 29.23 on page 1134

In addition, note the following:

- The variables `_VAR_` and `_SUBGRP_` are required. These must be character variables of length 8.
- The variable `_INDEX_` is required if you specify the `READINDEX=` option; this must be a character variable of length 16.
- The variables `_LIMITN_`, `_SIGMAS_` (or `_ALPHA_`), and `_TYPE_` are optional, but they are recommended to maintain a complete set of control limit information. The variable `_TYPE_` must be a character variable of length 8; valid values are **ESTIMATE**, **STANDARD**, **STDMU**, and **STDSIGMA**.
- BY variables are required if specified with a BY statement.

For an example, see “Reading Preestablished Control Limits” on page 1121.

HISTORY= Data Set

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

A HISTORY= data set used with the BOXCHART statement must contain the following:

*In Release 6.09 and in earlier releases, it is necessary to specify the READLIMITS option.

Part 9. The CAPABILITY Procedure

- the *subgroup-variable*
- a subgroup minimum variable for each *process*
- a subgroup first-quartile variable for each *process*
- a subgroup median variable for each *process*
- a subgroup mean variable for each *process*
- a subgroup third-quartile variable for each *process*
- a subgroup maximum variable for each *process*
- a subgroup sample size variable for each *process*
- either a subgroup range variable or a subgroup standard deviation variable for each *process*

If you specify the STDDEVIATIONS option, the subgroup standard deviation variable must be included; otherwise, the subgroup range variable must be included.

The names of the subgroup summary statistics variables must be the *process* name concatenated with the following special suffix characters:

Subgroup Summary Statistic	Suffix Character
subgroup minimum	L
subgroup first-quartile	1
subgroup median	M
subgroup mean	X
subgroup third-quartile	3
subgroup maximum	H
subgroup sample size	N
subgroup range	R
subgroup standard deviation	S

For example, consider the following statements:

```
proc shewhart history=summary;  
    boxchart (weight yldstren)*batch;  
run;
```

The data set SUMMARY must include the variables BATCH, WEIGHTL, WEIGHT1, WEIGHTM, WEIGHTX, WEIGHT3, WEIGHTH, WEIGHTR, WEIGHTN, YLDSREN1, YLDSREN1, YLDSRENM, YLDSRENX, YLDSREN3, YLDSRENH, YLDSRENR, and YLDSRENN.

If the STDDEVIATIONS option were specified in the preceding BOXCHART statement, it would be necessary for SUMMARY to include the variables WEIGHTS and YLDSRENS rather than WEIGHTR and YLDSRENR.

Note that if you specify a *process* name that contains eight characters, the names of the summary variables must be formed from the first four characters and the last three characters of the *process* name, suffixed with the appropriate character.

Other variables that can be read from a HISTORY= data set include

- `_PHASE_` (if READPHASES= is specified)
- *block-variables*

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

By default, the SHEWHART procedure reads all of the observations in a HISTORY= data set. However, if the data set includes the variable _PHASE_, you can read selected groups of observations (referred to as *phases*) with the READPHASES= option (see “Displaying Stratification in Phases” on page 1814 for an example).

For an example of a HISTORY= data set, see “Creating Box Charts from Subgroup Summary Data” page 1113.

TABLE= Data Set

You can read summary statistics and control limits from a TABLE= data set specified in the PROC SHEWHART statement. This enables you to reuse an OUTTABLE= data set created in a previous run of the SHEWHART procedure. Because the SHEWHART procedure simply displays the information in a TABLE= data set, you can use TABLE= data sets to create specialized control charts. Examples are provided in Chapter 46, “Specialized Control Charts,”.

The following table lists the variables required in a TABLE= data set used with the BOXCHART statement:

Table 29.26. Variables Required in a TABLE= Data Set

Variable	Description
LCLM	lower control limit for median
LCLX	lower control limit for mean
LIMITN	nominal sample size associated with the control limits
MEAN	process mean
<i>subgroup-variable</i>	values of the <i>subgroup-variable</i>
SUBMAX	subgroup maximum
SUBMIN	subgroup minimum
SUBMED	subgroup median
SUBN	subgroup sample size
SUBQ1	subgroup first quartile (25 th percentile)
SUBQ3	subgroup third quartile (75 th percentile)
SUBX	subgroup mean
UCLM	upper control limit for median
UCLX	upper control limit for mean

Note that if you specify CONTROLSTAT=MEDIAN, the variables _LCLM_, _SUBMED_, and _UCLM_ are required; otherwise, the variables _LCLX_, _SUBX_, and _UCLX_ are required.

Other variables that can be read from a TABLE= data set include

- *block-variables*
- *symbol-variable*

- BY variables
- ID variables
- `_PHASE_` (if the `READPHASES=` option is specified). This variable must be a character variable of length 16.
- `_TESTS_` (if the `TESTS=` option is specified). This variable is used to flag tests for special causes and must be a character variable of length 8.
- `_VAR_`. This variable is required if more than one *process* is specified or if the data set contains information for more than one *process*. This variable must be a character variable of length 8.

For an example of a `TABLE=` data set, see “Saving Control Limits” on page 1118.

Methods for Estimating the Standard Deviation

When control limits are computed from the input data, three methods (referred to as default, `MVLUE` and `RMSDF`) are available for estimating the process standard deviation σ . The method depends on whether you specify the `STDDEVIATIONS` option. If you specify this option, σ is estimated using subgroup standard deviations, and otherwise, σ is estimated using subgroup ranges. For further details and formulas, see “Methods for Estimating the Standard Deviation” on page 1593.

Percentile Definitions

You can use the `PCTLDEF=` option to specify one of five definitions for computing quantile statistics (percentiles). Let n equal the number of nonmissing values for a variable, and let x_1, x_2, \dots, x_n represent the ordered values of the process variable. For the t^{th} percentile, set $p = t/100$, and express np as

$$np = j + g$$

where j is the integer part of np , and g is the fractional part of np .

The t^{th} percentile (call it y) can be defined in five ways, as described in the next five sections.

`PCTLDEF=1`

This uses the weighted average at x_{np}

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

where x_0 is taken to be x_1 .

`PCTLDEF=2`

This uses the observation numbered closest to np

$$y = x_i$$

where i is the integer part of $np + 1/2$.

PCTLDEF=3

This uses the empirical distribution function

$$\begin{aligned} y &= x_j && \text{if } g = 0 \\ y &= x_{j+1} && \text{if } g > 0 \end{aligned}$$

PCTLDEF=4

This uses the weighted average aimed at $x_{p(n+1)}$

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

where $(n + 1)p = j + g$, and where x_{n+1} is taken to be x_n .

PCTLDEF=5

This uses the empirical distribution function with averaging

$$\begin{aligned} y &= (x_j + x_{j+1})/2 && \text{if } g = 0 \\ y &= x_{j+1} && \text{if } g > 0 \end{aligned}$$

Axis Labels

You can specify axis labels by assigning labels to particular variables in the input data set, as summarized in the following table:

Axis	Input Data Set	Variable
Horizontal	all	<i>subgroup-variable</i>
Vertical (box chart)	DATA=	<i>process</i>
Vertical (box chart)	HISTORY=	subgroup mean variable
Vertical (box chart)	TABLE=	_SUBX_

Note that if you specify the CONTROLSTAT=MEDIAN option, you should assign the label to the subgroup median variable in a HISTORY= data set or to the variable _SUBMED_ in an TABLE= data set.

If you specify the TRENDVAR= option, you can provide distinct labels for the vertical axes of the box and trend charts by breaking the vertical axis into two parts with a split character. Specify the split character with the SPLIT= option. The first part labels the vertical axis of the box chart, and the second part labels the vertical axis of the trend chart.

For an example, see “Labeling Axes” on page 1849.

Missing Values

An observation read from a DATA=, HISTORY=, or TABLE= data set is not analyzed if the value of the subgroup variable is missing. For a particular process variable, an observation read from a DATA= data set is not analyzed if the value of the process variable is missing. Missing values of process variables generally lead to unequal subgroup sample sizes. For a particular process variable, an observation read from

Part 9. The CAPABILITY Procedure

a HISTORY= or TABLE= data set is not analyzed if the values of any of the corresponding summary variables are missing.

Examples

This section provides advanced examples of the BOXCHART statement.

Example 29.1. Using Box Charts to Compare Subgroups

In this example, a box chart 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 cancelled, the delay is recorded as a missing value.

See SHWBOX4
in the SAS/QC
Sample Library

```

data times;
  informat day date7. ;
  format   day date7. ;
  input day @ ;
  do flight=1 to 25;
    input delay @ ;
    output;
  end;

cards;
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
;

```

First, the MEANS procedure is used to count the number of cancelled flights for each day. This information is then added to the data set TIMES.

```

proc means data=times noprint;
  var delay;

```

Part 9. The CAPABILITY Procedure

```
      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 & ;
cards;
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;
```

Next, control limits are established using the delays prior to the holiday period.

```
proc shewhart data=times;
  where day <= '21DEC88'D;
  boxchart delay * day /
    nochart
    stddeviations
    outlimits=timelim;
run;
```

The OUTLIMITS= option names a data set (TIMELIM) that saves the control limits. The STDDEVIATIONS option specifies that the estimate of σ is to be calculated from subgroup standard deviations. This, in turn, affects the calculation of the control limits. The NOCHART option suppresses the display of the chart.

The following statements create a box chart for the complete set of data using the control limits in TIMELIM:

```
symbol1 c=black v=plus;
symbol2 c=black v=square;
symbol3 c=black v=triangle;
```

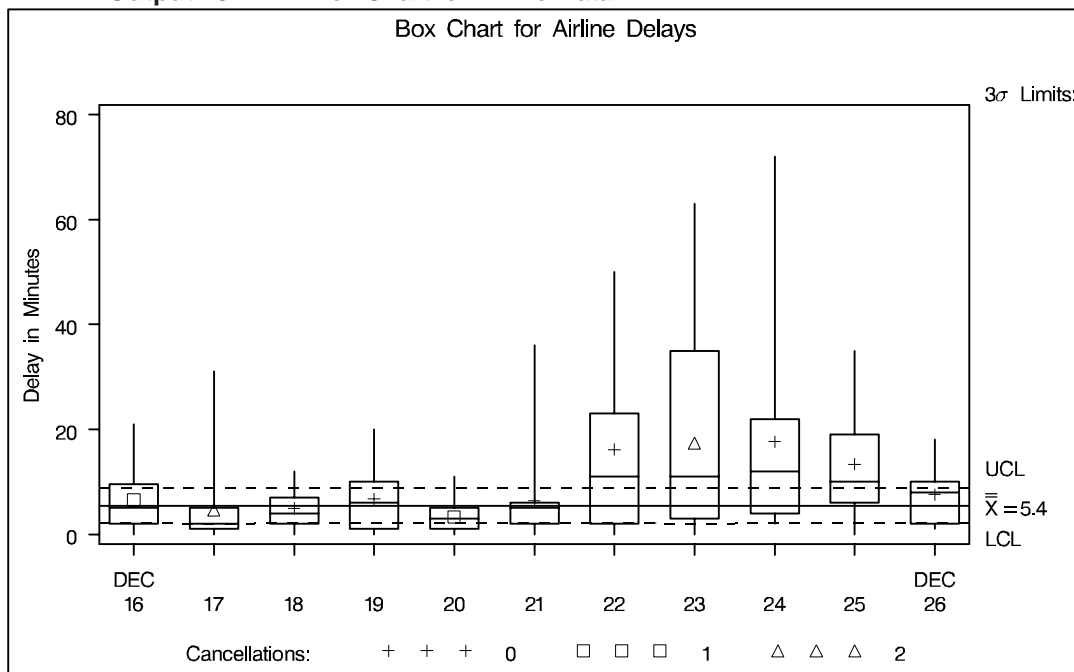
```

title 'Box Chart for Airline Delays';
proc shewhart data=times limits=timelim ;
  boxchart delay * day = ncancel /
    stddeviations
    nohlabel
    nolegend
    symbollegend=legend1;
  legend1 label=('Cancellations:');
  label delay = 'Delay in Minutes';
run;

```

The box chart is shown in Output 29.1.1. The level of the *symbol-variable* NCANCEL determines the symbol marker for each subgroup mean, and the SYMBOLLEGEND= option controls the appearance of the legend for the symbols. The NOHLABEL option suppresses the label for the horizontal axis, and the NOLEGEND option suppresses the default legend for subgroup sample sizes.

Output 29.1.1. Box Chart 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 29.2. Creating Various Styles of Box-and-Whisker Plots

This example uses the flight delay data of the preceding example to illustrate how you can create box charts with various styles of box-and-whisker plots. For simplicity, the control limits are suppressed. The following statements create a chart, shown in Output 29.2.1, that displays *skeletal box-and-whisker plots*:

```

title 'Analysis of Airline Departure Delays';
symbol v=plus;
proc shewhart data=times limits=timelim ;

```

See SHWBOX5
in the SAS/QC
Sample Library

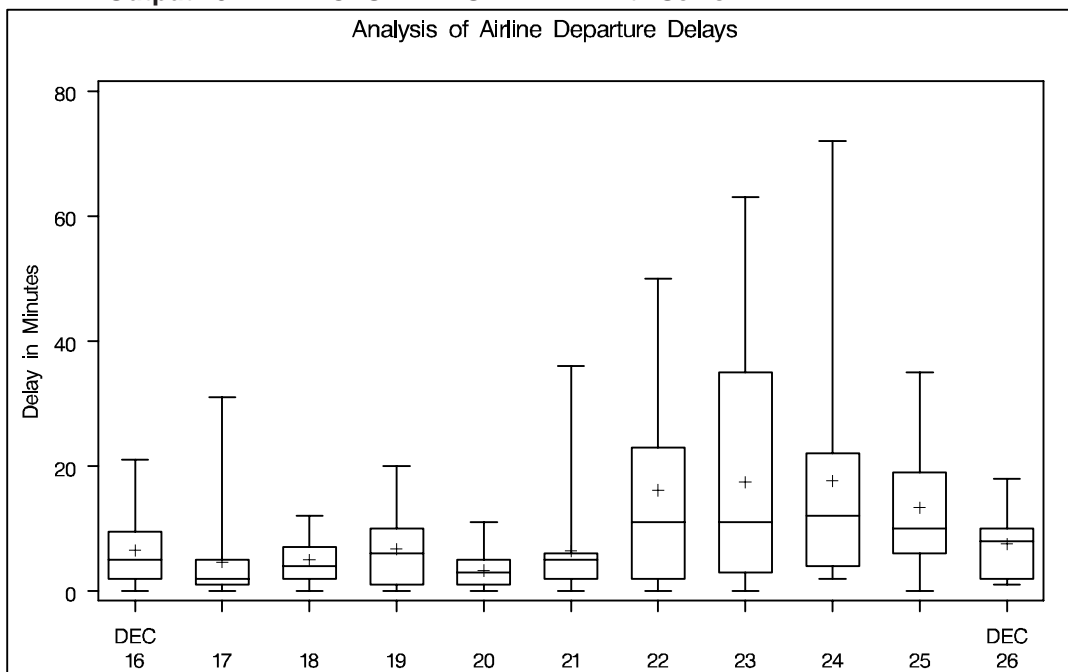
```

boxchart delay * day /
  boxstyle=skeletal
  serifs
  stddeviations
  nolimits
  nohlabel
  nolegend;
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 subgroup sample. You can also request this style by omitting the BOXSTYLE= option, since this style is the default. The SERIFS option adds serifs to the whiskers (by default, serifs are omitted with the skeletal style). The NOLIMITS option suppresses the display of the control limits. The STDDEVIATIONS option specifies that σ is to be estimated from subgroup standard deviations rather than subgroup ranges (you should specify this option with sample sizes greater than 10 or when using the NOLIMITS option to create standard side-by-side box-and-whisker plots).

Output 29.2.1. BOXSTYLE=SKELETAL with Serifs



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

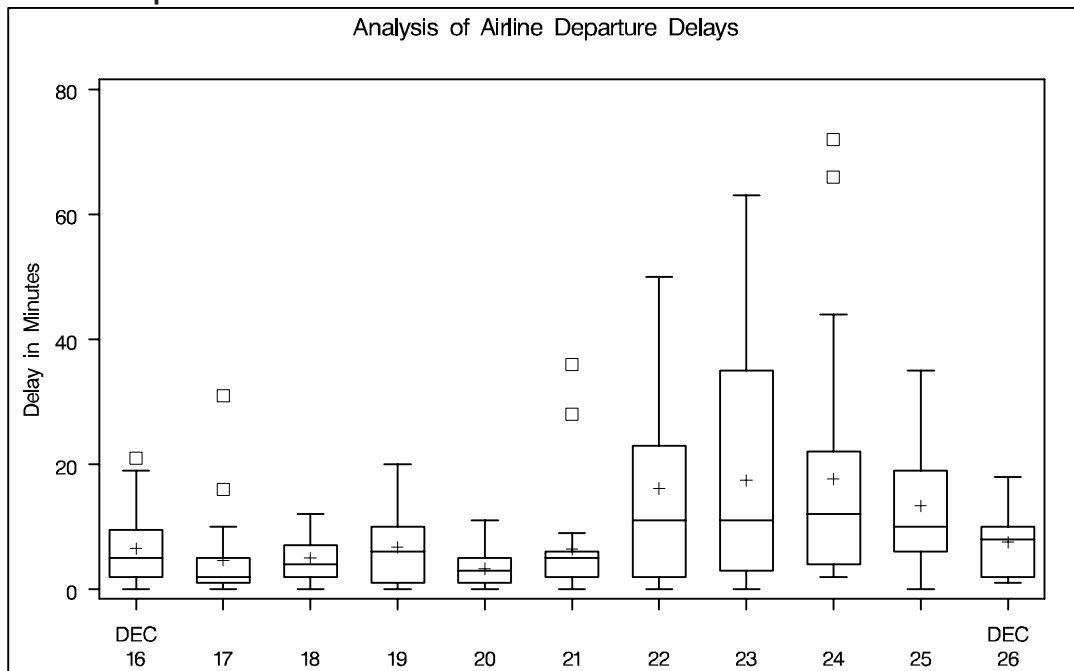
```

proc shewhart data=times limits=timelim ;
  boxchart delay * day /
    boxstyle=schematic
    stddeviations
    nolimits
    nohlabel
    nolegend;
label delay = 'Delay in Minutes';
run;

```

The chart is shown in Output 29.2.2. When `BOXSTYLE=SCHEMATIC` is specified, the whiskers are drawn to the most extreme points in the subgroup sample that lie within so-called “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 1739.

Output 29.2.2. `BOXSTYLE=SCHEMATIC`

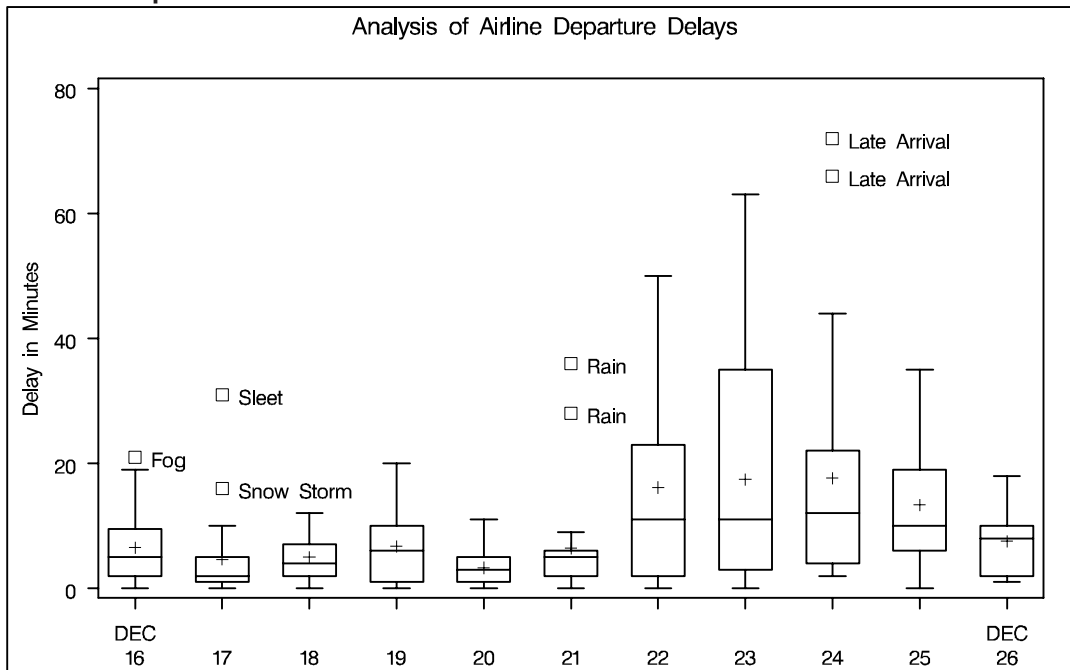


The following statements create a box chart with schematic box-and-whisker plots in which the observations outside the fences are labeled:

```
proc shewhart data=times limits=timelim ;
  boxchart delay * day /
    boxstyle=schematicid
    stddeviations
    nolimits
    nohlabel
    nolegend;
  id reason;
  label delay = 'Delay in Minutes';
run;
```

The chart is shown in Output 29.2.3. 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.

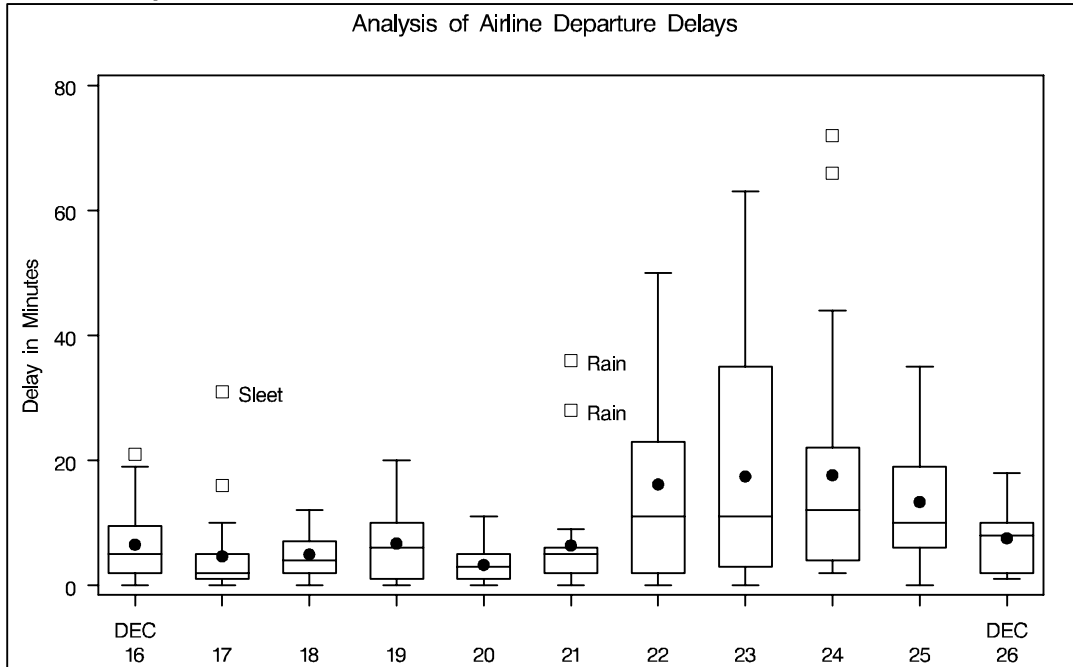
Output 29.2.3. BOXSTYLE=SCHEMATICID



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

```
proc shewhart data=times limits=timelim ;
  boxchart delay * day /
    boxstyle=schematicidfar
    stddeviations
    nolimits
    nohlabel
    nolegend;
  id reason;
  label delay = 'Delay in Minutes';
run;
```

The chart is shown in Output 29.2.4. If you specify BOXSTYLE=SCHEMATICIDFAR, schematic box-and-whisker 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 $3 \times \text{IQR}$ 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 29.2.4. BOXSTYLE=SCHEMATICIDFAR

Other options for controlling the display of box-and-whisker plots include the BOXWIDTH=, BOXWIDTHSCALE=, CBOXES=, CBOXFILL=, and LBOXES= options. For details, see the corresponding entries in Chapter 43, “Dictionary of Options.”

Example 29.3. Creating Notched Box-and-Whisker Plots

The following statements use the flight delay data of Example 29.1 to illustrate how to create side-by-side box-and-whisker plots with notches:

```

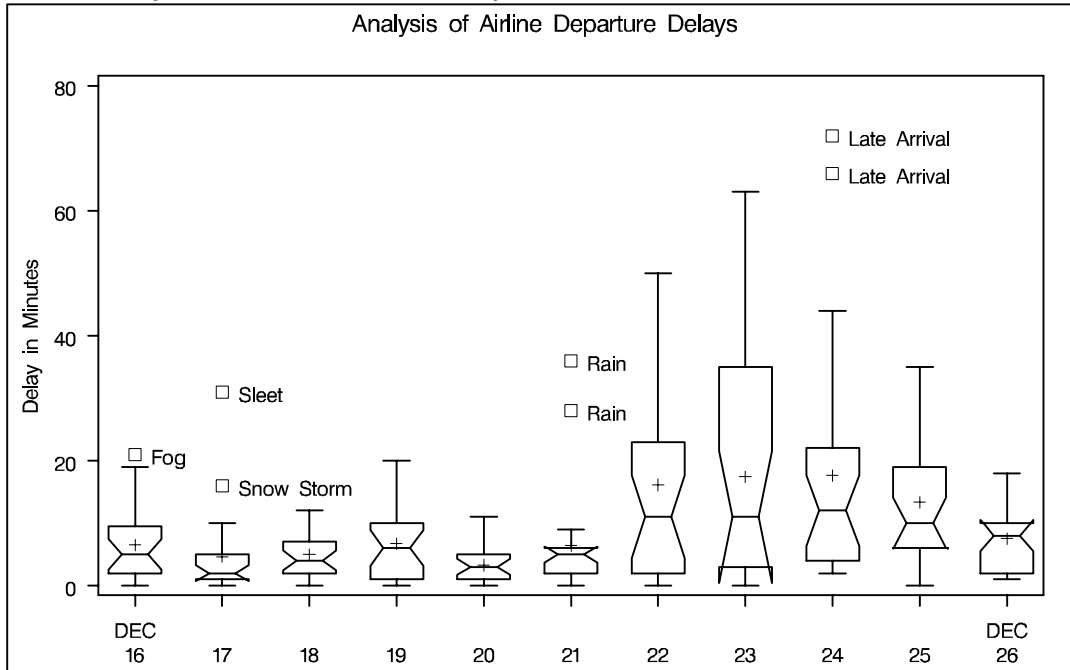
title 'Analysis of Airline Departure Delays';
symbol v=plus;
proc shewhart data=times limits=timelim ;
  boxchart delay * day /
    boxstyle = schematicid
    cboxfill = ligr
    stddeviations
    nohlabel
    nolegend
    nolimits
    notches;
  id reason;
  label delay = 'Delay in Minutes';
run;

```

See SHWBOX4
in the SAS/QC
Sample Library

The control limits are suppressed with the NOLIMITS option. The notches, requested with the NOTCHES option, measure the significance of the difference between two medians. The medians are significantly different at approximately the 95% level if the notches do not overlap. For details, see the entry for the NOTCHES option in Chapter 43, “Dictionary of Options.”

Output 29.3.1. Notched Side-by-Side Box-and-Whisker Plots



Example 29.4. Creating Box-and-Whisker Plots with Varying Widths

See SHWBOX7
in the SAS/QC
Sample Library

This example shows how to create a box chart with box-and-whisker plots whose widths vary proportionately with the subgroup sample 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;

cards;
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

```

```

08MAR90      3 . . . 7 . 6 4 0 . . .
             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 chart shown in Output 29.4.1:

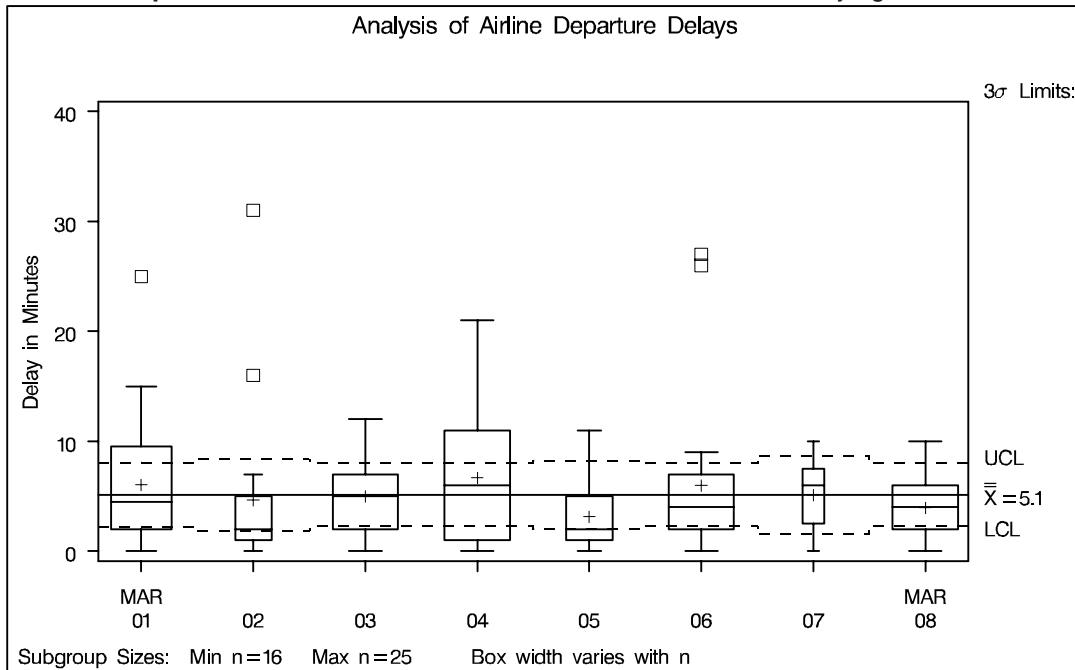
```

title 'Analysis of Airline Departure Delays';
symbol v=plus;
proc shewhart data=times2;
  boxchart delay * day /
    stddeviations
    nohlabel
    boxstyle      = schematic
    boxwidthscale = 1;
run;

```

The BOXWIDTHSCALE=1 option specifies that the widths of the box-and-whisker plots are to vary proportionately to the subgroup sample size n . This option is useful in situations where the sample size varies widely across subgroups. For further details, see the entry for the BOXWIDTHSCALE= option in Chapter 43, “Dictionary of Options.”

Output 29.4.1. Box Chart with Box-and-Whisker Plots of Varying Widths



Example 29.5. Creating Box-and-Whisker Plots with Different Line Styles and Colors

See SHWBOX7
in the SAS/QC
Sample Library

The control limits in Output 29.4.1 apply to the subgroup means. This example illustrates how you can modify the chart to indicate whether the variability of the process is in control. The following statements create a box chart for DELAY in which a dashed outline and a light gray fill color are used for a box-and-whisker plot if the corresponding subgroup standard deviation exceeds its 3σ limits.

First, the SHEWHART procedure is used to create an OUTTABLE= data set (DELAYTAB) that contains a variable (_EXLIMS_) that records which standard deviations exceed their 3σ limits.

```
proc shewhart data=times2;
  xschart delay * day / nochart
                        outtable = delaytab;
run;
```

Then, this information is used to set the line styles and fill colors as follows:

```
data delaytab;
  length boxcol $ 8;
  set delaytab;
  keep day lnstyle boxcol;
  if _exlims_ = 'UPPER' or _exlims_ = 'LOWER' then do;
    lnstyle = 2;
    boxcol  = 'ligr';
  end;
  else do;
    lnstyle = 1;
    boxcol  = 'dagr';
  end;

data times2;
  merge times2 delaytab;
  by day;
run;
```

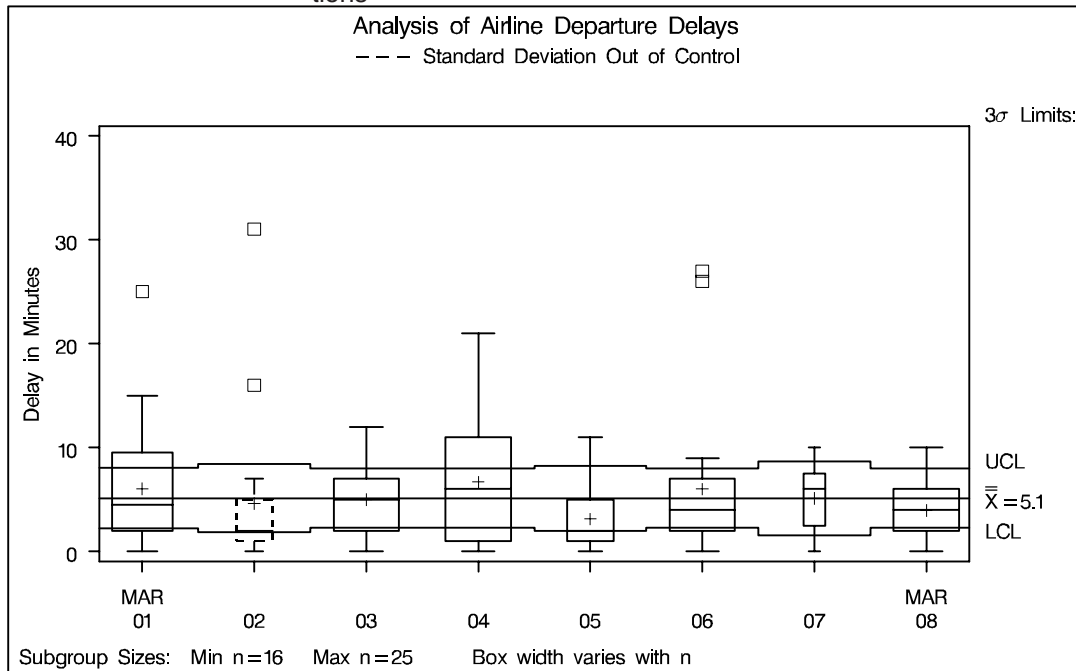
The following statements create the modified box chart:

```
title 'Analysis of Airline Departure Delays' ;
title2 '--- Standard Deviation Out of Control';
symbol v=plus c=yellow;

proc shewhart data=times2;
  boxchart delay * day /
    stddeviations
    nohlabel
    boxstyle      = schematic
    llimits       = 1
    cboxfill      = ( boxcol )
    lboxes        = ( lnstyle )
    boxwidthscale = 1 ;
run;
```

The chart is shown in Output 29.5.1. The values of the variable LNSTYLE specified with the LBOXES= option determine the outline styles for the box-and-whisker plots. The values of the variable BOXCOL specified with the CBOXFILL= option determines the fill colors. For further details, see the entries for these options in Chapter 43, “Dictionary of Options,”. The chart indicates that the large variability for March 2 should be checked.

Output 29.5.1. Box Chart Displaying Out-of-Control Subgroup Standard Deviations



Example 29.6. Computing the Control Limits for Subgroup Maximums

This example illustrates how to compute and display control limits for the *maximum* of a subgroup sample. Subgroup samples of 20 metal braces are collected daily, and the lengths of the braces are measured in centimeters. These data are analyzed extensively in Example 41.3 on page 1706. The box chart for LOGLENG (the log of length) shown in Output 41.3.3 on page 1709 indicates that the subgroup mean is in control and that the subgroup distributions of LOGLENG are approximately normal. The following statements save the control limits for the mean of the LOGLENG in a data set named LOGLLIMS:

See SHWBOX3
 in the SAS/QC
 Sample Library

```
data lengdata;
  set lengdata;
  logleng=log(length-105);

proc shewhart data=lengdata;
  xchart logleng*day /
  nochart
  outlimits=logllims;
run;
```

Part 9. The CAPABILITY Procedure

The next statements replace the control limits for the mean of LOGLENG with control limits for the maximum of LOGLENG:

```
data maxlim;
  set lengdata;
  set logllims;
  drop avgmax stdmax;
  label _lclx_ = 'Lower Limit for Maximum of 20'
        _uclx_ = 'Upper Limit for Maximum of 20'
        _mean_ = 'Central Line for Maximum of 20';
  avgmax = _stddev_*1.86747 + _mean_;
  stdmax = _stddev_*0.52509;
  _lclx_ = avgmax - _sigmas_*stdmax;
  _uclx_ = avgmax + _sigmas_*stdmax;
  _mean_ = avgmax;
  call symput('avgmax',left(put(avgmax,8.1)));
run;
```

The control limits are computed using the fact that the maximum of a sample of size 20 from a normal population with zero mean and unit standard deviation has an expected value of 1.86747 and a standard deviation of 0.52509; refer to Teichroew (1956) and see Table 29.27 on page 1159. Finally, the following statements create a box chart for LOGLENG that displays control limits for the subgroup maximum:

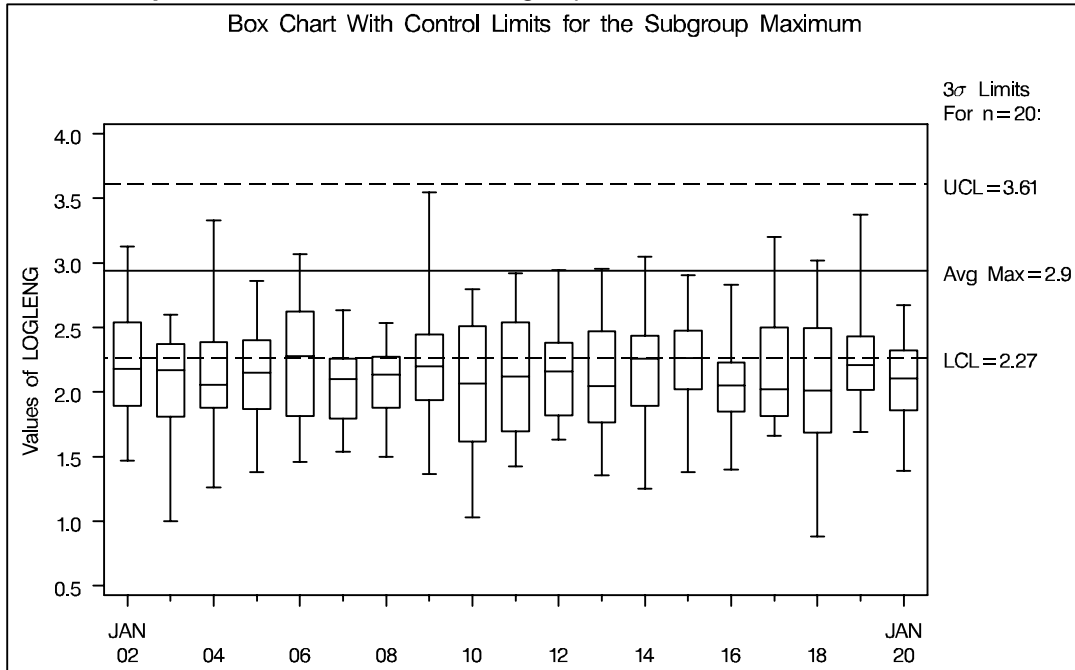
```
title 'Box Chart With Control Limits for the Subgroup Maximum';
symbol v=none;

proc shewhart data=lengdata limits=maxlim;
boxchart logleng*day /
  cboxfill = gray
  cinfill  = ligr
  serifs
  nolegend
  nohlabel
  xsymbol  = "Avg Max=&AVGMAX";
  label logleng='Values of LOGLENG';
run;
```

The box chart, shown in Output 29.6.1, indicates that the maximum is in control since the tips of the upper whiskers fall within the control limits.

The SYMPUT call is used to pass the value of `_MEAN_` in a macro variable to the SHEWHART procedure so that this value can be used to label the central line.

You can apply the variable replacement method shown here to data with sample sizes other than 20 by replacing the constants 1.86747 and 0.52509 with the appropriate values from Table 29.27. Austin (1973) describes a method for approximating these values. You can also use the preceding statements to display control limits for the subgroup minimum by changing the sign of the expected values in Table 29.27.

Output 29.6.1. Box Chart for Subgroup Maximum

The variable replacement method can also be used to create a variety of box charts, including the modifications suggested by Iglewicz and Hoaglin (1987) and Rocke (1989).

Table 29.27. Expected Values and Standard Deviations of Maximum of a Normal Sample

n	Expected Value	Standard Deviation
2	0.56418	0.82565
3	0.84628	0.74798
4	1.02937	0.70123
5	1.16296	0.66899
6	1.26720	0.64494
7	1.35217	0.62605
8	1.42360	0.61065
9	1.48501	0.59780
10	1.53875	0.58681
11	1.58643	0.57730
12	1.62922	0.56891
13	1.66799	0.56144
14	1.70338	0.55474
15	1.73591	0.54869
16	1.76599	0.54316
17	1.79394	0.53809
18	1.82003	0.53342
19	1.84448	0.52910
20	1.86747	0.52509

Example 29.7. Constructing Multi-Vari Charts

“Multi-vari” charts* are used in a variety of industries to analyze process data with nested (hierarchical) patterns of variation

- within-sample variation (for example, position within wafer)
- sample-to-sample variation within batches of samples (for example, wafer within lot)
- batch-to-batch variation (for example, across lots)

This example illustrates the construction of a “multi-vari” display. The following statements create a SAS data set named PARM that contains the value of a measured parameter (MEASURE) recorded at each of five positions on wafers produced in lots.

```
data parm;
  length _phase_ $ 5 wafer $ 2 position $ 1;
  input  _phase_ $ & wafer $ & position $ measure ;
datalines;
  Lot A    01    L    2.424
  Lot A    01    B    2.441
  Lot A    01    C    2.421
  Lot A    01    T    2.449
  Lot A    01    R    2.500
  Lot A    02    L    2.681
  Lot A    02    B    2.571
  Lot A    02    C    2.546
  Lot A    02    T    2.659
  Lot A    02    R    2.692
  Lot A    03    L    2.180
  Lot A    03    B    2.135
  Lot A    03    C    2.443
  Lot A    03    T    2.290
  Lot A    03    R    2.259
  Lot B    01    L    2.465
  Lot B    01    B    2.448
  Lot B    01    C    2.523
  Lot B    01    T    2.744
  Lot B    01    R    2.883
  Lot B    02    L    2.372
  Lot B    02    B    2.145
  Lot B    02    C    2.531
  Lot B    02    T    2.474
  Lot B    02    R    2.562
  Lot B    03    L    2.933
  Lot B    03    B    2.325
  Lot B    03    C    2.528
```

*Multi-vari charts should not be confused with multivariate control charts , which are discussed on page 1927.

```

Lot B    03    T    2.603
Lot B    03    R    2.684
.        .    .    .
.        .    .    .
.        .    .    .
Lot G    03    R    2.843
run;

```

The following statements create an ordinary side-by-side box-and-whisker display for the measurements.

```

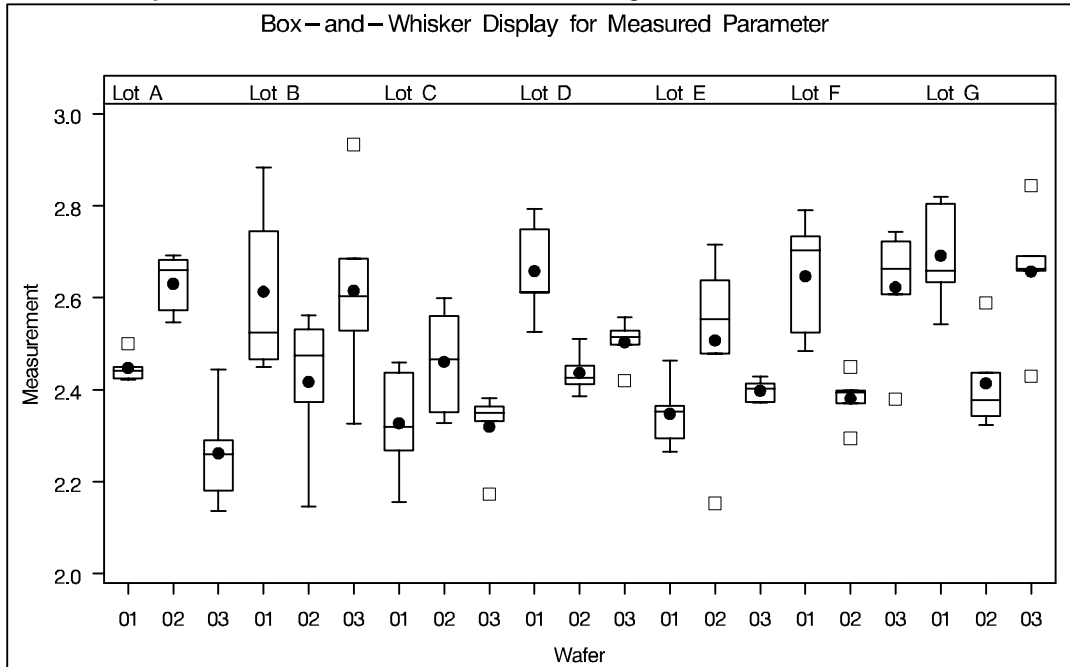
title 'Box-and-Whisker Display for Measured Parameter';

proc shewhart data=parm;
  boxchart parm*wafer /
    stddevs
    nolimits
    cboxes    = black
    boxstyle  = schematic
    idsymbol  = square
    readphase = all
    phaselegend
    nolegend;
  label measure = 'Measurement'
         wafer   = 'Wafer Within Lot';
run;

```

The display is shown in Output 29.7.1. Here, the *subgroup-variable* is WAFER, and the option BOXSTYLE=SCHEMATIC is specified to request schematic box-and-whisker plots for the measurements in each subgroup (wafer) sample. The lot values are provided as the values of the special variable `_PHASE_`, which is read when the option READPHASE=ALL is specified. The option PHASELEGEND requests the legend for phase (lot) values at the top of the chart, and the NOLEGEND option suppresses the default legend for sample sizes. The NOLIMITS option suppresses the display of control limits, and the STDDEVS option is used to base the estimate of the process standard deviation on subgroup standard deviations rather than subgroup ranges. These two options are recommended whenever you are using the BOXCHART statement to create side-by-side box-and-whisker plots.

Output 29.7.1. Box-and-Whisker Plot Using BOXSTYLE=SCHEMATIC



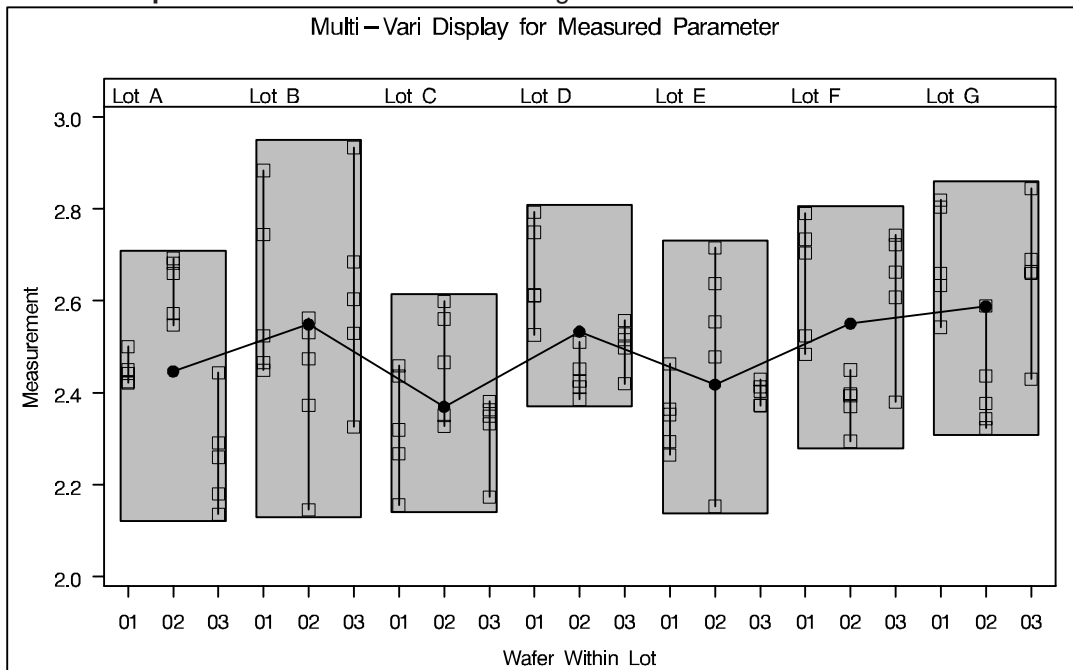
The box-and-whisker display in Output 29.7.1 is not particularly appropriate for these data since there are only five measurements in each wafer and since the variation within each wafer may depend on the position, which is not indicated. The next statements use the BOXCHART statement to produce a multi-vari chart for the same data.

```

title 'Multi-Vari Display for Measured Parameter';
symbol v=none;
proc shewhart data=parm;
  boxchart measure*wafer /
    stddevs
    nolimits
    boxstyle          = pointsjoin
    cboxes            = black
    idsymbol          = square
    cphaseboxfill    = ligr
    cphasebox        = black
    cphasemeanconnect = black
    phasemeanconnect = black
    phasemeanconnect = black
    phasemeanconnect = black
    phasemeanconnect = black
    readphase        = all
    phaselegend
    nolegend;
  label measure = 'Measurement'
        wafer   = 'Wafer Within Lot';
run;

```

The display is shown in Output 29.7.2.

Output 29.7.2. Multi-Vari Chart Using BOXSTYLE=POINTSJOIN

The option `BOXSTYLE=POINTSJOIN` specifies that the values for each wafer are to be displayed as points joined by a vertical line. The `IDSYMBOL=` option specifies the symbol marker for the points. The option `V=NONE` in the `SYMBOL` statement is specified to suppress the symbol for the wafer averages shown in Output 29.7.1. The option `CPHASEBOX=BLACK` specifies that the points for each lot are to be enclosed in a black box, and the `CPHASEBOXFILL=` option specifies the fill color for the box. The option `CPHASEMEANCONNECT=BLACK` specifies that the means of the lots are to be connected with black lines, and the `PHASEMEANSYMBOL=` option specifies the symbol marker for the lot means.

The following statements create a slightly different multi-vari chart using the values of the variable `POSITION` to identify the measurements for each wafer. Note that the option `BOXSTYLE=POINTS` is specified and that `POSITION` is specified as the ID variable. The display is shown in Output 29.7.3.

```
proc shewhart data=parm;
  boxchart measure*wafer /
    nolimits
    stddevs
    cboxes          = black
    cphaseboxfill  = ligr
    cphasemeanconnect = black
    boxstyle       = pointsid
    phasemeanconnect = black
    phasemeanconnect = black
    readphase      = all
    phaselegend
    nolegend;
  label measure = 'Measurement'
        wafer   = 'Wafer Within Lot';
  id position;
```

Output 29.7.3. Multi-Vari Chart Using BOXSTYLE=POINTSID

