

Chapter 3

COMPHISTOGRAM Statement

Chapter Table of Contents

OVERVIEW	117
GETTING STARTED	118
Creating a One-Way Comparative Histogram	118
Adding Fitted Normal Curves to a Comparative Histogram	119
SYNTAX	121
Summary of Options	122
Dictionary of Options	126
EXAMPLES	140
Example 3.1 Adding Insets with Descriptive Statistics	140
Example 3.2 Creating a Two-Way Comparative Histogram	141

Chapter 3

COMPHISTOGRAM Statement

Overview

Comparative histograms are useful for comparing the distribution of a process variable across levels of classification variables. You can use the COMPHISTOGRAM statement to create one-way and two-way comparative histograms. When used with a single classification variable, the COMPHISTOGRAM statement displays an array of component histograms (stacked or side-by-side), one for each level of the classification variable. When used with two classification variables, the COMPHISTOGRAM statement displays a matrix of component histograms, one for each combination of levels of the classification variables.

In quality improvement applications, typical uses of comparative histograms include

- comparing the capability of a process before and after an improvement
- comparing process capabilities of two or more suppliers
- exploring stratification in process data due to different lots, machines, manufacturing methods, and so forth
- studying the evolution of process capability over successive time periods

You can use options in the COMPHISTOGRAM statement to

- specify the midpoints for histogram intervals
- specify the number of rows and/or columns of component histograms
- display specification limits on the component histograms
- display density curves for fitted normal distributions
- display kernel density estimates
- request graphical enhancements
- inset summary statistics and process capability indices on the component histograms

Getting Started

This section introduces the COMPHISTOGRAM statement with examples that illustrate commonly used options. Complete syntax for the COMPHISTOGRAM statement is presented in the “Syntax” section on page 121, and advanced examples are given in the “Examples” section on page 140.

Creating a One-Way Comparative Histogram

The effective channel length (in microns) is measured for 1225 field effect transistors. Both the channel length (LENGTH) and the lot source (LOT) are saved in a SAS data set named CHANNEL. A partial listing of CHANNEL is shown in Figure 3.1.

Obs	lot	length
1	Lot 1	0.90979
2	Lot 1	1.01131
3	Lot 1	0.95001
4	Lot 1	1.12591
5	Lot 1	1.11707
6	Lot 1	0.86177
7	Lot 1	0.96033
8	Lot 1	1.16649
9	Lot 1	1.35797
10	Lot 1	1.09681
.	.	.
.	.	.
.	.	.
1224	Lot 3	1.74088
1225	Lot 3	1.91107

Figure 3.1. Partial Listing of the Data Set CHANNEL

The data set CHANNEL is also used in Example 4.5 on page 203, where a kernel density estimate is superimposed on the histogram of channel lengths. The display in Output 4.5.2 on page 204 reveals that there are three distinct peaks in the process distribution. To investigate whether these peaks (modes) in the histogram are related to the lot source, you can create a comparative histogram using LOT as a classification variable. The following statements create the comparative histogram shown in Figure 3.2:

```

title 'Comparative Analysis of Lot Source';
proc capability data=channel noprint;
  spec lsl = 0.8  llsl = 2
      usl = 2.0  lusl = 3;
  comphistogram length / class      = lot
                        nrows      = 3
                        nlegend    = 'Lot Size'
                        nlegendpos = nw ;
  label lot = 'Transistor Source' ;
run;

```

The COMPHISTOGRAM statement requests a comparative histogram for the process variable LENGTH. The CLASS= option requests a component histogram for each level (distinct value) of the classification variable LOT. The option NROWS=3

stacks the histograms three to a page. The NLEGEND= option adds a sample size legend to each component histogram, and the option NLEGENDPOS=NW positions each legend in the northwest corner. The SPEC statement provides the specification limits displayed as vertical reference lines. See “Dictionary of Options” on page 126 for descriptions of these options, and see “Syntax for the SPEC Statement” on page 54 for details of the SPEC statement.

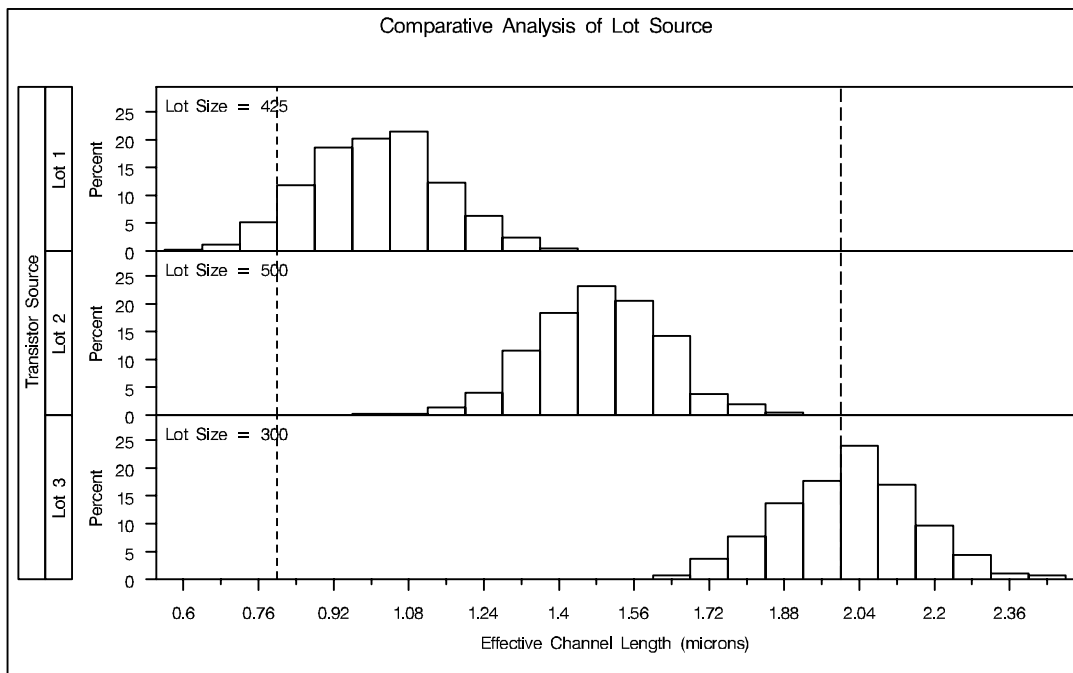


Figure 3.2. Comparison by Lot Source

Adding Fitted Normal Curves to a Comparative Histogram

In Figure 3.2, it appears that each lot produces transistors with channel lengths that are normally distributed. The following statements use the NORMAL option to fit a normal distribution to the data for each lot (the observations corresponding to a specific level of the classification variable are referred to as a *cell*). The normal parameters μ and σ are estimated from the data for each lot, and the curves are superimposed on each component histogram.

```
proc capability data=channel noprint;
  spec lsl = 0.8  lls1 = 2
        usl = 2.0  lus1 = 3;
  comphistogram length / class      = lot
                          nrows     = 3
                          intertile = 1
                          cprop     = orange
                          normal ;
  label lot = 'Transistor Source';
run;
```

See CAPCMH1
in the SAS/QC
Sample Library

Part 1. The CAPABILITY Procedure

The comparative histogram is displayed in Figure 3.3.

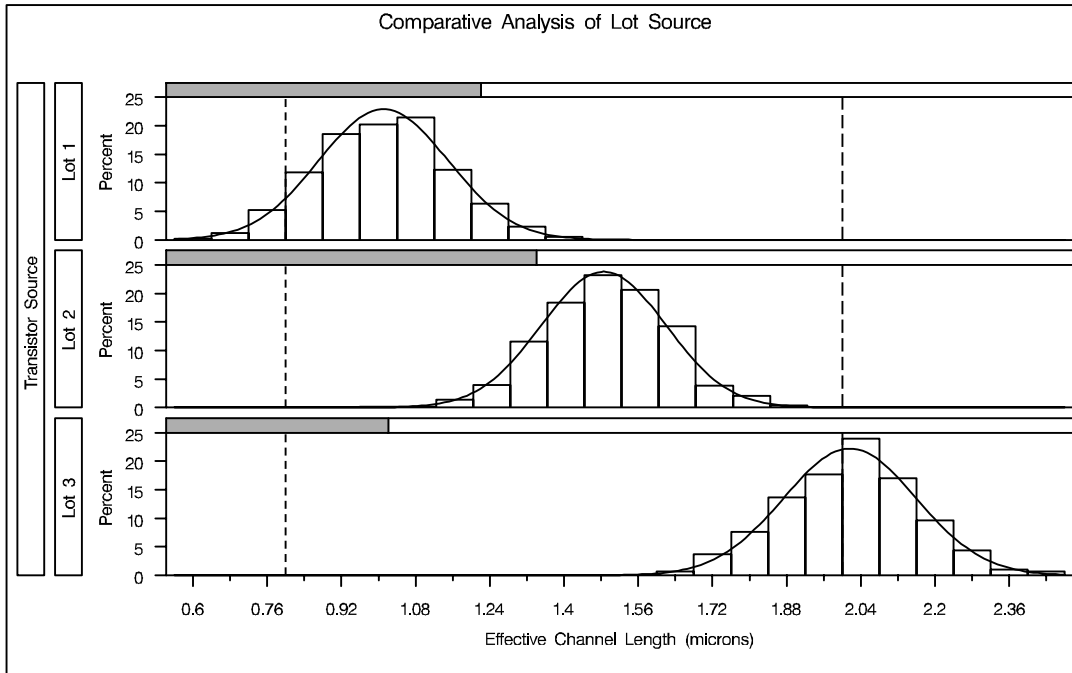


Figure 3.3. Fitting Normal Curves

Specifying `INTERTILE=1` inserts a space of one percent screen unit between the framed areas, which are referred to as *tiles*. The shaded bars, added with the `CPROP=` option, represent the relative frequency of observations in each cell. See “Dictionary of Options” on page 126 for details concerning these options.

Syntax

The syntax for the COMPHISTOGRAM statement is as follows:

```
COMPHISTOGRAM <variables > / CLASS=(class-variables) <options >;
```

You can specify the keyword COMPHIST as an alias for COMPHISTOGRAM. You can use any number of COMPHISTOGRAM statements after a PROC CAPABILITY statement.

To create a comparative histogram, you must specify at least one *variable* and either one or two *class-variables* (also referred to as *classification variables*). * The COMPHISTOGRAM statement displays a component histogram for each level of the *class-variables* using the values of the *variable*. The observations in a given level are referred to as a *cell*.

The components of the COMPHISTOGRAM statement are described as follows:

variables

are the process variables for which comparative histograms are to be created. If you specify a VAR statement, the *variables* must also be listed in the VAR statement. Otherwise, *variables* can be any numeric variables in the input data set that are not also listed as *class-variables*. If you do not specify *variables* in a COMPHISTOGRAM statement or a VAR statement, then by default a comparative histogram is created for each numeric variable in the DATA= data set that is not used as a *class-variable*. If you use a VAR statement and do not specify *variables* in the COMPHISTOGRAM statement, then by default a comparative histogram is created for each variable listed in the VAR statement.

For example, suppose a data set named STEEL contains two process variables named LENGTH and WIDTH, a numeric classification variable named LOT, and a character classification variable named DAY. The following statements create two comparative histograms, one for LENGTH and one for WIDTH:

```
proc capability data=steel;
  comphist / class = lot;
run;
```

Likewise, the following statements create comparative histograms for LENGTH and WIDTH:

```
proc capability data=steel;
  var length width;
  comphist / class = day;
run;
```

The following statements create three comparative histograms (for LENGTH, WIDTH, and LOT):

*In Release 6.12 and in previous releases of SAS/QC software, the keyword GRAPHICS was required in the PROC CAPABILITY statement since the COMPHISTOGRAM statement creates output only for high resolution graphics devices.

Part 1. The CAPABILITY Procedure

```
proc capability data=steel;
  comphist / class = day;
run;
```

The following statements create a comparative histogram for WIDTH only:

```
proc capability data=steel;
  var length width;
  comphist width / class=lot;
run;
```

class-variables

are one or two required classification variables. For example, the following statements create a one-way comparative histogram for WIDTH using the classification variable LOT:

```
proc capability data=steel;
  comphist width / class=lot;
run;
```

The following statements create a two-way comparative histogram for WIDTH classified by LOT and DAY:

```
proc capability data=steel;
  comphist width / class=(lot day);
run;
```

Note that the parentheses surrounding the *class-variables* are needed only if two classification variables are specified. See Output 3.1.1 on page 141 and Output 3.2.1 on page 143 for further examples.

options

control the features of the comparative histogram. All *options* are specified after the slash (/) in the COMPHIST statement. In the following example, the CLASS= option specifies the classification variable, the NORMAL option fits a normal density curve in each cell, and the CTEXT= option specifies the color of the text:

```
proc capability data=steel;
  comphist length / class = lot
                    normal
                    ctext = yellow;
run;
```

Summary of Options

The following tables list the COMPHIST statement options by function. For complete descriptions, see “Dictionary of Options” on page 126.

Normal Curve Options

Table 3.1 summarizes options that specify features of fitted normal distributions requested with the NORMAL option. Specify these options in parentheses after the NORMAL option.

Table 3.1. Normal-Options

COLOR= <i>color</i>	specifies color of normal curve
FILL	fills area under normal curve
L= <i>linetype</i>	specifies line type of normal curve
MU= <i>value</i>	specifies mean μ for fitted normal curve
SIGMA= <i>value</i>	specifies standard deviation σ for fitted normal curve
W= <i>n</i>	specifies width of normal curve

For example, the following statements use the NORMAL option to fit a normal curve in each cell of the comparative histogram:

```
proc capability;
  comphistogram / class = machine
                 normal(color=red l=2);
run;
```

The COLOR= *normal-option* draws the curve in red, and the L= *normal-option* specifies a line style of 2 (a dashed line) for the curve. In this example, maximum likelihood estimates are computed for the normal parameters μ and σ for each cell since these parameters are not specified.

Kernel Options

You can specify the options listed in Table 3.2 in parentheses after the keyword KERNEL to control features of kernel density estimates requested with the KERNEL option.

Table 3.2. Kernel Options

C= <i>value-list</i> MISE	specifies standardized bandwidth parameter c for kernel density estimate
COLOR= <i>color</i>	specifies color of the kernel density curve
FILL	fills area under kernel density curve
K= <i>keyword</i>	specifies NORMAL, TRIANGULAR, or QUADRATIC kernel
L= <i>linetype</i>	specifies line type used for kernel density curve
W= <i>n</i>	specifies line width for kernel density curve

General Options**Table 3.3.** Comparative Histogram Layout Options

ANNOKEY	applies annotation requested in ANNOTATE= data set to key cell only
BARWIDTH= <i>n</i>	specifies width for the bars
CLASS=(<i>variables</i>)	specifies classification variables
CLASSKEY=('values')	specifies key cell
HOFFSET= <i>value</i>	specifies offset for horizontal axis
INTERTILE= <i>value</i>	specifies distance between tiles
MAXNBIN= <i>n</i>	specifies maximum number of bins displayed
MAXSIGMAS= <i>value</i>	limits number of bins displayed to range of <i>value</i> standard deviations above and below mean of data in key cell
MIDPOINTS= <i>values</i> KEY UNIFORM	specifies how midpoints are determined
MISSING1	requests that missing values of first CLASS= variable be treated as a level of that CLASS= variable
MISSING2	requests that missing values of second CLASS= variable be treated as a level of that CLASS= variable
NCOLS= <i>n</i>	specifies number of columns in comparative histogram
NOBARS	suppresses histogram bars
NOFRAME	suppresses frame around plotting area
NOKEYMOVE	suppresses rearrangement of cells that occurs by default with the CLASSKEY= option
NOPLOT	suppresses plot
NROWS= <i>n</i>	specifies number of rows in comparative histogram
ORDER1=DATA FORMATTED FREQ INTERNAL	specifies display order for values of the first CLASS= variable
ORDER2=DATA FORMATTED FREQ INTERNAL	specifies display order for values of the second CLASS= variable
RTINCLUDE	includes right endpoint in interval
TILELEGLABEL='string'	specifies label displayed when _CTILE_ and _TILELG_ variables are provided in the CLASSSPEC= data set
TURNVLABELS	turns and strings out vertically characters in labels for vertical axis
WBARLINE= <i>n</i>	specifies line thickness for bar outlines

Table 3.4. Reference Line Options

HREF= <i>value-list</i>	specifies reference lines perpendicular to horizontal axis
HREFLABELS= <i>'label1' ... 'labeln'</i>	specifies labels for HREF= lines
HREFLABPOS= <i>n</i>	specifies vertical position of labels for HREF= lines
LHREF= <i>linetype</i>	specifies line style for HREF= lines
LVREF= <i>linetype</i>	specifies line style for VREF= lines
VREF= <i>value-list</i>	specifies reference lines perpendicular to vertical axis
VREFLABELS= <i>'label1' ... 'labeln'</i>	specifies labels for VREF= lines
VREFLABPOS= <i>n</i>	specifies horizontal position of labels for VREF= lines

Table 3.5. Text Enhancement Options

FONT= <i>font</i>	specifies software font for text
HEIGHT= <i>value</i>	specifies height of text used outside framed areas
INFONT= <i>font</i>	specifies software font for text inside framed areas
INHEIGHT= <i>value</i>	specifies height of text inside framed areas

Table 3.6. Axis and Legend Options

GRID	adds grid corresponding to vertical axis
LGRID= <i>linetype</i>	specifies line style for grid requested with GRID option
NLEGEND<= <i>'string'</i> >	specifies form of the legend displayed inside tiles
NLEGENDPOS=NE NW	specifies position of legend displayed inside tiles
NOHLABEL	suppresses label for horizontal axis
NOVLABEL	suppresses label for vertical axis
NOVTICK	suppresses tick marks and tick mark labels for vertical axis
VAXIS= <i>value-list</i>	specifies tick mark values for vertical axis
VAXISLABEL= <i>'string'</i>	specifies label for vertical axis
VOFFSET= <i>value</i>	specifies length of offset at upper end of vertical axis
VSCALE=COUNT PERCENT PROPORTION	specifies scale for vertical axis
WAXIS= <i>n</i>	specifies line thickness for axes and frame
WGRID= <i>n</i>	specifies line thickness for grid

Table 3.7. Graphics Catalog Options

DESCRIPTION= <i>'string'</i>	specifies description for graphics catalog member
NAME= <i>'string'</i>	specifies name for plot in graphics catalog

Table 3.8. Color and Pattern Options

CAXIS= <i>color</i>	specifies color for axis
CBARLINE= <i>color</i>	specifies color for outline of the bars
CFILL= <i>color</i>	specifies color for filling bars
CFRAME= <i>color</i>	specifies color for frame
CFRAMENLEG= <i>color</i>	specifies the color for the frame requested by the NLEGEND option
CFRAMESIDE= <i>color</i>	specifies color for filling frame for row labels
CFRAMETOP= <i>color</i>	specifies color for filling frame for column labels
CGRID= <i>color</i>	specifies color for grid lines
CHREF= <i>color</i>	specifies color for HREF= lines
CPROP= <i>color</i>	specifies color for proportion of frequency bar
CTEXT= <i>color</i>	specifies color for text
CVREF= <i>color</i>	specifies color for VREF= lines
PFILL= <i>pattern</i>	specifies pattern used to fill bars

Table 3.9. Input and Output Data Sets

ANNOTATE= <i>SAS-data-set</i>	annotate data set
CLASSSPEC= <i>SAS-data-set</i>	data set with specification limit information for each cell
OUTHISTOGRAM= <i>SAS-data-set</i>	information on histogram intervals

Dictionary of Options

The following entries describe the *options* in detail. All options apply with high resolution graphics output.

ANNOKEY

specifies that annotation requested with the ANNOTATE= option is to be applied only to the *key cell*. By default, annotation is applied to all of the cells. Use the CLASSKEY= option to specify the key cell.

ANNOTATE=*SAS-data-set*

ANNO=*SAS-data-set*

specifies an input data set containing annotate variables as described in *SAS/GRAPH Software: Reference*. You can use this data set to add features to the comparative his-

togram. The ANNOTATE= data set you specify in the COMPHISTOGRAM statement is used for all plots created by the statement. You can also specify an ANNOTATE= data set in the PROC CAPABILITY statement to enhance all plots created by the procedure; for more information, see “ANNOTATE= Data Sets” on page 59.

BARWIDTH=*value*

specifies the width of the histogram bars in screen percent units.

C=*value-list* | **MISE**

specifies the standardized bandwidth parameter c for kernel density estimates requested with the KERNEL option. You can specify up to five *values* to display multiple estimates in each cell. You can also specify the keyword MISE to request the bandwidth parameter that minimizes the estimated mean integrated square error (MISE). For example, consider the following statements (for more information, see “Kernel Density Estimates” on page 179):

```
proc capability;
  comphist length / class=batch kernel(c = 0.5 1.0 mise);
run;
```

The KERNEL option displays three density estimates. The first two have standardized bandwidths of 0.5 and 1.0, respectively. The third has a bandwidth parameter that minimizes the MISE. You can also use the C= and K= options (K= specifies kernel type) to display multiple estimates. For example, consider the following statements:

```
proc capability;
  comphist length / class = batch
                    kernel(c = 0.75 k = normal triangular);
run;
```

Here two estimates are displayed. The first uses a normal kernel and bandwidth parameter of 0.75, and the second uses a triangular kernel and a bandwidth parameter of 0.75. In general, if more kernel types are specified than bandwidth parameters, the last bandwidth parameter in the list will be repeated for the remaining estimates. Likewise, if more bandwidth parameters are specified than kernel types, the last kernel type will be repeated for the remaining estimates. The default is MISE.

CAXIS=*color***CAXES=***color***CA=***color*

specifies the color for the axes, tick marks, and target line. The default is the first color in the device color list.

CBARLINE=*color*

specifies the color of the outline of the histogram bars. This option overrides the C= option in the SYMBOL1 statement. The default is the first color in the device color list.

CFILL=*color*

specifies a color used to fill the bars of the histograms (or the areas under a fitted

curve if you also specify the FILL option). See the entry for the FILL option for additional details. See Output 3.1.1 on page 141 and Example 3.2 on page 141 for examples. Refer to *SAS/GRAPH Software: Reference* for a list of colors. By default, bars and curve areas are not filled.

CFRAME=*color*

specifies the color for the area enclosed by the axes and the frame. This area is not filled by default. The CFRAME= option cannot be used with the NOFRAME option, the CTILES= option, or the variable `_CTILE_` in a CLASSSPEC= data set.

CFRAMENLEG=*color* | **EMPTY**

specifies that the legend requested with the NLEGEND option (or the variable `_TILELB_` in a CLASSSPEC= data set) is to be framed and that the frame is to be filled with the color indicated. If you specify CFRAMENLEG=EMPTY, a frame is drawn but not filled with a color.

CFRAMESIDE=*color*

specifies the color for filling the frame area for the row labels displayed along the left side of a comparative histogram requested with the CLASS= option. This color is also used to fill the frame area for the label of the corresponding CLASS= variable (if a label is associated with the variable.) See Output 3.2.1 on page 143 for an example. By default, these areas are not filled.

CFRAMETOP=*color*

specifies the color for filling the frame area for the column labels displayed across the top of a comparative histogram requested with the CLASS= option. This color is also used to fill the frame area for the label of the corresponding CLASS= variable (if a label is associated with the variable.) See Output 3.2.1 on page 143 for an example. By default, these areas are not filled.

CGRID=*color*

specifies the color for grid lines requested with the GRID option. The default is the first color in the device color list. If you use CGRID=, you do not need to specify the GRID option.

CHREF=*color*

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

CLASS=*variable*

CLASS=(*variable1 variable2***)**

specifies that a comparative histogram is to be created using the levels of the *variables* (also referred to as *class-variables* or *classification variables*).

If you specify a single *variable*, a one-way comparative histogram is created. The observations in the input data set are sorted by the formatted values (levels) of the variable. A separate histogram is created for the process variable values in each level, and these component histograms are arranged in an array to form the comparative histogram. Uniform horizontal and vertical axes are used to facilitate comparisons. For an example, see Figure 3.2 on page 119.

If you specify two *classification variables*, a two-way comparative histogram is created. The observations in the input data set are cross-classified according to the values (levels) of these variables. A separate histogram is created for the process variable values in each cell of the cross-classification, and these component histograms are arranged in a matrix to form the comparative histogram. The levels of *variable1* are used to label the rows of the matrix, and the levels of *variable2* are used to label the columns of the matrix. Uniform horizontal and vertical axes are used to facilitate comparisons. For an example, see Output 3.2.1 on page 143.

Classification variables can be numeric or character, and the length of a character variable cannot exceed 16. Formatted values are used to determine the levels. You can specify whether missing values are to be treated as a level with the MISSING1 and MISSING2 options.

If a label is associated with a classification variable, the label is displayed on the comparative histogram. The variable label is displayed parallel to the column (or row) labels. For an example, see Figure 3.2 on page 119.

CLASSKEY=*'value'*

CLASSKEY=(*'value1' 'value2'*)

specifies the *key cell* in a comparative histogram requested with the CLASS= option. The bin size and midpoints are first determined for the key cell, and then the midpoint list is extended to accommodate the data ranges for the remaining cells. Thus, the choice of the key cell determines the uniform horizontal axis used for all cells.

If you specify CLASS=*variable*, you can specify CLASSKEY=*'value'* to identify the key cell as the level for which *variable* is equal to *value*. The *value* can have up to 16 characters, and you must specify a formatted *value*. By default, the levels are sorted in the order determined by the ORDER1= option, and the key cell is the level that occurs first in this order. The cells are displayed in this order from top to bottom (or left to right), and, consequently, the key cell is displayed at the top or at the left. If you specify a different key cell with the CLASSKEY= option, this cell is displayed at the top or at the left unless you also specify the NOKEYMOVE option.

If you specify CLASS=(*variable1 variable2*), you can specify CLASSKEY=(*'value1' 'value2'*) to identify the key cell as the level for which *variable1* is equal to *value1* and *variable2* is equal to *value2*. Here, *value1* and *value2* must be formatted values, and they must be enclosed in quotes. For an example of the CLASSKEY= option with a two-way comparative histogram, see Output 3.2.1 on page 143. By default, the levels of *variable1* are sorted in the order determined by the ORDER1= option, and within each of these levels, the levels of *variable2* are sorted in the order determined by the ORDER2= option. The default key cell is the combination of levels of *variable1* and *variable2* that occurs first in this order. The cells are displayed in order of *variable1* from top to bottom and in order of *variable2* from left to right. Consequently, the default key cell is displayed in the upper left corner. If you specify a different key cell with the CLASSKEY= option, this cell is displayed in the upper left corner unless you also specify the NOKEYMOVE option.

CLASSSPEC=*SAS-data-set*

CLASSSPECs=*SAS-data-set*

Part 1. The CAPABILITY Procedure

specifies a data set that provides distinct specification limits for each cell, as well as a color, legend, and label for the corresponding tile. The following table lists the variables that are read from a CLASSSPEC= data set:

Variable Name	Description
BY variables	subsets the data set
Classification variables	specifies the structure of the comparative histogram
VAR	specifies name of process variable (must be character variable of length 8)
LSL	specifies lower specification limit for tile
TARGET	specifies target value for tile
USL	specifies upper specification limit for tile
CTILE	specifies background color for tiles (must be character variable of length 8)
TILELG	specifies text displayed in color tile legend at bottom of comparative histogram (character variable of length not greater than 16)
TILELB	specifies text displayed in corner of each tile (character variable of length not greater than 16)

If you specify a CLASSSPEC= data set, you cannot use the SPEC statement or a SPEC= data set. If you use a BY statement, the CLASSSPEC= data set must contain one observation for each unique combination of process and classification variables within each BY group. See Example 3.1 on page 140 for an example of a CLASSSPEC= data set.

Also note that

- you can suppress the background color for a tile by assigning the value **EMPTY** or a blank value to the variable _CTILE_
- you can use the NLEGENDPOS= option to specify the corner of the tile in which the _TILELB_ label is displayed. You can frame the label with the CFRAMENLEG= option.
- you cannot use the variable _TILELG_ unless you specify the variable _CTILE_
- the variable _TILELB_ takes precedence over the NLEGEND option

COLOR=*color*

specifies the color of the normal density curve or the kernel density estimate curve. Enclose the COLOR= option in parentheses after the NORMAL option or the KERNEL option. See Output 3.1.1 on page 141 for an example.

CPROP=*color*

specifies the color for a horizontal bar whose length (relative to the width of the tile) indicates the proportion of the total frequency that is represented by the corresponding

cell. For an example, see Figure 3.3 on page 120. Empty bars are displayed if you specify CPROP=EMPTY. By default, bars are not displayed.

CTEXT=*color*

CT=*color*

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

CVREF=*color*

specifies the color for lines requested with the VREF= option. The default is the first color in the device color list.

DESCRIPTION='*string*'

DES='*string*'

specifies a description, up to 40 characters, that appears in the PROC GREPLAY master menu. The default is the variable name.

FILL

fills areas under a fitted density curve with colors and patterns. Enclose the FILL option in parentheses after the keyword NORMAL or KERNEL. Depending on the area to be filled (outside or between the specification limits), you can specify the color and pattern with options in the SPEC statement and the COMPHISTOGRAM statement, as summarized in the following table:

Area Under Curve	Statement	Option
between specification limits	COMPHIST	CFILL= <i>color</i>
	COMPHIST	PFILL= <i>pattern</i>
left of lower specification limit	SPEC	CLEFT= <i>color</i>
	SPEC	PLEFT= <i>pattern</i>
right of upper specification limit	SPEC	CRIGHT= <i>color</i>
	SPEC	PRIGHT= <i>pattern</i>

If you do not display specification limits, you can use the CFILL= and PFILL= options to specify the color and pattern for the entire area under the curve. Solid fills are used by default if patterns are not specified. You can specify the FILL option with only one fitted curve. For an example, see Output 3.1.1 on page 141. Refer to *SAS/GRAPH Software: Reference* for a list of available patterns and colors. If you do not specify the FILL option but you do specify the options in the preceding table, the colors and patterns are applied to the corresponding areas under the histogram.

FONT=*font*

specifies a software font for text used outside the framed areas of a comparative histogram (labels for axes, tick marks, and so forth). This font takes precedence over the FTEXT= font specified in a GOPTIONS statement. Refer to *SAS/GRAPH Software: Reference* for a list of fonts.

GRID

adds a grid to the comparative histogram. Grid lines are horizontal lines positioned at major tick marks on the vertical axis.

HEIGHT=*value*

specifies the height in percent screen units of text for axis labels, tick mark labels, and legends. The HEIGHT= option takes precedence over the HTEXT= option in the GOPTIONS statement.

HOFFSET=*value*

specifies the offset in percent screen units at both ends of the horizontal axis. Specify HOFFSET=0 to eliminate the default offset.

HREF=*value-list*

draws reference lines perpendicular to the horizontal axis at the values specified. For an illustration, see Output 4.1.1 on page 194.

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

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

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

specifies labels for the lines requested with the HREF= option. The number of labels must equal the number of lines. Enclose the labels in quotes. Labels can be up to 16 characters. For an illustration, see Output 4.1.1 on page 194.

HREFLABPOS=*n*

specifies the vertical position of HREFLABELS= labels as follows: 1 positions the labels along the top of the histogram; 2 staggers the labels from top to bottom; 3 positions the labels along the bottom. The default is 1.

INFONT=*font*

specifies a software font for text used inside the framed areas of the comparative histogram (such as sample size legends). The INFONT= option takes precedence over the FTEXT= option in the GOPTIONS statement. Refer to *SAS/GRAPH Software: Reference* for a list of fonts.

INHEIGHT=*value*

specifies the height in percent screen units of text used inside the framed areas of the comparative histogram (such as sample size legends). The default height is the height you specify with the HEIGHT= option. If you do not specify the HEIGHT= option, the default height is the height you specify with the HTEXT= option in the GOPTIONS statement.

INTERTILE=*value*

specifies the distance in horizontal percent screen units between tiles. For an example, see Figure 3.3 on page 120. By default, the tiles are contiguous.

K=NORMAL | TRIANGULAR | QUADRATIC

specifies the type of kernel (normal, triangular, or quadratic) used to compute kernel density estimates requested with the KERNEL option. Enclose the K= option in parentheses after the keyword KERNEL. You can specify a single type or a list of types. If you specify more estimates than types, the last kernel type in the list is used for the remaining estimates. By default, a normal kernel is used.

KERNEL<(*kernel-options*)>

requests a kernel density estimate for each cell of the comparative histogram. You can specify the *kernel-options* described in the following table:

FILL	specifies that the area under the curve is to be filled
COLOR=	specifies the color of the curve
L=	specifies the line style for the curve
W=	specifies the width of the curve
K=	specifies the type of kernel
C=	specifies the smoothing parameter

See Output 3.1.1 on page 141 for an example. By default, the estimate is based on the AMISE method. For more information, see “Kernel Density Estimates” on page 179.

L=*linetype*

specifies the line type for a normal or kernel density estimate curve. Enclose the L= option in parentheses after the NORMAL option or the KERNEL option. If you use the L= option with the KERNEL option, you can specify a single line type or a list of line types. Refer to *SAS/GRAPH Software: Reference* for a list of available line types. The default is 1, which produces a solid line.

LGRID=*n*

specifies the line type for the grid requested with the GRID option. If you use the LGRID= option, you do not need to specify the GRID option. The default is 1, which produces a solid line.

LHREF=*n***LH=***n*

specifies the line type for lines requested with the HREF= option. The default is 2, which produces a dashed line.

LVREF=*n***LV=***n*

specifies the line type for lines requested with the VREF= option. The default is 2, which produces a dashed line.

MAXNBIN=*n*

specifies the maximum number of bins to be displayed. This option is useful in situations where the scales or ranges of the data distributions differ greatly from cell to cell. By default, the bin size and midpoints are determined for the key cell, and then the midpoint list is extended to accommodate the data ranges for the remaining cells. However, if the cell scales differ considerably, the resulting number of bins may be so great that each cell histogram is scaled into a narrow region. By limiting the number of bins with the MAXNBIN= option, you can narrow the window about the data distribution in the key cell. Note that the MAXNBIN= option provides an alternative to the MAXSIGMAS= option.

MAXSIGMAS=*value*

limits the number of bins to be displayed to a range of *value* standard deviations (of the data in the key cell) above and below the mean of the data in the key cell. This option is useful in situations where the scales or ranges of the data distributions differ greatly from cell to cell. By default, the bin size and midpoints are determined for the key cell, and then the midpoint list is extended to accommodate the data ranges for the remaining cells. If the cell scales differ considerably, however, the resulting

number of bins may be so great that each cell histogram is scaled into a narrow region. By limiting the number of bins with the MAXSIGMAS= option, you narrow the window about the data distribution in the key cell. Note that the MAXSIGMAS= option provides an alternative to the MAXNBIN= option.

MIDPOINTS=*value-list* | KEY | UNIFORM

specifies how midpoints are determined for the bins in the comparative histogram. The method you specify is used for all process variables analyzed with the COMPHISTOGRAM statement.

If you specify MIDPOINTS=*value-list*, the *values* must be listed in increasing order and must be evenly spaced. The difference between consecutive midpoints is used as the width of the histogram bars. If the range of the *values* does not cover the range of the data as well as any specification limits (LSL and USL) that are given, the list is extended in either direction as necessary. See Example 3.1 on page 140 for an illustration.

If you specify MIDPOINTS=KEY, the procedure first determines the midpoints for the data in the key cell. The initial number of midpoints is based on the number of observations in the key cell using the method of Terrell and Scott (1985). The midpoint list for the key cell is then extended in either direction as necessary until it spans the data in the remaining cells.

If you specify MIDPOINTS=UNIFORM, the procedure determines the midpoints using all the observations as if there were no cells. In other words, the number of midpoints is based on the total sample size using the method of Terrell and Scott (1985).

By default, MIDPOINTS=KEY. However, if the key cell contains no observations, the default is MIDPOINTS=UNIFORM.

MISSING1

specifies that missing values of the first CLASS= variable are to be treated as a level of the CLASS= variable. If the first CLASS= variable is a character variable, a missing value is defined as a blank internal (unformatted) value. If the process variable is numeric, a missing value is defined as any of the SAS System missing values. If you do not specify MISSING1, observations for which the first CLASS= variable is missing are excluded from the analysis.

MISSING2

specifies that missing values of the second CLASS= variable are to be treated as a level of the CLASS= variable. If the second CLASS= variable is a character variable, a missing value is defined as a blank internal (unformatted) value. If the process variable is numeric, a missing value is defined as any of the SAS System missing values. If you do not specify MISSING2, observations for which the second CLASS= variable is missing are excluded from the analysis.

MU=*value*

specifies the parameter μ for the normal density curves requested with the NORMAL option. Enclose the MU= option in parentheses after the NORMAL option. The default value is the sample mean of the observations in the cell.

NAME='string'

specifies a name for the plot, up to eight characters, that appears in the PROC GREPLAY master menu. The default is 'CAPABILI'.

NCOLS=*n***NCOL=*n***

specifies the number of columns in a comparative histogram. You can use the NCOLS= option with the NROWS= option if you specify two CLASS= variables. See Output 3.2.1 on page 143 for an example of a two-way comparative histogram using the NCOLS= option. By default, NCOLS=1 (and NROWS=2) if you specify only one CLASS= variable, and NCOLS=2 (and NROWS=2) if you specify two CLASS= variables.

NLEGEND<='label'>

specifies the form of a legend that is displayed inside each tile and indicates the sample size of the cell. The following two forms are available:

- If you specify the NLEGEND option, the form is $N = n$ where n is the cell sample size.
- If you specify the NLEGEND='label' option, the form is $label = n$ where n is the cell sample size. The label can be up to 16 characters and must be enclosed in quotes. For instance, you might specify NLEGEND='Number of Parts' to request a label of the form *Number of Parts = n*.

See Figure 3.2 on page 119 for an example. You can use the CFRAMENLEG= option to frame the sample size legend. The variable _TILELB_ in a CLASSSPEC= data set overrides the NLEGEND option. By default, no legend is displayed.

NLEGENDPOS=NW | NE

specifies the position of the legend requested with the NLEGEND option or the variable _TILELB_ in a CLASSSPEC= data set. If NLEGENDPOS=NW, the legend is displayed in the northwest corner of the tile; if NLEGENDPOS=NE, the legend is displayed in the northeast corner of the tile. See Figure 3.2 on page 119 for an illustration. The default is NE.

NOBARS

suppresses the display of the bars in a comparative histogram.

NOCHART

suppresses the creation of a comparative histogram. This is an alias for NOPLOT.

NOFRAME

suppresses the frame around each tile. The NOFRAME option cannot be specified with the CFRAME= option.

NOHLABEL

suppresses the label for the horizontal axis. This is useful for avoiding clutter.

NOKEYMOVE

suppresses the rearrangement of cells that occurs by default when you use the CLASSKEY= option to specify the key cell. For details, see the entry for the

Part 1. The CAPABILITY Procedure

CLASSKEY= option.

NOPLOT

suppresses the creation of a comparative histogram. This option is useful when you are using the COMPHISTOGRAM statement solely to create an output data set.

NORMAL<(normal-options)>

displays a normal density curve for each cell of the comparative histogram. The equation of the normal density curve is

$$p(x) = \frac{h \times 100\%}{\sigma \sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad \text{for } -\infty < x < \infty$$

where

μ = mean
 σ = standard deviation ($\sigma > 0$)
 h = width of histogram interval

If you specify values for μ and σ with the MU= and SIGMA= *normal-options*, the same curve is displayed for each cell. By default, a distinct curve is displayed for each cell based on the sample mean and standard deviation for that cell. For example, the following statements display a distinct curve for each level of the variable SUPPLIER:

```
proc capability noprint;  
  comphist width / class=supplier normal(color=red l=2);  
run;
```

The curves are drawn in red with a line style of 2 (a dashed line). See Figure 3.3 on page 120 for another illustration. Table 3.1 on page 123 lists options that can be specified in parentheses after the NORMAL option.

NOVLABEL

suppresses the label for the vertical axis.

NOVTICK

suppresses the tick marks and tick mark labels for the vertical axis. If you specify the NOVTICK option, the NOVLABEL option is assumed.

NROWS=*n*

NROW=*n*

specifies the number of rows in a comparative histogram. You can use the NROWS= option with the NCOLS= option if you specify two CLASS= variables. See Figure 3.2 on page 119 for a *one-way* comparative histogram using the NROWS= option, and see Output 3.2.1 on page 143 for a *two-way* comparative histogram using the NROWS= and NCOLS= options. The default is 2.

ORDER1=INTERNAL | FORMATTED | DATA | FREQ

specifies the display order for the values of the first CLASS= variable.

The levels of the first CLASS= variable are always constructed using the *formatted* values of the variable, and the formatted values are always used to label the rows (columns) of a comparative histogram. You can use the ORDER1= option to determine the order of the rows (columns) corresponding to these values, as follows:

- **If you specify ORDER1=INTERNAL**, the rows (columns) are displayed from top to bottom (left to right) in increasing order of the internal (unformatted) values of the first CLASS= variable. If there are two or more distinct internal values with the same formatted value, then the order is determined by the internal value that occurs first in the input data set.

For example, suppose that you specify a numeric CLASS= variable called DAY (with values 1, 2, and 3). Suppose also that a format (created with the FORMAT procedure) is associated with DAY and that the formatted values are as follows: 1 = 'Wednesday', 2 = 'Thursday', and 3 = 'Friday'. If you specify ORDER1=INTERNAL, the rows of the comparative histogram will appear in day-of-the-week order (*Wednesday, Thursday, Friday*) from top to bottom.

- **If you specify ORDER1=FORMATTED**, the rows (columns) are displayed from top to bottom (left to right) in increasing order of the formatted values of the first CLASS= variable. In the preceding illustration, if you specify ORDER1=FORMATTED, the rows will appear in alphabetical order (*Friday, Thursday, Wednesday*) from top to bottom.
- **If you specify ORDER1=DATA**, the rows (columns) are displayed from top to bottom (left to right) in the order in which the values of the first CLASS= variable first appear in the input data set.
- **If you specify ORDER1=FREQ**, the rows (columns) are displayed from top to bottom (left to right) in order of *decreasing* frequency count. If two or more classes have the same frequency count, the order is determined by the formatted values.

By default, ORDER1=INTERNAL.

ORDER2=INTERNAL | FORMATTED | DATA | FREQ

specifies the display order for the values of the second CLASS= variable.

The levels of the second CLASS= variable are always constructed using the *formatted* values of the variable, and the formatted values are always used to label the columns of a two-way comparative histogram. You can use the ORDER2= option to determine the order of the columns.

The layout of a two-way comparative histogram is determined by using the ORDER1= option to obtain the order of the rows from top to bottom (recall that ORDER1=INTERNAL by default). Then the ORDER2= option is applied to the observations corresponding to the first row to obtain the order of the columns from left to right. If any columns remain unordered (that is, the categories are *unbalanced*), the ORDER2= option is applied to the observations in the second row, and so on, until all the columns have been ordered.

The values of the ORDER2= option are interpreted as described for the ORDER1= option. By default, ORDER2=INTERNAL.

OUTHISTOGRAM=SAS-data-set

creates a SAS data set that saves the midpoints of the histogram intervals, the observed percent of observations in each interval, and (optionally) the percent of observations in each interval estimated from a fitted normal distribution.

PFILL=pattern

specifies a pattern used to fill the bars of the histograms (or the areas under a fitted curve if you also specify the FILL option). See the entries for the CFILL= and FILL options for additional details. Refer to *SAS/GRAPH Software: Reference* for a list of pattern values. By default, the bars and curve areas are not filled.

RTINCLUDE

includes the right endpoint of each histogram interval in that interval. The left endpoint is included by default.

SIGMA=value

specifies the parameter σ for normal density curves requested with the NORMAL option. Enclose the SIGMA= option in parentheses after the NORMAL option. The default value is the sample standard deviation of the observations in the cell.

TILELEGLABEL='label'

specifies a label displayed to the left of the legend that is created when you provide _CTILE_ and _TILELG_ variables in a CLASSSPEC= data set. The *label* can be up to 16 characters and must be enclosed in quotes. The default *label* is *Tiles:*.

TURNVLABEL

TURNVLABELS

specifies that the characters in the labels for the vertical axis are to be turned and strung out vertically. This happens by default when a hardware font is used.

VAXIS=value-list

specifies tick mark values for the vertical axis. The values must be equally spaced and in increasing order, and the first value must be zero. You must scale the values in the same units as the bars (see the VSCALE= option), and the last value must be greater than or equal to the height of the largest bar. See Output 3.2.1 on page 143 for an example.

VAXISLABEL='label'

specifies a label (up to 40 characters) for the vertical axis.

VOFFSET=value

specifies the offset in percent screen units at the upper end of the vertical axis.

VREF=value-list

draws reference lines perpendicular to the vertical axis at the values specified. For an illustration, see Output 2.2.1 on page 112.

VREFLABELS='label1'... 'labeln'

VREFLABEL='label1'... 'labeln'

VREFLAB='label1'... 'labeln'

specifies labels for the lines requested with the VREF= option. The number of labels must equal the number of lines. Enclose the labels in quotes. Labels can be up to 16 characters. For an illustration, see Output 2.2.1 on page 112.

VREFLABPOS=*n*

specifies the horizontal position of VREFLABELS= labels as follows: VREFLABPOS=1 positions the labels at the left of the tile, and VREFLABPOS=2 positions the labels at the right. The default is 1.

VSCALE=PERCENT | COUNT | PROPORTION

specifies the scale of the vertical axis. The value COUNT scales the data in units of the number of observations per data unit. The value PERCENT scales the data in units of percent of observations per data unit. The value PROPORTION scales the data in units of proportion of observations per data unit. The default is PERCENT.

W=*n*

specifies the width in pixels of the curve. Enclose the W= option in parentheses after the NORMAL option or the KERNEL option. The default is 1.

WAXIS=*n*

specifies the line thickness (in pixels) for the axes and frame. The default is 1.

Examples

This section provides advanced examples of comparative histograms.

Example 3.1. Adding Insets with Descriptive Statistics

Three similar machines are used to attach a part to an assembly. One hundred assemblies are sampled from the output of each machine, and a part position is measured in millimeters. The following statements save the measurements in a SAS data set named MACHINES:

```

data machines;
  input position @@;
  label position='Position in Millimeters';
  if (_n_ <= 100) then machine = 'Machine 1';
  else if (_n_ <= 200) then machine = 'Machine 2';
  else machine = 'Machine 3';
  cards;
-0.17 -0.19 -0.24 -0.24 -0.12
 0.07 -0.61  0.22  1.91 -0.08
-0.59  0.05 -0.38  0.82 -0.14
 0.32  0.12 -0.02  0.26  0.19
-0.07  0.13 -0.49  0.07  0.65
 0.94 -0.51 -0.61 -0.57 -0.51
  .   .   .   .   .
  .   .   .   .   .
  .   .   .   .   .
 0.58  0.46  0.58  0.92  0.70
 0.81  0.07  0.33  0.82  0.62
 0.48  0.41  0.78  0.58  0.43
 0.07  0.27  0.49  0.79  0.92
 0.79  0.66  0.22  0.71  0.53
 0.57  0.90  0.48  1.17  1.03
;

```

Distinct specification limits for the three machines are provided in a data set named SPECLIMS.

```

data speclims;
  input @1 machine $9. _lsl_ _usl_;
  _var_ = 'position';
  cards;
Machine 1 -0.5 0.5
Machine 2  0.0 1.0
Machine 3  0.0 1.0
;

```

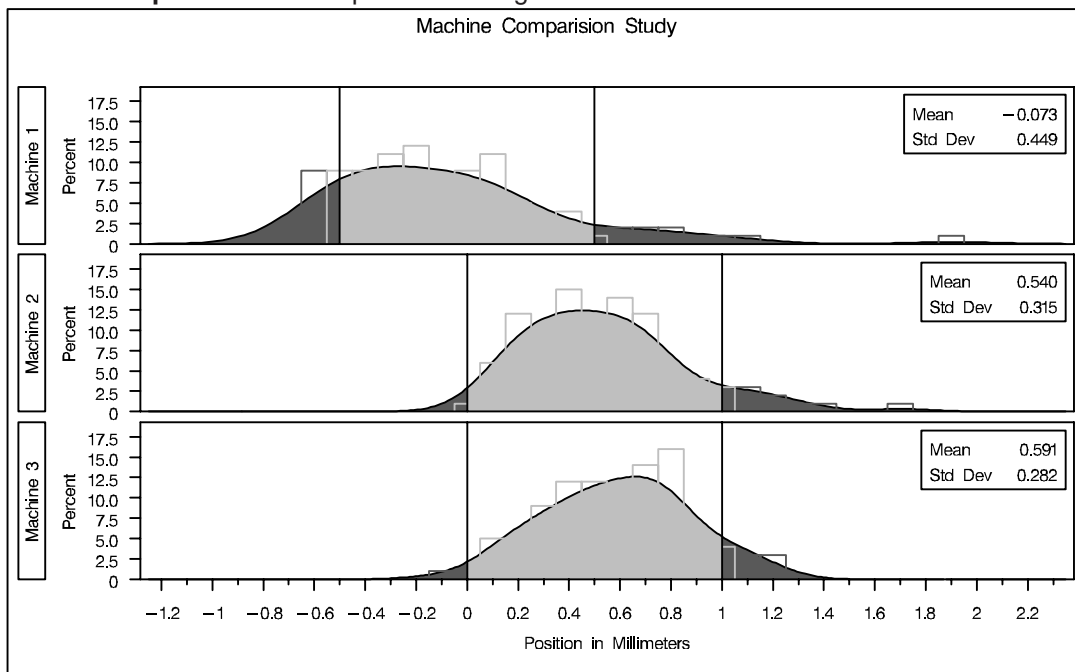
The following statements create a comparative histogram for the measurements in MACHINES that displays the specification limits in SPECLIMS. The display is shown in Output 3.1.1.

```

title 'Machine Comparision Study';
proc capability data=machines noprint;
  spec clsl=black c usl=black cleft=dagr cright=dagr;
  comphist position / class      = machine
                        nrows    = 3
                        intertile = 1
                        midpoints = -1.2 to 2.2 by 0.1
                        kernel (color=blue fill)
                        cfill     = ligr
                        classspecs = speclims;
  inset mean std='Std Dev' / pos = ne format = 6.3;
run;

```

Output 3.1.1. Comparative Histograms



The INSET statement is used to inset the sample mean and standard deviation for each machine in the corresponding tile. The MIDPOINTS= option specifies the midpoints of the histogram bins. Kernel density estimates are displayed using the KERNEL option. The curve areas outside the specification limits are filled using the CLEFT= and CRIGHT= options in the SPEC statement, and the area between the limits is filled using the CFILL= option in COMPHISTOGRAM statement.

Example 3.2. Creating a Two-Way Comparative Histogram

Two suppliers (A and B) provide disk drives for a computer manufacturer. The manufacturer measures the disk drive opening width to compare the process capabilities of the suppliers and determine whether there has been an improvement from 1992 to 1993.

See CAPCMH3
in the SAS/QC
Sample Library

Part 1. The CAPABILITY Procedure

The following statements save the measurements in a data set named DISK. There are two classification variables, SUPPLIER and YEAR, and a format is associated with YEAR.

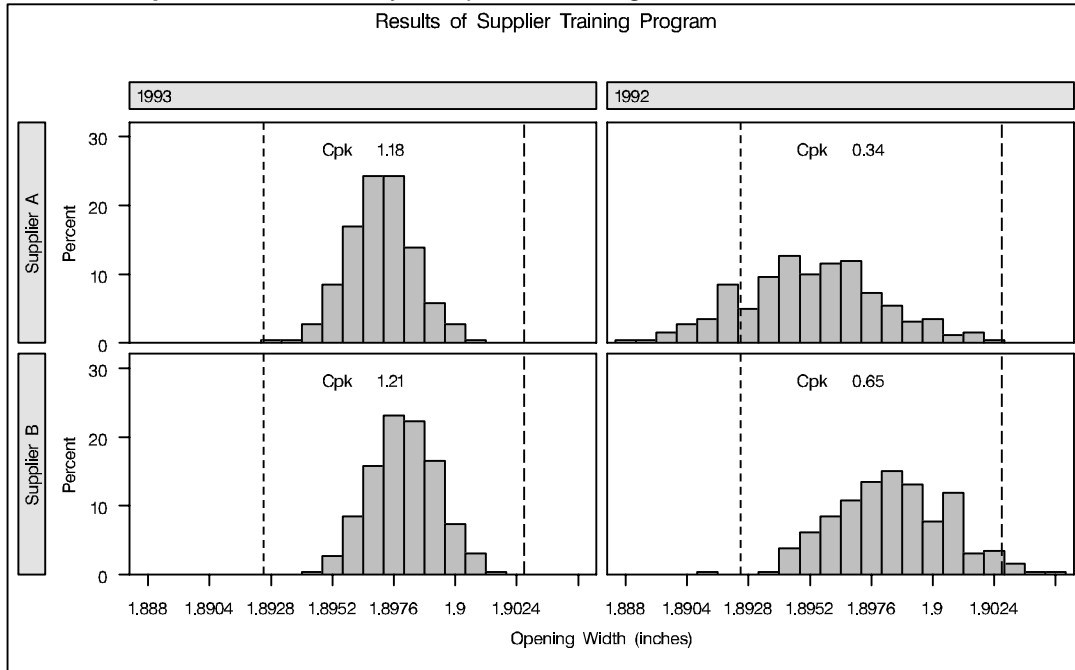
```
proc format ;
  value mytime 1 = '1992'
              2 = '1993' ;

data disk;
  input @1 supplier $10. year width;
  label width = 'Opening Width (inches)';
  format year mytime.;
  cards;
Supplier A   1   1.8932
Supplier A   1   1.8952
.           .   .
.           .   .
Supplier B   1   1.8980
Supplier B   1   1.8986
Supplier A   2   1.8978
Supplier A   2   1.8966
.           .   .
.           .   .
Supplier B   2   1.8967
Supplier B   2   1.8997
;
```

The following statements create the comparative histogram in Output 3.2.1:

```
title 'Results of Supplier Training Program';
proc capability data=disk noprint;
  specs  lsl = 1.8925  lls1 = 2
         usl = 1.9027  lus1 = 3;
  comphist width / class      = ( supplier year )
                             classkey = ('Supplier A' '1993')
                             intertile = 1.0
                             vaxis     = 0 10 20 30
                             ncols     = 2
                             nrows     = 2
                             cfill     = ligr
                             cframetop = blue
                             cframeside = blue ;
  inset cpk (4.2) / noframe pos = n;
run;
```

Output 3.2.1. Two-Way Comparative Histogram



The CLASSKEY= option specifies the key cell as the observations for which SUPPLIER is equal to **SUPPLIER A** and YEAR is equal to **2**. This cell determines the binning for the other cells, and (since the NOKEYMOVE option is not specified) the columns are interchanged so that this cell is displayed in the upper left corner. Note that if the CLASSKEY= option were not specified, the default key cell would be the observations for which SUPPLIER is equal to **SUPPLIER A** and YEAR is equal to **1**. If the CLASSKEY= option were not specified (or if the NOKEYMOVE option were specified), the column labeled *1992* would be displayed to the left of the column labeled *1993*. See the entry for the CLASSKEY= option on page 129 for details.

The VAXIS= option specifies the tick mark labels for the vertical axis, while NROWS=2 and NCOLS=2 specify a 2×2 arrangement for the tiles. The CFRAMESIDE= and CFRAMETOP= options specify fill colors for the row and column labels, and the CFILL= option specifies a fill color for the bars. The INSET statement is used to display the capability index C_{pk} for each cell. Output 3.2.1 provides evidence that both suppliers have reduced variability from 1992 to 1993.