SAS/QC® 14.3 User's Guide
The CAPABILITY Procedure

This document is an individual chapter from *SAS/QC® 14.3 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2017. *SAS/QC® 14.3 User's Guide*. Cary, NC: SAS Institute Inc.

SAS/QC® 14.3 User's Guide

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

September 2017

Chapter 6
The CAPABILITY Procedure

Contents

Introduction: CAPABILITY Procedure

A process capability analysis compares the distribution of output from an in-control process to its specification limits to determine the consistency with which the specifications can be met. The CAPABILITY procedure provides the following:

- process capability indices, such as C_p and C_{pk}
- descriptive statistics based on moments, including skewness and kurtosis. Other descriptive information provided includes quantiles or percentiles (such as the median), frequency tables, and details on extreme values.
- histograms. Optionally, these can be superimposed with specification limits, fitted probability density curves for various distributions, and kernel density estimates.
- cumulative distribution function plots (cdf plots). Optionally, these can be superimposed with specification limits and probability distribution curves for various distributions.
- quantile-quantile plots (Q-Q plots), probability plots, and probability-probability plots (P-P plots). These plots facilitate the comparison of a data distribution with various theoretical distributions. Optionally, Q-Q plots and probability plots can be superimposed with specification limits.
- comparative histograms, cdf plots, Q=Q plots, probability plots, and P-P plots. These are composite graphs that are composed of plots that correspond to the different levels of specified CLASS variables.
- goodness-of-fit tests for a variety of distributions including the normal. The assumption of normality is critical to the interpretation of capability indices.
- statistical intervals (prediction, tolerance, and confidence intervals) for a normal population
- the ability to produce plots either as traditional graphics, ODS Graphics output, or legacy line printer plots. Traditional graphics can be saved, replayed, and annotated.
- the ability to inset summary statistics and capability indices in graphical output

- the ability to analyze data sets with a frequency variable
- the ability to read specification limits from a data set
- the ability to create output data sets containing summary statistics, capability indices, histogram intervals, parameters of fitted curves, and statistical intervals

You can use the PROC CAPABILITY statement, together with the VAR and SPEC statements, to compute summary statistics and process capability indices. See "Getting Started: CAPABILITY Procedure" on page 197 for introductory examples. In addition, you can use the statements summarized in Table 6.1 to request plots and specialized analyses:

Table 6.1 Statements for Plots and Specialized Analyses

Statement	Result
CDFPLOT	cumulative distribution function plot
COMPHISTOGRAM	comparative histogram
HISTOGRAM	histogram
INSET	inset table on plot
INTERVALS	statistical intervals
OUTPUT	output data set with summary statistics and capability indices
PPPLOT	probability-probability plot
PROBPLOT	probability plot
QQPLOT	quantile-quantile plot

You have three alternatives for producing plots with the CAPABILITY procedure:

- ODS Graphics output is produced if ODS Graphics is enabled, for example by specifying the ODS GRAPHICS ON statement prior to the PROC statement.
- Otherwise, traditional graphics are produced by default if SAS/GRAPH is licensed.
- Legacy line printer charts are produced when you specify the LINEPRINTER option in the PROC statement.

See Chapter 4, "SAS/QC Graphics," for more information about producing these different kinds of graphs.

You can use the INSET statement with any of the plot statements to enhance the plot with an inset table of summary statistics. The INSET statement is not applicable when you produce line printer plots.

Learning about the CAPABILITY Procedure

To learn about the CAPABILITY procedure, first select the appropriate statement Table 6.1. Then refer to the corresponding "Getting Started" section for introductory examples:

- "Getting Started: CDFPLOT Statement" on page 256

To broaden your knowledge of the procedure, read "PROC CAPABILITY and General Statements" on page 195 which summarizes the syntax for the entire procedure and describes the PROC CAPABILITY statement, the VAR statement, the CLASS statement, and the SPEC statement. Subsequent chapters describe the statements listed in Table 6.1. In addition to introductory examples, each chapter provides syntax summaries, descriptions of options, computational details, and advanced examples. Although the chapters are self-contained, much of what you learn about one plot statement, including the syntax, is transferable to other plot statements.

PROC CAPABILITY and General Statements

Overview: CAPABILITY Procedure

This chapter describes several statements that are generally used with the CAPABILITY procedure:

- The PROC CAPABILITY statement is required to invoke the CAPABILITY procedure. You can use this statement by itself to compute summary statistics.
- The VAR statement, which is optional, specifies the variables in the input data set that are to be analyzed. These are called the analysis or *process* variables. By default, all of the numeric variables are analyzed.
- The CLASS statement, which is optional, specifies one or two variables that group the data into classification levels. A separate analysis is carried out for each combination of levels, and you can use the CLASS statement with plot statements (such as HISTOGRAM) to create comparative displays. [1]
- The SPEC statement, which is optional, provides specification limits for the variables that are to be analyzed. When you use a SPEC statement, the procedure computes process capability indices in addition to summary statistics. Furthermore, the specification limits are displayed in plots created with plot statements that are described in subsequent chapters.

[1] You can use the COMPHISTOGRAM statement to create comparative histograms without applying classification levels to the overall analysis.

You can use the PROC CAPABILITY statement to request a variety of statistics for summarizing the data distribution of each analysis variable:

- sample moments
- basic measures of location and variability
- confidence intervals for the mean, standard deviation, and variance
- tests for location
- tests for normality
- trimmed and Winsorized means
- robust estimates of scale
- quantiles and related confidence intervals
- extreme observations and extreme values
- frequency counts for observations
- missing values

You can use the PROC CAPABILITY and SPEC statements together to request a variety of statistics for process capability analysis:

- percents of measurements within and outside specification limits
- confidence intervals for the probabilities of exceeding the specification limits
- standard capability indices and related confidence intervals
- tests of normality in conjunction with capability indices
- specialized capability indices

In addition, you can use options in the PROC CAPABILITY statement to

- specify the input data set to be analyzed
- specify an input data set containing specification limits
- specify a graphics catalog for saving traditional graphics output
- specify rounding units for variable values
- specify the definition used to calculate percentiles
- specify the divisor used to calculate variances and standard deviations
- request legacy line printer plots and define special printing characters used for features

- suppress tables

You can use options in the SPEC statement to

- provide lower and upper specification limits and target values
- control the appearance of specification lines on plots
- control the appearance of the areas under a histogram outside the specification limits

Getting Started: CAPABILITY Procedure

This section introduces the PROC CAPABILITY, VAR, and SPEC statements with examples that illustrate the most commonly used options.

Computing Descriptive Statistics

NOTE: See *Computing Summary Stats and Capability Indices* in the SAS/QC Sample Library.

The fluid weights of 100 drink cans are measured in ounces. The filling process is assumed to be in statistical control. The measurements are saved in a SAS data set named Cans.

```
data Cans;
   label Weight = "Fluid Weight (ounces)";
   input Weight @@;
   datalines;
12.07  12.02  12.00  12.01  11.98  11.96  12.04  12.05  12.01  11.97
12.03  12.03  12.00  12.04  11.96  12.02  12.06  12.00  12.02  11.91
12.05  11.98  11.91  12.01  12.06  12.02  12.05  11.90  12.07  11.98
12.02  12.11  12.00  11.99  11.95  11.98  12.05  12.00  12.10  12.04
12.06  12.04  11.99  12.06  11.99  12.07  11.96  11.97  12.00  11.97
12.09  11.99  11.95  11.99  11.99  11.96  11.94  12.03  12.09  12.03
11.99  12.00  12.05  12.04  12.05  12.01  11.97  11.93  12.00  11.97
12.13  12.07  12.00  11.96  11.99  11.97  12.05  11.94  11.99  12.02
11.95  11.99  11.91  12.06  12.03  12.06  12.05  12.04  12.03  11.98
12.05  12.05  12.11  11.96  12.00  11.96  11.96  12.00  12.01  11.98
;
```

You can use the PROC CAPABILITY and VAR statements to compute summary statistics for the weights.

```
title 'Process Capability Analysis of Fluid Weight';
proc capability data=Cans normaltest;
   var Weight;
run;
```

The input data set is specified with the DATA= option. The NORMALTEST option requests tests for normality. The VAR statement specifies the variables to analyze. If you omit the VAR statement, all numeric variables in the input data set are analyzed.

The descriptive statistics for Weight are shown in Figure 6.1. For instance, the average weight (labeled *Mean*) is 12.0093. The Shapiro-Wilk test statistic labeled *W* is 0.987876, and the probability of a more extreme

value of W (labeled $Pr < W$) is 0.499. Compared to the usual cutoff value of 0.05, this probability (referred to as a p-value) indicates that the weights are normally distributed.

Figure 6.1 Descriptive Statistics

Process Capability Analysis of Fluid Weight

The CAPABILITY Procedure
Variable: Weight (Fluid Weight (ounces))

Moments			
N	100	**Sum Weights**	100
Mean	12.0093	**Sum Observations**	1200.93
Std Deviation	0.04695269	**Variance**	0.00220456
Skewness	0.05928405	**Kurtosis**	-0.1717404
Uncorrected SS	14422.5469	**Corrected SS**	0.218251
Coeff Variation	0.39096946	**Std Error Mean**	0.00469527

Basic Statistical Measures			
Location		**Variability**	
Mean	12.00930	**Std Deviation**	0.04695
Median	12.00000	**Variance**	0.00220
Mode	12.00000	**Range**	0.23000
		Interquartile Range	0.07000

Tests for Location: Mu0=0				
Test		**Statistic**	**p Value**	
Student's t	**t**	2557.745	**Pr > |t|**	<.0001
Sign	**M**	50	**Pr >= |M|**	<.0001
Signed Rank	**S**	2525	**Pr >= |S|**	<.0001

Tests for Normality				
Test		**Statistic**	**p Value**	
Shapiro-Wilk	**W**	0.987876	**Pr < W**	0.4991
Kolmogorov-Smirnov	**D**	0.088506	**Pr > D**	0.0522
Cramer-von Mises	**W-Sq**	0.079055	**Pr > W-Sq**	0.2179
Anderson-Darling	**A-Sq**	0.457672	**Pr > A-Sq**	>0.2500

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	12.130
99%	12.120
95%	12.090
90%	12.065
75% Q3	12.050
50% Median	12.000
25% Q1	11.980
10%	11.955
5%	11.935
1%	11.905
0% Min	11.900

Figure 6.1 *continued*

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
11.90	28	12.09	59
11.91	83	12.10	39
11.91	23	12.11	32
11.91	20	12.11	93
11.93	68	12.13	71

Computing Capability Indices

NOTE: See *Computing Summary Stats and Capability Indices* in the SAS/QC Sample Library.

This example is a continuation of the previous example and shows how you can provide specification limits with a SPEC statement to request capability indices in addition to descriptive statistics.

```
proc capability data=Cans normaltest freq;
   spec lsl=11.95 target=12 usl=12.05;
   var Weight;
run;
```

The options LSL=, TARGET=, and USL= specify the lower specification limit, target value, and upper specification limit for the weights. These statements produce the output shown in Figure 6.2 in addition to the output shown in Figure 6.1.

Figure 6.2 Capability Indices and Frequency Table

Process Capability Analysis of Fluid Weight

The CAPABILITY Procedure
Variable: Weight (Fluid Weight (ounces))

Specification Limits			
Limit		Percent	
Lower (LSL)	11.95000	% < LSL	7.00000
Target	12.00000	% Between	77.00000
Upper (USL)	12.05000	% > USL	16.00000

Process Capability Indices			
Index	Value	95% Confidence Limits	
Cp	0.354967	0.305565	0.404288
CPL	0.420991	0.332644	0.508117
CPU	0.288943	0.211699	0.365112
Cpk	0.288943	0.212210	0.365677
Cpm	0.348203	0.301472	0.398228

Figure 6.2 *continued*

Frequency Counts			
		Percents	
Value	Count	Cell	Cum
11.90	1	1.0	1.0
11.91	3	3.0	4.0
11.93	1	1.0	5.0
11.94	2	2.0	7.0
11.95	3	3.0	10.0
11.96	8	8.0	18.0
11.97	6	6.0	24.0
11.98	6	6.0	30.0
11.99	10	10.0	40.0
12.00	11	11.0	51.0
12.01	5	5.0	56.0
12.02	6	6.0	62.0
12.03	6	6.0	68.0
12.04	6	6.0	74.0
12.05	10	10.0	84.0
12.06	6	6.0	90.0
12.07	4	4.0	94.0
12.09	2	2.0	96.0
12.10	1	1.0	97.0
12.11	2	2.0	99.0
12.13	1	1.0	100.0

In Figure 6.2, the table labeled *Specification Limits* lists the specification limits and target value, together with the percents of observations outside and between the limits. The table labeled *Process Capability Indices* lists estimates for the standard process capability indices C_p, CPL, CPU, C_{pk}, and C_{pm}, along with 95% confidence limits. The index C_{pm} is not computed unless you specify a TARGET= value. See "Standard Capability Indices" on page 235 for formulas used to compute the indices.

If you specify more than one variable in the VAR statement, you can provide corresponding specification limits and target values by specifying lists of values for the LSL=, USL=, and TARGET= options. As an alternative to the SPEC statement, you can read specification limits and target values from a data set specified with the SPEC= option in the PROC CAPABILITY statement. This is illustrated in Example 6.1.

The FREQ option in the PROC CAPABILITY statement requests the table labeled *Frequency Counts* in Figure 6.2.

Syntax: CAPABILITY Procedure

The following are the primary statements that control the CAPABILITY procedure:

PROC CAPABILITY < *options* > **;**
BY *variables* **;**
CDFPLOT < *variables* > < / *options* > **;**
CLASS *variable-1* < *variable-2* >< / *options* > **;**
COMPHISTOGRAM < *variables* > / **CLASS= (***class-variables***)** < *options* > **;**
FREQ *variable* **;**
HISTOGRAM < *variables* > < / *options* > **;**
ID *variables* **;**
INSET *keyword-list* < / *options* > **;**
INTERVALS < *variables* > < / *options* > **;**
OUTPUT < **OUT=** *SAS-data-set* > < *keyword1=names . . . keywordk=names* > **;**
PPPLOT < *variables* > < / *options* > **;**
PROBPLOT < *variables* > < / *options* > **;**
QQPLOT < *variables* > < / *options* > **;**
SPEC < *options* > **;**
VAR *variables* **;**
WEIGHT *variable* **;**

The PROC CAPABILITY statement invokes the procedure. The VAR statement specifies the numeric variables to be analyzed, and it is required if the OUTPUT statement is used to save summary statistics and capability indices in an output data set. If you do not use the VAR statement, all numeric variables in the data set are analyzed. The SPEC statement provides specification limits.

The plot statements (CDFPLOT, COMPHISTOGRAM, HISTOGRAM, PPPLOT, PROBPLOT, and QQPLOT) create graphical displays, and the INSET statement enhances these displays by adding a table of summary statistics directly on the graph. The INTERVALS statement computes statistical intervals. You can specify one or more of each of the plot statements, the INSET statement, the INTERVALS statement, and the OUTPUT statement. If you use a VAR statement, the variables listed in a plot statement must be a subset of the variables listed in the VAR statement.

PROC CAPABILITY Statement

The syntax for the PROC CAPABILITY statement is as follows:

PROC CAPABILITY < *options* > **;**

The following section lists all *options*. See the section "Dictionary of Options" on page 204 for detailed information.

Summary of Options

Table 6.2 lists all the PROC CAPABILITY *options* by function.

Table 6.2 PROC CAPABILITY Statement Options

Option	Description
Input Data Set Options	
ANNOTATE=	specifies input data set containing annotation information
DATA=	specifies input data set
EXCLNPWGT	specifies that non-positive weights are to be excluded
NOBYSPECS	specifies that specification limits in SPEC= data set are to be applied to all BY groups
SPEC=	specifies input data set with specification limits
Plotting and Graphics Options	
FORMCHAR(*index*)=	defines characters used for features on legacy line printer plots
GOUT=	specifies catalog for saving traditional graphics output
LINEPRINTER	requests legacy line printer plots
Computational Options	
FORCEQN	forces calculation of the robust estimator of scale Q_n
FORCESN	forces calculation of the robust estimator of scale S_n
PCTLDEF=	specifies definition used to calculate percentiles
ROUND=	specifies units used to round variable values
VARDEF=	specifies divisor used to calculate variances and standard deviations
Data Summary Options	
ALL	requests all tables
FREQ	requests frequency table
MODES	requests table of modes
NEXTROBS=	requests table of n lowest, n highest observations
NEXTRVAL=	requests table of n lowest, n highest values
Output Options	
NOPRINT	suppresses printed output
OUTTABLE=	creates an output data set containing univariate statistics and capability indices in tabular form
Hypothesis Testing Options	
MU0=	specifies mean for null hypothesis in tests for location
LOCCOUNT	requests table of counts used in sign test and signed rank test
NORMALTEST	performs tests for normality
Robust Estimation Options	
ROBUSTSCALE	requests table of robust measures of scale
TRIMMED=(*trimmed-options*)	requests table of trimmed means
WINSORIZED=(*Winsorized-options*)	requests table of Winsorized means
TRIMMED-Options	
ALPHA=	specifies confidence level
TYPE=	specifies type of confidence limit

Table 6.2 *continued*

Option	Description
WINSORIZED-Options	
ALPHA=	specifies confidence level
TYPE=	specifies type of confidence limit
Capability Index Options	
CPMA=	(obsolete) specifies a for Cpm(a)
CHECKINDICES	requests test of normality in conjunction with standard indices
SPECIALINDICES	requests table of specialized indices including Boyles' C_{pm}, C_{jkp}, C_{pmk}, $C_{pm}(a)$, and Wright's C_s
CHECKINDICES-Options	
ALPHA=	specifies cutoff probability for p-values for test for normality used in conjunction with process capability indices
TEST=	specifies test for normality (Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling, Cramér-von Mises, or no test)
Confidence Limit Options	
ALPHA=	specifies level for all confidence limits
CIBASIC	requests confidence limits for the mean, standard deviation, variance
CIINDICES	specifies level and type of confidence limits for capability indices
CIPCTLDF	requests distribution-free confidence limits for percentiles
CIPCTLNORMAL	requests confidence limits for percentiles assuming normality
CIPROBEX	requests confidence limits for the probability of exceeding specifications
CIBASIC-Options	
ALPHA=	specifies confidence level
TYPE=	specifies type of confidence limit
CIINDICES-Options	
ALPHA=	specifies confidence level
TYPE=	specifies type of confidence limit
CIPCTLDF-Options	
ALPHA=	specifies confidence level
TYPE=	specifies type of confidence limit
CIPCTLNORMAL-Options	
ALPHA=	specifies confidence level
TYPE=	specifies type of confidence limit
CIPROBEX-Options	
ALPHA=	specifies confidence level
TYPE=	specifies type of confidence limit

Dictionary of Options

The following entries provide detailed descriptions of the *options* in the PROC CAPABILITY statement.

ALL

requests all of the tables generated by the FREQ, MODES, NEXTRVAL=5, CIBASIC, CIPCTLDF, and CIPCTLNORMAL options. If a WEIGHT statement is not used, the ALL option also requests the tables generated by the LOCCOUNT, NORMALTEST, ROBUSTSCALE, TRIMMED=.25, and WINSORIZED=.25 options. PROC CAPABILITY uses any values that you specify with the ALPHA=, MUO=, NEXTRVAL=, CIBASIC, CIPCTLDF, CIPCTLNORMAL, TRIMMED=, or WINSORIZED= options in conjunction with the ALL option.

ALPHA=*value*

specifies the default confidence level for all confidence limits computed by the CAPABILITY procedure. The coverage percent for the confidence limits is $(1 - value)100$. For example, ALPHA=0.10 results in 90% confidence limits. The default *value* is 0.05.

Note that specialized ALPHA= options are available for a number of confidence interval options. For example, you can specify CIBASIC(ALPHA=0.10) to request a table of *Basic Confidence Limits* at the 90% level. The default values of these options default to the value of the general ALPHA= option.

ANNOTATE=*SAS-data-set*

ANNO=*SAS-data-set*

specifies an input data set containing annotate variables as described in SAS/GRAPH documentation. You can use this data set to add features to traditional graphics. Use this data set only when creating traditional graphics; it is ignored when the LINEPRINTER option is specified and when ODS Graphics is in effect. Features provided in this data set are added to every plot produced in the current run of the procedure.

CHECKINDICES< (TEST = SW | KS | AD | CVM | NONE > < ALPHA=*value* **>)**

specifies the test of normality used in conjunction with process capability indices that are displayed in the *Process Capability Indices* table. If the p-value for the test is less than the cutoff probability value specified with the ALPHA= option, a warning is added to the table, as illustrated in Figure 6.3. See "Tests for Normality" on page 227 for details concerning the test.

```
proc capability data=Process;
var p2;
specs lsl=10
      usl=275;
run;
```

Figure 6.3 Warning Message Printed with Capability Indices

Process Capability Analysis of Fluid Weight

The CAPABILITY Procedure
Variable: P2

Process Capability Indices			
Index	**Value**	**95% Confidence Limits**	
Cp	0.541072	0.388938	0.692946
CPL	0.642426	0.417087	0.862984
CPU	0.439718	0.257339	0.617184
Cpk	0.439718	0.259310	0.620126

Warning: Normality is rejected for alpha = 0.05 using the Shapiro-Wilk test

ALPHA=*value*
specifies the cutoff probability for p-values for a test for normality used in conjunction with process capability indices. The *value* must be between zero and 0.5. The default value is 0.05.

TEST = SW | KS | AD | CVM | NONE
specifies the test of normality used in conjunction with process capability indices that are displayed in the *Process Capability Indices* table. The tests available are Shapiro-Wilk (SW), Kolmogorov-Smirnov (KS), Anderson-Darling (AD), and Cramér-von Mises (CVM). The default test is the Shapiro-Wilk test if the sample size is less than or equal to 2000 and the Kolmogorov-Smirnov test if the sample size is greater than 2000.

CIBASIC< (< TYPE=*keyword* **> < ALPHA=***value* **>) >**
requests confidence limits for the mean, standard deviation, and variance based on the assumption that the data are normally distributed. With large sample sizes, this assumption is not required for confidence limits for the mean.

ALPHA=*value*
specifies the confidence level. The coverage percent for the confidence limits is $(1 - value)100$. For example, ALPHA=0.10 requests 90% confidence limits. The default value is 0.05.

TYPE=*keyword*
specifies the type of confidence limit, where *keyword* is LOWER, UPPER, or TWOSIDED. The default value is TWOSIDED.

CIINDICES< (TYPE=*keyword* **>< ALPHA=***value* **>)**
specifies the type and level of the confidence limits for standard capability indices displayed in the table labeled *Process Capability Indices*.

ALPHA=*value*
specifies the confidence level. The coverage percent for the confidence limits is $(1 - value)100$. For example, ALPHA=0.10 requests 90% confidence limits. The default value is 0.05.

TYPE=*keyword*
specifies the type of confidence limit, where *keyword* is LOWER, UPPER, or TWOSIDED. The default value is TWOSIDED.

CIPCTLDF< (TYPE=*keyword* **>< ALPHA=***value* **>)**

CIQUANTDF< (TYPE=*keyword* **>< ALPHA=***value* **>)**

requests confidence limits for quantiles computed using a distribution-free method. In other words, no specific parametric distribution (such as the normal) is assumed for the data. Order statistics are used to compute the confidence limits as described in Section 5.2 of Hahn and Meeker (1991). This option is not available if you specify a WEIGHT statement.

ALPHA=*value*

specifies the confidence level. The coverage percent for the confidence limits is $(1 - value)100$. For example, ALPHA=0.10 requests 90% confidence limits. The default value is 0.05.

TYPE=*keyword*

specifies the type of confidence limit, where *keyword* is LOWER, UPPER, SYMMETRIC, or ASYMMETRIC. The default value is SYMMETRIC.

CIPCTLNORMAL< (TYPE=*keyword* **>< ALPHA=***value* **>)**

CIQUANTNORMAL< (TYPE=*keyword* **> < ALPHA=***value* **>)**

requests confidence limits for quantiles based on the assumption that the data are normally distributed. The computational method is described in Section 4.4.1 of Hahn and Meeker (1991) and uses the noncentral *t* distribution as given by Odeh and Owen (1980). This option is not available if you specify a WEIGHT statement.

ALPHA=*value*

specifies the confidence level. The coverage percent for the confidence limits is $(1 - value)100$. For example, ALPHA=0.10 requests 90% confidence limits. The default value is 0.05.

TYPE=*keyword*

specifies the type of confidence limit, where *keyword* is LOWER, UPPER, or TWOSIDED. The default value is TWOSIDED.

CIPROBEX< (TYPE=*keyword* **>< ALPHA=***value* **>)**

requests confidence limits for $Pr[X \leq \text{LSL}]$ and $Pr[X \geq \text{USL}]$, where X is the analysis variable, LSL is the lower specification limit, and USL is the upper specification limit. The computational method, which assumes that X is normally distributed, is described in Section 4.5 of Hahn and Meeker (1991) and uses the noncentral *t* distribution as given by Odeh and Owen (1980). This option is not available if you specify a WEIGHT statement.

ALPHA=*value*

specifies the confidence level. The coverage percent for the confidence limits is $(1 - value)100$. For example, ALPHA=0.10 requests 90% confidence limits. The default value is 0.05.

TYPE=*keyword*

specifies the type of confidence limit, where *keyword* is LOWER, UPPER, or TWOSIDED. The default value is TWOSIDED.

CPMA=*value*

specifies the *value* of the parameter a for the capability index $C_{pm}(a)$. This option has been superseded by the SPECIALINDICES(CPMA=) option.

DATA=*SAS-data-set*
specifies the input data set containing the observations to be analyzed. If the DATA= option is omitted, the procedure uses the most recently created SAS data set.

DEF=*index*
is an alias for the PCTLDEF= option. See the entry for the PCTLDEF= option.

EXCLNPWGT
excludes observations with non-positive weight values (zero or nonnegative) for the analysis. By default, PROC CAPABILITY treats observations with negative weights like those with zero weights and counts them in the total number of observations. This option is applicable only if you specify a WEIGHT statement.

FORCEQN
forces calculation of the robust estimate of scale Q_n. Because this calculation is very computationally intensive, by default Q_n is not computed for a variable that has more than 65,526 nonmissing observations. On some hosts, Q_n cannot be computed at all when there are more than 65,526 nonmissing observations.

FORCESN
forces calculation of the robust estimate of scale S_n. Because this calculation is computationally intensive, by default S_n is not computed for a variable that has more than 1 million nonmissing observations.

FORMCHAR(*index***)='***string***'**
defines characters used for features on legacy line printer plots, where *index* is a number ranging from 1 to 11, and *string* is a character or hexadecimal string. This option is ignored unless you specify the LINEPRINTER option in the PROC CAPABILITY statement.

The *index* identifies which features are controlled with the *string* characters, as discussed in the table that follows. If you specify the FORMCHAR= option omitting the *index*, the *string* controls all 11 features.

By default, the form character list specified with the SAS system option FORMCHAR= is used; otherwise, the default is FORMCHAR='|---|+|--'. If you print to a PC screen or your device supports the ASCII symbol set (1 or 2), the following is recommended:

```
formchar='B3,C4,DA,C2,BF,C3,C5,B4,C0,C1,D9'X
```

As an example, suppose you want to plot the data values of the empirical cumulative distribution function with asterisks (*). You can change the appropriate character by using the following:

```
formchar(2)='*'
```

Note that the FORMCHAR= option in the PROC CAPABILITY statement enables you to temporarily override the values of the SAS system option with the same name. The values of the SAS system option are not altered by using the FORMCHAR= option in PROC CAPABILITY statement.

The features associated with values of *index* are as follows:

Value of *index*	Description of Character	Chart Feature
1	vertical bar	frame, ecdf line, HREF= lines
2	horizontal bar	frame, ecdf line, VREF= lines
3	box character (upper left)	frame, ecdf line, histogram bars
4	box character (upper middle)	histogram bars, tick marks (horizontal axis)
5	box character (upper right)	frame, histogram bars
6	box character (middle left)	histogram bars
7	box character (middle middle)	not used
8	box character (middle right)	histogram bars, tick marks (vertical axis)
9	box character (lower left)	frame
10	box character (lower middle)	histogram bars
11	box character (lower right)	frame, ecdf line

FREQ

requests a frequency table in the printed output that contains the variable values, frequencies, percentages, and cumulative percentages. See Figure 6.2 for an example.

GOUT=*graphics-catalog*

specifies a graphics catalog in which to save traditional graphics output. This option is ignored unless you are producing traditional graphics.

LINEPRINTER

requests that legacy line printer plots be produced by the CDFPLOT, HISTOGRAM, PROBPLOT, PPPLOT, and QQPLOT statements. The CLASS and COMPHISTOGRAM statements cannot be used when the LINEPRINTER option is specified.

LOCCOUNT

requests a table with the number of observations greater than, not equal to, and less than the value of MUO=. PROC CAPABILITY uses these values to construct the sign test and signed rank test. This option is not available if you specify a WEIGHT statement.

MODES

MODE

requests a table of all possible modes. By default, when the data contains multiple modes, PROC CAPABILITY displays the lowest mode in the table of basic statistical measures. When all values are unique, PROC CAPABILITY does not produce a table of modes.

MU0=*value(s)*

LOCATION=*value(s)*

specifies the value of the mean or location parameter (μ_o) in the null hypothesis for the tests summarized in the table labeled *Tests for Location: Mu0=value*. If you specify a single value, PROC CAPABILITY tests the same null hypothesis for all analysis variables. If you specify multiple values, a VAR statement is required, and PROC CAPABILITY tests a different null hypothesis for each analysis variable by matching the VAR variables with the values in the corresponding order. The default value is 0.

NEXTROBS=*n*

specifies the number of extreme observations in the table labeled *Extreme Observations*. The table lists the *n* lowest observations and the *n* highest observations. The default value is 5. The value of *n* must be an integer between 0 and half the number of observations. You can specify NEXTROBS=0 to suppress the table.

NEXTRVAL=*n*

requests the table labeled *Extreme Values* and specifies the number of extreme values in the table. The table lists the *n* lowest unique values and the *n* highest unique values. The value of *n* must be an integer between 0 and half the maximum number of observations. By default, $n = 0$ and no table is displayed.

NOBYSPECS

specifies that specification limits in SPEC= data set be applied to all BY groups. If you use a BY statement and specify a SPEC= data set that does not contain the BY variables, you must specify the NOBYSPECS option.

NOPRINT

suppresses the tables of descriptive statistics and capability indices which are created by the PROC CAPABILITY statement. The NOPRINT option does not suppress the tables created by the INTERVALS or plot statements. You can use the NOPRINT options in these statements to suppress the creation of their tables.

NORMALTEST

NORMAL

requests a table of *Tests for Normality* for each of the analysis variables. The table provides test statistics and *p*-values for the Shapiro-Wilk test (provided the sample size is less than or equal to 2000), the Kolmogorov-Smirnov test, the Anderson-Darling test, and the Cramér-von Mises test. See "Tests for Normality" on page 227 for details. If specification limits are provided, the NORMALTEST option is assumed.

OUTTABLE=*SAS-data-set*

specifies an output data set that contains univariate statistics and capability indices arranged in tabular form. See "OUTTABLE= Data Set" on page 222 for details.

PCTLDEF=*index*

DEF=*index*

specifies one of five definitions used to calculate percentiles. The value of *index* can be 1, 2, 3, 4, or 5. See "Percentile Computations" on page 230 for details. By default, PCTLDEF=5.

ROBUSTSCALE

requests a table of robust measures of scale. These measures include the interquartile range, Gini's mean difference, the median absolute deviation about the median (*MAD*), and two statistics proposed by Rousseeuw and Croux (1993), Q_n, and S_n. This option is not available if you specify a WEIGHT statement.

ROUND=*value-list*

specifies units used to round variable values. The ROUND= option reduces the number of unique values for each variable and hence reduces the memory required for temporary storage. *Values* must be greater than 0 for rounding to occur.

If you use only one value, the procedure uses this unit for all variables. If you use a list of values, you must also use a VAR statement. The procedure then uses the roundoff values for variables in the order given in the VAR statement. For example, the following statements specify a roundoff value of 1 for Yieldstrength and a roundoff value of 0.5 for TENSTREN.

```
proc capability round=1 0.5;
   var Yieldstrength tenstren;
run;
```

When a variable value is midway between the two nearest rounded points, the value is rounded to the nearest even multiple of the roundoff value. For example, with a roundoff value of 1, the variable values of –2.5, –2.2, and –1.5 are rounded to –2; the values of –0.5, 0.2, and 0.5 are rounded to 0; and the values of 0.6, 1.2, and 1.4 are rounded to 1.

SPECIALINDICES

requests a table of specialized process capability indices. These indices include k, Boyles' modified C_{pm} (also denoted as $C_{pm}+$), C_{jkp}, $C_{pm}(a)$, $C_p(5.15)$, $C_{pk}(5.15)$, C_{pmk}, Wright's C_s, Boyles' S_{jkp}, C_{pp}, C''_{pp}, C_{pg}, C_{pq}, C_p^W, C_{pk}^W, C_{pm}^W, C_{pc}, and Vännmann's $C_p(u,v)$ and $C_p(v)$.

You can provide values for the parameters a for $C_{pm}(a)$, u and v for $C_p(u,v)$ and $C_p(v)$, and for the γ multiplier for C_s by specifying the following options in parentheses after the SPECIALINDICES option.

CPMA=*value*

specifies the *value* of the parameter a for the capability index $C_{pm}(a)$ described in Section 3.7 of Kotz and Johnson (1993). The *value* must be positive. The default *value* is 0.5. The existing CPMA= option in the PROC CAPABILITY statement is considered obsolete but still works.

CPU=*value*

specifies the *value* of the parameter u for Vännmann's capability index $C_p(u,v)$. The *value* must be greater than or equal to zero. The default *value* is zero.

CPV=*value*

specifies the *value* of the parameter v for Vännmann's capability indices $C_p(u,v)$ and $C_p(v)$. The *value* must be greater than or equal to zero. The default *value* is 4.

CSGAMMA=*value*

specifies the *value* of the γ multiplier suggested by Chen and Kotz (1996) for Wright's capability index C_s. The *value* must be greater than zero. The default *value* is 1.

SPEC=*SAS-data-set*

SPECS=*SAS-data-set*

specifies an input data set containing specification limits for each of the variables in the VAR statement. This option is an alternative to the SPEC statement, which also provides specification limits. See "SPEC= Data Set" on page 220 for details on SPEC= data sets, and Example 6.1 for an example. If you use both the SPEC= option and a SPEC statement, the SPEC= option is ignored.

TRIMMED=*values(s)* **< (TYPE=***keyword* **>< ALPHA=***value* **>)**

requests a table of trimmed means, where each *value* specifies the number or the proportion of trimmed observations. If the *value* is the number n of trimmed observations, n must be between 0 and half the number of nonmissing observations. If the *value* is a proportion p between 0 and 0.5, the number of observations trimmed is the smallest integer greater than or equal to np, where n is the number of observations. To obtain confidence limits for the mean and the student t-test, you must use the default value of VARDEF= which is DF. The TRIMMED= option is not available if you specify a WEIGHT statement.

ALPHA=*value*

specifies the confidence level. The coverage percent is $(1 - value)100$. For example, ALPHA=0.10 requests a 90% confidence limit. The default value is 0.05.

TYPE=*keyword*

specifies the type of confidence limit, where *keyword* is LOWER, UPPER, or TWOSIDED. The default value is TWOSIDED.

VARDEF=DF | N | WDF | WEIGHT | WGT

specifies the divisor used in calculating variances and standard deviations. The values and associated divisors are shown in the following table. By default, VARDEF=DF.

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

WINSORIZED=*values(s)* **< (< TYPE=***keyword* **> < ALPHA=***value* **>) >**

WINSOR=*values(s)* **< (< TYPE=***keyword* **> < ALPHA=***value* **>) >**

requests a table of winsorized means, where each *value* specifies the number or the proportion of winsorized observations. If the *value* is the number n of winsorized observations, n must be between 0 and half the number of nonmissing observations. If the *value* is a proportion p between 0 and 0.5, the number of observations winsorized is the smallest integer greater than or equal to np, where n is the number of observations. To obtain confidence limits for the mean and the student t-test, you must use the default value of VARDEF= which is DF. The WINSORIZED= option is not available if you specify a WEIGHT statement.

ALPHA=*value*

specifies the confidence level. The coverage percent is $(1 - value)100$. For example, ALPHA=0.10 results in a 90% confidence limit. The default value is 0.05.

TYPE=*keyword*

specifies the type of confidence limit, where *keyword* is LOWER, UPPER, or TWOSIDED. The default value is TWOSIDED.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC CAPABILITY to obtain separate analyses of observations in groups that are defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. If you specify more than one BY statement, only the last one specified is used.

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

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

For more information about BY-group processing, see the discussion in *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the discussion in the *SAS Visual Data Management and Utility Procedures Guide*.

CLASS Statement

CLASS *variable-1* < **(***v-options***)** > < *variable-2* < **(***v-options***)** > >
< **/ KEYLEVEL=** *value1* | **(***value1 value2***)** > **;**

The CLASS statement specifies one or two variables used to group the data into classification levels. Variables in a CLASS statement are referred to as *CLASS variables*. CLASS variables can be numeric or character. Class variables can have floating point values, but they typically have a few discrete values that define levels of the variable. You do not have to sort the data by CLASS variables. PROC CAPABILITY uses the formatted values of the CLASS variables to determine the classification levels.

NOTE: You cannot specify a COMPHISTOGRAM statement together with a CLASS statement.

You can specify the following *v-options* enclosed in parentheses after a CLASS variable:

MISSING
specifies that missing values for the CLASS variable are to be treated as valid classification levels. Special missing values that represent numeric values (‘.A’ through ‘.Z’ and ‘._’) are each considered as a separate value. If you omit MISSING, PROC CAPABILITY excludes the observations with a missing CLASS variable value from the analysis. Enclose this option in parentheses after the CLASS variable.

ORDER=DATA | FORMATTED | FREQ | INTERNAL
specifies the display order for the CLASS variable values. The default value is INTERNAL. You can specify the following values with the ORDER= option:

DATA
: orders values according to their order in the input data set. When you use a plot statement, PROC CAPABILITY displays the rows (columns) of the comparative plot from top to bottom (left to right) in the order that the CLASS variable values first appear in the input data set.

FORMATTED
: orders values by their ascending formatted values. This order might depend on your operating environment. When you use a plot statement, PROC CAPABILITY displays the rows (columns) of the comparative plot from top to bottom (left to right) in increasing order of the formatted CLASS variable values. For example, suppose a numeric CLASS variable DAY (with values 1, 2, and 3) has a user-defined format that assigns Wednesday to the value 1, Thursday to the value 2, and Friday to the value 3. The rows of the comparative plot will appear in alphabetical order (Friday, Thursday, Wednesday) from top to bottom.

  If there are two or more distinct internal values with the same formatted value, then PROC CAPABILITY determines the order by the internal value that occurs first in the input data set. For numeric variables without an explicit format, the levels are ordered by their internal values.

FREQ
: orders values by descending frequency count so that levels with the most observations are listed first. If two or more values have the same frequency count, PROC CAPABILITY uses the formatted values to determine the order.

  When you use a plot statement, PROC CAPABILITY displays the rows (columns) of the comparative plot from top to bottom (left to right) in order of decreasing frequency count for the CLASS variable values.

INTERNAL
: orders values by their unformatted values, which yields the same order as PROC SORT. This order may depend on your operating environment.

  When you use a plot statement, PROC CAPABILITY displays the rows (columns) of the comparative plot from top to bottom (left to right) in increasing order of the internal (unformatted) values of the CLASS variable. The first CLASS variable is used to label the rows of the comparative plots (top to bottom). The second CLASS variable is used to label the columns of the comparative plots (left to right). For example, suppose a numeric CLASS variable DAY (with values 1, 2, and 3) has a user-defined format that assigns Wednesday to the value 1, Thursday to the value 2, and Friday to the value 3. The rows of the comparative plot will appear in day-of-the-week order (Wednesday, Thursday, Friday) from top to bottom.

You can specify the following options after the slash (/) in the CLASS statement.

KEYLEVEL=*value* | (*value1 value2*)
: specifies the *key cells* in comparative plots. For each plot, PROC CAPABILITY first determines the horizontal axis scaling for the key cell, and then extends the axis using the established tick interval to accommodate the data ranges for the remaining cells, if necessary. Thus, the choice of the key cell determines the uniform horizontal axis that PROC CAPABILITY uses for all cells.

  If you specify only one CLASS variable and use a plot statement, KEYLEVEL=*value* identifies the key cell as the level for which the CLASS variable is equal to *value*. By default, PROC CAPABILITY sorts the levels in the order determined by the ORDER= option, and the key cell is the first occurrence of a level in this order. The cells display in order from top to bottom or left to right. Consequently,

the key cell appears at the top (or left). When you specify a different key cell with the KEYLEVEL= option, this cell appears at the top (or left).

If you specify two CLASS variables, use KEYLEVEL= (*value1 value2)* to identify the key cell as the level for which CLASS variable *n* is equal to *valuen*. By default, PROC CAPABILITY sorts the levels of the first CLASS variable in the order that is determined by its ORDER= option. Then, within each of these levels, it sorts the levels of the second CLASS variable in the order that is determined by its ORDER= option. The default key cell is the first occurrence of a combination of levels for the two variables in this order. The cells display in the order of the first CLASS variable from top to bottom and in the order of the second CLASS variable from left to right. Consequently, the default key cell appears at the upper left corner. When you specify a different key cell with the KEYLEVEL= option, this cell appears at the upper left corner.

The length of the KEYLEVEL= value cannot exceed 16 characters and you must specify a formatted value.

The KEYLEVEL= option has no effect unless you specify a plot statement.

NOKEYMOVE

specifies that the location of the key cell in a comparative plot be unchanged by the CLASS statement KEYLEVEL= option. By default, the key cell is positioned as the first cell in a comparative plot.

The NOKEYMOVE option has no effect unless you specify a plot statement.

FREQ Statement

FREQ *variable* **;**

The FREQ statement names a variable that provides frequencies for each observation in the input data set. If *n* is the value of the FREQ variable for a given observation, then that observation is used *n* times. If the value of the FREQ variable is missing or is less than one, the observation is not used in the analysis. If the value is not an integer, only the integer portion is used.

ID Statement

ID *variables* **;**

The ID statement specifies one or more variables to include in the table of extreme observations. The corresponding values of the ID variables appear beside the *n* largest and *n* smallest observations, where *n* is the value of the NEXTROBS= option.

SPEC Statement

The syntax for the SPEC statement is as follows:

SPEC < *options* > **;**

You can use at most one SPEC statement in the CAPABILITY procedure. When you provide specification limits and target values in a SPEC statement, the tabular output produced by the PROC CAPABILITY statement includes process capability indices as well as summary statistics. You can use the SPEC statement in conjunction with the CDFPLOT, COMPHISTOGRAM, HISTOGRAM, PROBPLOT, and QQPLOT statements to add specification limit and target lines to the plots produced with these statements.

options

control features of the specification limits and target values. Table 6.3 lists all options by function.

Summary of Options

Table 6.3 SPEC Statement Options

Option	Description
Lower Specification Limit Options	
CLEFT=	color used to fill area left of lower specification limit (histograms only)
CLSL=	color of lower specification limit line
LLSL=	line type of lower specification limit line
LSL=	lower specification limit values
LSLSYMBOL=	character used for lower specification limit line in line printer plots
PLEFT=	pattern type used to fill area left of lower specification limit (histograms only)
WLSL=	width of lower specification limit line
Target Options	
CTARGET=	color of target line
LTARGET=	line type of target line
TARGET=	target value
TARGETSYMBOL=	character used for target in line printer plots
WTARGET=	width of target line
Upper Specification Limit Options	
CRIGHT=	color used to fill area right of upper specification limit (histograms only)
CUSL=	color of upper specification limit line
LUSL=	line type of upper specification limit line
PRIGHT=	pattern type used to fill area right of upper specification limit (histograms only)
USL=	upper specification limit values
USLSYMBOL=	character used for upper specification limit in line printer plots
WUSL=	width of upper specification limit line

General Options

You can specify the following options whether you are producing ODS Graphics output or traditional graphics:

CLEFT=*color*

CLEFT

determines the *color* used to fill the area under a histogram to the left of the lower specification limit. You can specify the CLEFT option without an argument to fill this area with an appropriate color from the ODS style. If you are producing ODS Graphics output, an explicit color specification is ignored. This option is applicable only when the SPEC statement is used in conjunction with a HISTOGRAM or COMPHISTOGRAM statement. See Output 6.2.1 for an example. The CLEFT= option also applies to the area under a fitted curve;for an example, see Output 6.8.1.

CRIGHT=*color*

CRIGHT

determines the *color* used to fill the area under a histogram to the right of the upper specification limit. You can specify the CRIGHT option without an argument to fill this area with an appropriate color from the ODS style. If you are producing ODS Graphics output, an explicit color specification is ignored. This option is applicable only when the SPEC statement is used in conjunction with a HISTOGRAM or COMPHISTOGRAM statement. See Output 6.2.1 for an example. The CRIGHT= option also applies to the area under a fitted curve; for an example, see Output 6.8.1.

LSL=*value-list*

specifies the lower specification limits for the variables listed in the VAR statement, or for all numeric variables in the input data set if no VAR statement is used. If you specify only one lower limit, it is used for all of the variables; otherwise, the number of limits must match the number of variables. See the section "Computing Capability Indices" on page 199 for an example.

TARGET=*value-list*

specifies target values for the variables listed in the VAR statement, or for all numeric variables in the input data set if no VAR statement is used. If you specify only one target value, it is used for all of the variables; otherwise, the number of values must match the number of variables. See the section "Computing Capability Indices" on page 199 for an example.

USL=*value-list*

specifies the upper specification limits for the variables listed in the VAR statement, or for all numeric variables in the input data set if no VAR statement is used. If you specify only one upper limit, it is used for all of the variables; otherwise, the number of limits must match the number of variables. See the section "Computing Capability Indices" on page 199 for an example.

Options for Traditional Graphics

You can specify the following options if you are producing traditional graphics:

CLSL=*color*

specifies the color of the lower specification line displayed in plots created with the CDFPLOT, COMPHISTOGRAM, HISTOGRAM, PROBPLOT, and QQPLOT statements.

CTARGET=*color*

specifies the color of the target line displayed in plots created with the CDFPLOT, COMPHISTOGRAM, HISTOGRAM, PROBPLOT, and QQPLOT statements.

CUSL=*color*

specifies the color of the upper specification line displayed in plots created with the CDFPLOT, COMPHISTOGRAM, HISTOGRAM, PROBPLOT, and QQPLOT statements.

LLSL=*linetype*

specifies the line type for the lower specification line displayed in plots created with the CDFPLOT, COMPHISTOGRAM, HISTOGRAM, PROBPLOT, and QQPLOT statements. See Output 6.2.1 for an example. The default is 1, which produces a solid line.

LTARGET=*linetype*
specifies the line type for the target line in plots created with the CDFPLOT, COMPHISTOGRAM, HISTOGRAM, PROBPLOT, and QQPLOT statements. See Output 6.2.1 for an example. The default is 1, which produces a solid line.

LUSL=*linetype*
specifies the line type for the upper specification line displayed in plots created with the CDFPLOT, COMPHISTOGRAM, HISTOGRAM, PROBPLOT, and QQPLOT statements. See Output 6.2.1 for an example. The default is 1, which produces a solid line.

PLEFT=*pattern*
specifies the pattern used to fill the area under a histogram to the left of the lower specification limit. This option is applicable only when the SPEC statement is used in conjunction with a HISTOGRAM or COMPHISTOGRAM statement. For an example, see Output 6.2.1. The PLEFT= option also applies to the area under a fitted curve; for an example, see Output 6.8.1. The default pattern is a solid fill.

PRIGHT=*pattern*
specifies the pattern used to fill the area under a histogram to the right of the upper specification limit. This option is applicable only when the SPEC statement is used in conjunction with a HISTOGRAM or COMPHISTOGRAM statement. For an example, see Output 6.2.1. The PRIGHT= option also applies to the area under a fitted curve; for an example, see Output 6.8.1. The default pattern is a solid fill.

WLSL=*n*
specifies the width in pixels of the lower specification line in plots created with the CDFPLOT, COMPHISTOGRAM, HISTOGRAM, PROBPLOT, and QQPLOT statements. See Output 6.2.1 for an illustration. The default is 1.

WTARGET=*n*
specifies the width in pixels of the target line in plots created with the CDFPLOT, COMPHISTOGRAM, HISTOGRAM, PROBPLOT, and QQPLOT statements. See Output 6.2.1 for an illustration. The default is 1.

WUSL=*n*
specifies the width in pixels of the upper specification line in plots created with the CDFPLOT, COMPHISTOGRAM, HISTOGRAM, PROBPLOT, and QQPLOT statements. See Output 6.2.1 for an illustration. The default is 1.

Options for Legacy Line Printer Plots

You can specify the following options if you are producing legacy line printer plots:

LSLSYMBOL='*character*'
specifies the character used to display the lower specification line in line printer plots created with the CDFPLOT, HISTOGRAM, PROBPLOT, and QQPLOT statements. The default character is 'L'.

TARGETSYMBOL='*character*'
TARGETSYM='*character*'
specifies the character used to display the target line in line printer plots created with the CDFPLOT, HISTOGRAM, PROBPLOT, and QQPLOT statements. The default character is 'T'.

USLSYMBOL='*character*'
specifies the character used to display the upper specification line in line printer plots created with the CDFPLOT, HISTOGRAM, PROBPLOT, and QQPLOT statements. The default character is 'U'.

VAR Statement

VAR *variables* **;**

The VAR statement specifies the analysis variables and their order in the results. By default, if you omit the VAR statement, PROC CAPABILITY analyzes all numeric variables that are not listed in the other statements.

You must provide a VAR statement when you use an OUTPUT statement. To store the same statistic for several analysis variables in the OUT= data set, you specify a list of names in the OUTPUT statement. PROC CAPABILITY makes a one-to-one correspondence between the order of the analysis variables in the VAR statement and the list of names that follow a statistic keyword.

WEIGHT Statement

WEIGHT *variable* **;**

The WEIGHT statement names a variable that provides weights for each observation in the input data set. The CAPABILITY procedure uses the values w_i of the WEIGHT variable to modify the computation of a number of summary statistics by assuming that the variance of the ith value X_i of the analysis variable is equal to σ^2/w_i, where σ is an unknown parameter. This assumption is rarely applicable in process capability analysis, and the purpose of the WEIGHT statement is simply to make the CAPABILITY procedure consistent with other data summarization procedures, such as the UNIVARIATE procedure.

The values of the WEIGHT variable do not have to be integers and are typically positive. By default, observations with non-positive or missing values of the WEIGHT variable are handled as follows:

- If the value is zero, the observation is counted in the total number of observations.
- If the value is negative, it is converted to zero, and the observation is counted in the total number of observations.
- If the value is missing, the observation is excluded from the analysis.

To exclude observations that contain negative and zero weights from the analysis, specify the option EXCLNPWGT in the PROC statement. Note that most SAS/STAT procedures, such as PROC GLM, exclude negative and zero weights by default.

When you specify a WEIGHT variable, the procedure uses its values, w_i, to compute weighted versions of the statistics provided in the *Moments* table. For example, the procedure computes a weighted mean $\overline{X}_w$ and a weighted variance s_w^2 as $\overline{X}_w = \frac{\sum_i w_i x_i}{\sum_i w_i}$ and $s_w^2 = \frac{1}{d}\sum_i w_i (x_i - \overline{X}_w)^2$ where x_i is the ith variable value. The divisor d is controlled by the VARDEF= option in the PROC CAPABILITY statement.

When you use both the WEIGHT and SPEC statements, capability indices are computed using $\overline{X}_w$ and s_w in place of $\overline{X}$ and s. Again, note that weighted capability indices are seldom needed in practice.

When you specify a WEIGHT statement, the procedure also computes a weighted standard error and a weighted version of Student's t test. This test is the only test of location that is provided when weights are specified.

The WEIGHT statement does not affect the determination of the mode, extreme values, extreme observations, or the number of missing values of the analysis variables. However, the weights w_i are used to compute weighted percentiles.

The WEIGHT variable has no effect on the calculation of extreme values, and it has no effect on graphical displays produced with the plot statements.

Graphical Enhancement Statements

You can use TITLE, FOOTNOTE, and NOTE statements to enhance printed output. If you are creating traditional graphics, you can also use AXIS, LEGEND, PATTERN, and SYMBOL statements to enhance your plots. For details, see SAS/GRAPH documentation and the chapter for the plot statement that you are using.

Details: CAPABILITY Procedure

This section provides details on the following topics:

- input data sets specified with the DATA= option, the SPEC= option, and the ANNOTATE= option
- the output data set specified with the OUTTABLE= option
- descriptive statistics
- the tests for normality requested with the NORMALTEST option
- percentile definitions controlled using the PCTLDEF= option
- robust estimators
- computing the mode
- assumptions and terminology for capability indices
- standard capability indices
- specialized capability indices

Input Data Sets

DATA= Data Set

The DATA= data set contains a set of variables that represent measurements from a process. The CAPABILITY procedure must have a DATA= data set. If you do not specify one with the DATA= option in the PROC CAPABILITY statement, the procedure uses the last data set created.

SPEC= Data Set

The SPEC= option in the PROC CAPABILITY statement identifies a SPEC= data set, which contains specification limits. This option is an alternative to using the SPEC statement. If you use both the SPEC= option and a SPEC statement, the SPEC= option is ignored. The SPEC= option is especially useful when:

- the number of variables is large
- the same specification limits are referred to in more than one analysis
- a BY statement is used
- batch processing is used

The following variables are read from a SPEC= data set:

Variable	Description
LSL	lower specification limit
TARGET	target value
USL	upper specification limit
VAR	name of the variable

You may omit either _LSL_ or _USL_ but not both. _TARGET_ is optional. If the SPEC= data set contains both _LSL_ and _USL_, you can assign missing values to _LSL_ or _USL_ to indicate one-sided specifications. You can assign missing values to _TARGET_ when the variable does not use a target value. _LSL_, _USL_, and _TARGET_ must be numeric variables. _VAR_ must be a character variable.

You can include the following optional variables in a SPEC= data set to control the appearance of specification limits on charts:

Variable	Description
CLEFT	color used to fill area left of LSL (histograms only)
CLSL	color of LSL line
CRIGHT	color used to fill area right of USL (histograms only)
CTARGET	color of target line
CUSL	color of USL line
LLSL	line type of LSL line
LSLSYM	character used for LSL line in line printer plots
LTARGET	line type of target line
LUSL	line type of USL line
PLEFT	pattern type used to fill area left of LSL (histograms only)
PRIGHT	pattern type used to fill area right of USL (histograms only)
TARGETSYM	character used for target in line printer plots
USLSYM	character used for USL line in line printer plots
WLSL	width of LSL line
WTARGET	width of target line
WUSL	width of USL line

If you are using the HISTOGRAM statement to create "clickable" histograms in HTML, you can also provide the following variables in a SPEC= data set:

Variable	Description
LOURL	URL associated with area to left of lower specification limit
HIURL	URL associated with area to right of upper specification limit
URL	URL associated with area between specification limits

These are character variables whose values are Uniform Resource Locators (URLs) linked to areas on a histogram. When you view the ODS HTML output with a browser, you can click on an area, and the browser will bring up the page specified by the corresponding URL.

If you use a BY statement, the SPEC= data set must also contain the BY variables. The SPEC= data set must be sorted in the same order as the DATA= data set. Within a BY group, specification limits for each variable plotted are read from the first observation where _VAR_ matches the variable name.

See the section "Examples: CAPABILITY Procedure" on page 248 for an example of reading specification limits from a SPEC= data set.

ANNOTATE= Data Sets

In the CAPABILITY procedure, you can add features to traditional graphics plots by specifying ANNOTATE= data sets either in the PROC CAPABILITY statement or in individual plot statements. Depending on where you specify an ANNOTATE= data set, however, the information is used for all plots or only for plots produced by a given statement.

Information contained in the ANNOTATE= data set specified in the PROC CAPABILITY statement is used for all plots produced in a given PROC step; this is a "global" ANNOTATE= data set. By using this global data set, you can keep information common to all plots in one data set.

Information contained in the ANNOTATE= data set specified in a plot statement is used for plots produced by that statement; this is a "local" ANNOTATE= data set. By using this data set, you can add statement-specific features to plots. For example, you can add different features to plots produced by the HISTOGRAM and QQPLOT statements by specifying an ANNOTATE= data set in each plot statement.

In addition, you can specify an ANNOTATE= data set in the PROC CAPABILITY statement and in plot statements. This enables you to add some features to all plots (those given in the data set specified in the PROC statement) and also add statement-specific features to plots (those given in the data set specified in the plot statement).

For complete details on the structure and content of Annotate type data sets, see SAS/GRAPH documentation.

Output Data Set

OUTTABLE= Data Set

The OUTTABLE= data set saves univariate statistics and capability indices. The following variables can be saved:

Table 6.4 OUTTABLE= Data Set

Variable	Description
_CP_	Capability index C_p
CPLCL	Lower confidence limit for C_p
CPUCL	Upper confidence limit for C_p
CPK	Capability index C_{pk}
CPKLCL	Lower confidence limit for C_{pk}
CPKUCL	Upper confidence limit for C_{pk}
CPL	Capability index CPL
CPLLCL	Lower confidence limit for CPL
CPLUCL	Upper confidence limit for CPL
CPM	Capability index C_{pm}
CPMLCL	Lower confidence limit for C_{pm}
CPMUCL	Upper confidence limit for C_{pm}
CPU	Capability index CPU
CPULCL	Lower confidence limit for CPU
CPUUCL	Upper confidence limit for CPU
CSS	Corrected sum of squares
_CV_	Coefficient of variation
GEOMEAN	Geometric mean
GINI	Gini's mean difference
_K_	Capability index K
KURT	Kurtosis
LSL	Lower specification limit
MAD	Median absolute difference about the median
MAX	Maximum
MEAN	Mean
MEDIAN	Median
MIN	Minimum
MODE	Mode
MSIGN	Sign statistic
NMISS	Number of missing observations
NOBS	Number of nonmissing observations
_P1_	1st percentile
_P5_	5th percentile
_P10_	10th percentile
_P90_	90th percentile
_P95_	95th percentile
_P99_	99th percentile
PCTGTR	Percentage of observations greater than upper specification limit
PCTLSS	Percentage of observations less than lower specification limit

Table 6.4 *(continued)*

Variable	Description
PROBM	p-value of sign statistic
PROBN	p-value of test for normality
PROBS	p-value of signed rank test
PROBT	p-value of t statistic
_Q1_	25th percentile (lower quartile)
_Q3_	75th percentile (upper quartile)
_QN_	Q_n (see "Robust Estimates of Scale" on page 233)
QRANGE	Interquartile range (upper quartile minus lower quartile)
RANGE	Range
SGNRNK	Centered sign rank
SKEW	Skewness
_SN_	S_n (see "Robust Estimates of Scale" on page 233)
STD	Standard deviation
STDGINI	Gini's standard deviation
STDMAD	MAD standard deviation
STDMEAN	Standard error of the mean
STDQN	Q_n standard deviation
STDQRANGE	Interquartile range standard deviation
STDSN	S_n standard deviation
SUMWGT	Sum of the weights
SUM	Sum
TARGET	Target value
USL	Upper specification limit
USS	Uncorrected sum of squares
VARI	Variance
VAR	Variable name

NOTE: The variables _CP_, _CPLCL_, _CPUCL_, _CPK_, _CPKLCL_, _CPKUCL_, _CPL_, _CPLLCL_, _CPLUCL_, _CPM_, _CPMLCL_, _CPMUCL_, _CPU_, _CPULCL_, _CPUUCL_, _K_, _LSL_, _PCTGTR_, _PCTLSS_, _TARGET_, and _USL_ are included if you provide specification limits.

The OUTTABLE= data set and the OUT= data set[2] contain essentially the same information. However, the structure of the OUTTABLE= data set may be more appropriate when you are computing summary statistics or capability indices for more than one process variable in the same invocation of the CAPABILITY procedure. Each observation in the OUTTABLE= data set corresponds to a different process variable, and the variables in the data set correspond to summary statistics and indices.

NOTE: See *Tabulating Results for Multiple Variables* in the SAS/QC Sample Library.

For example, suppose you have ten process variables (P1-P10). The following statements create an OUTTABLE= data set named Table, which contains summary statistics and capability indices for each of these variables:

[2] See "OUTPUT Statement: CAPABILITY Procedure" on page 423 for details on the OUT= data set.

```
proc capability data=Process outtable=Table noprint;
   var P1-P10;
   specs lsl=5 10 65 35 35 5 25 25 60 15
         usl=175 275 300 450 550 200 275 425 500 525;
run;
```

The following statements create the table shown in Figure 6.4, which contains the mean, standard deviation, lower and upper specification limits, and capability index C_{pk} for each process variable:

```
proc print data=Table label noobs;
   var _VAR_ _MEAN_ _STD_ _LSL_ _USL_ _CPK_;
   label _VAR_='Process';
run;
```

Figure 6.4 Tabulating Results for Multiple Process Variables

Process Capability Analysis of Fluid Weight

Process	Mean	Standard Deviation	Lower Specification Limit	Upper Specification Limit	Capability Index CPK
P1	90.76	57.024	5	175	0.49242
P2	167.32	81.628	10	275	0.43972
P3	224.56	96.525	65	300	0.26052
P4	258.08	145.218	35	450	0.44053
P5	283.48	157.033	35	550	0.52745
P6	107.48	52.437	5	200	0.58814
P7	153.20	90.031	25	275	0.45096
P8	217.08	130.031	25	425	0.49239
P9	280.68	140.943	60	500	0.51870
P10	243.24	178.799	15	525	0.42551

Descriptive Statistics

This section provides computational details for the descriptive statistics which are computed with the PROC CAPABILITY statement. These statistics can also be saved in the OUT= data set by specifying the keywords listed in Table 6.52 in the OUTPUT statement.

Standard algorithms (Fisher 1973) are used to compute the moment statistics. The computational methods used by the CAPABILITY procedure are consistent with those used by other SAS procedures for calculating descriptive statistics. For details on statistics also calculated by Base SAS software, see *SAS Visual Data Management and Utility Procedures Guide*.

The following sections give specific details on several statistics calculated by the CAPABILITY procedure.

Mean

The sample mean is calculated as

$$\frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

where n is the number of nonmissing values for a variable, x_i is the ith value of the variable, and w_i is the weight associated with the ith value of the variable. If there is no WEIGHT= variable, the formula reduces to $\frac{1}{n}\sum_{i=1}^{n} x_i$.

Sum

The sum is calculated as $\sum_{i=1}^{n} w_i x_i$, where n is the number of nonmissing values for a variable, x_i is the ith value of the variable, and w_i is the weight associated with the ith value of the variable. If there is no WEIGHT= variable, the formula reduces to $\sum_{i=1}^{n} x_i$.

Sum of the Weights

The sum of the weights is calculated as $\sum_{i=1}^{n} w_i$, where n is the number of nonmissing values for a variable and w_i is the weight associated with the ith value of the variable. If there is no WEIGHT= variable, the sum of the weights is n.

Variance

The variance is calculated as

$$\frac{1}{d}\sum_{i=1}^{n} w_i (x_i - \bar{X}_w)^2$$

where n is the number of nonmissing values for a variable, x_i is the ith value of the variable, $\bar{X}_w$ is the weighted mean, w_i is the weight associated with the ith value of the variable, and d is the divisor controlled by the VARDEF= option in the PROC CAPABILITY statement. If there is no WEIGHT= variable, the formula reduces to

$$\frac{1}{d}\sum_{i=1}^{n} (x_i - \bar{X}_w)^2$$

Standard Deviation

The standard deviation is calculated as

$$\sqrt{\frac{1}{d}\sum_{i=1}^{n} w_i (x_i - \bar{X}_w)^2}$$

where n is the number of nonmissing values for a variable, x_i is the ith value of the variable, $\bar{X}_w$ is the weighted mean, w_i is the weight associated with the ith value of the variable, and d is the divisor controlled by the VARDEF= option in the PROC CAPABILITY statement. If there is no WEIGHT= variable, the formula reduces to

$$\sqrt{\frac{1}{d}\sum_{i=1}^{n} (x_i - \bar{X}_w)^2}$$

Skewness

The sample skewness is calculated as

$$\frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{X}}{s} \right)^3$$

where n is the number of nonmissing values for a variable and must be greater than 2, x_i is the ith value of the variable, $\bar{X}$ is the sample average, and s is the sample standard deviation.

The sample skewness can be positive or negative; it measures the asymmetry of the data distribution and estimates the theoretical skewness $\sqrt{\beta_1} = \mu_3 \mu_2^{-\frac{3}{2}}$, where μ_2 and μ_3 are the second and third central moments. Observations that are normally distributed should have a skewness near zero.

Kurtosis

The sample kurtosis is calculated as

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{X}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

where $n > 3$. The sample kurtosis measures the heaviness of the tails of the data distribution. It estimates the adjusted theoretical kurtosis denoted as $\beta_2 - 3$, where $\beta_2 = \frac{\mu_4}{\mu_2{}^2}$, and μ_4 is the fourth central moment. Observations that are normally distributed should have a kurtosis near zero.

Coefficient of Variation (CV)

The coefficient of variation is calculated as

$$CV = \frac{100 \times s}{\bar{X}}$$

Geometric Mean

The geometric mean is calculated as

$$\left(\prod_{i=1}^{n} x_i^{w_i} \right)^{1/\sum_{i=1}^{n} w_i}$$

where n is the number of nonmissing values for a variable, x_i is the ith value of the variable, and w_i is the weight associated with the ith value of the variable.

If there is no WEIGHT variable, the formula reduces to

$$\left(\prod_{i=1}^{n} x_i \right)^{1/n}$$

If any x_i is negative, the geometric mean is set to missing.

Signed Rank Statistic

The signed rank statistic S is computed as

$$S = \sum_{i:x_i>\mu_0} r_i^+ - \frac{n(n+1)}{4}$$

where r_i^+ is the rank of $|x_i - \mu_0|$ after discarding values of $x_i = \mu_0$, and n is the number of x_i values not equal to μ_0. Average ranks are used for tied values.

If $n \leq 20$, the significance of S is computed from the exact distribution of S, where the distribution is a convolution of scaled binomial distributions. When $n > 20$, the significance of S is computed by treating

$$S\sqrt{\frac{n-1}{nV - S^2}}$$

as a Student t variate with $n - 1$ degrees of freedom. V is computed as

$$V = \frac{1}{24}n(n+1)(2n+1) - \frac{1}{48}\sum t_i(t_i+1)(t_i-1)$$

where the sum is over groups tied in absolute value and where t_i is the number of values in the ith group (Iman 1974, Conover 1980). The null hypothesis tested is that the mean (or median) is μ_0, assuming that the distribution is symmetric. Refer to Lehmann and D'Abrera (1975).

Tests for Normality

You can use the NORMALTEST option in the PROC CAPABILITY statement to request several tests of the hypothesis that the analysis variable values are a random sample from a normal distribution. These tests, which are summarized in the table labeled *Tests for Normality*, include the following:

- Shapiro-Wilk test
- Kolmogorov-Smirnov test
- Anderson-Darling test
- Cramér-von Mises test

Tests for normality are particularly important in process capability analysis because the commonly used capability indices are difficult to interpret unless the data are at least approximately normally distributed. Furthermore, the confidence limits for capability indices displayed in the table labeled *Process Capability Indices* require the assumption of normality. Consequently, the tests of normality are always computed when you specify the SPEC statement, and a note is added to the table when the hypothesis of normality is rejected. You can specify the particular test and the significance level with the CHECKINDICES option.

Shapiro-Wilk Test

If the sample size is 2000 or less, the procedure computes the Shapiro-Wilk statistic W (also denoted as W_n to emphasize its dependence on the sample size n). The statistic W_n is the ratio of the best estimator of the variance (based on the square of a linear combination of the order statistics) to the usual corrected sum of squares estimator of the variance. When n is greater than three, the coefficients to compute the linear

combination of the order statistics are approximated by the method of Royston (1992). The statistic W_n is always greater than zero and less than or equal to one ($0 < W \leq 1$).

Small values of W lead to rejection of the null hypothesis. The method for computing the p-value (the probability of obtaining a W statistic less than or equal to the observed value) depends on n. For $n = 3$, the probability distribution of W is known and is used to determine the p-value. For $n > 4$, a normalizing transformation is computed:

$$Z_n = \begin{cases} (-\log(\gamma - \log(1 - W_n)) - \mu)/\sigma & \text{if } 4 \leq n \leq 11 \\ (\log(1 - W_n) - \mu)/\sigma & \text{if } 12 \leq n \leq 2000 \end{cases}$$

The values of σ, γ, and μ are functions of n obtained from simulation results. Large values of Z_n indicate departure from normality, and because the statistic Z_n has an approximately standard normal distribution, this distribution is used to determine the p-values for $n > 4$.

EDF Tests for Normality

The Kolmogorov-Smirnov, Anderson-Darling and Cramér-von Mises tests for normality are based on the empirical distribution function (EDF) and are often referred to as EDF tests. EDF tests for a variety of non-normal distributions are available in the HISTOGRAM statement; see the section "EDF Goodness-of-Fit Tests" on page 350 for details. For a thorough discussion of these tests, refer to D'Agostino and Stephens (1986).

The empirical distribution function is defined for a set of n independent observations $X_1, \ldots, X_n$ with a common distribution function $F(x)$. Under the null hypothesis, $F(x)$ is the normal distribution. Denote the observations ordered from smallest to largest as $X_{(1)}, \ldots, X_{(n)}$. The empirical distribution function, $F_n(x)$, is defined as

$$F_n(x) = \begin{cases} 0, & x < X_{(1)} \\ \frac{i}{n}, & X_{(i)} \leq x < X_{(i+1)}, i = 1, \ldots, n-1 \\ 1, & X_{(n)} \leq x \end{cases}$$

Note that $F_n(x)$ is a step function that takes a step of height $\frac{1}{n}$ at each observation. This function estimates the distribution function $F(x)$. At any value x, $F_n(x)$ is the proportion of observations less than or equal to x, while $F(x)$ is the probability of an observation less than or equal to x. EDF statistics measure the discrepancy between $F_n(x)$ and $F(x)$.

The EDF tests make use of the probability integral transformation $U = F(X)$. If $F(X)$ is the distribution function of X, the random variable U is uniformly distributed between 0 and 1. Given n observations $X_{(1)}, \ldots, X_{(n)}$, the values $U_{(i)} = F(X_{(i)})$ are computed. These values are used to compute the EDF test statistics, as described in the next three sections. The CAPABILITY procedures computes the associated p-values by interpolating internal tables of probability levels similar to those given by D'Agostino and Stephens (1986).

Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov statistic (D) is defined as

$$D = \sup_x |F_n(x) - F(x)|$$

The Kolmogorov-Smirnov statistic belongs to the supremum class of EDF statistics. This class of statistics is based on the largest vertical difference between $F(x)$ and $F_n(x)$.

The Kolmogorov-Smirnov statistic is computed as the maximum of D^+ and D^-, where D^+ is the largest vertical distance between the EDF and the distribution function when the EDF is greater than the distribution function, and D^- is the largest vertical distance when the EDF is less than the distribution function.

$$\begin{aligned} D^+ &= \max_i \left(\frac{i}{n} - U_{(i)}\right) \\ D^- &= \max_i \left(U_{(i)} - \frac{i-1}{n}\right) \\ D &= \max\left(D^+, D^-\right) \end{aligned}$$

PROC CAPABILITY uses a modified Kolmogorov D statistic to test the data against a normal distribution with mean and variance equal to the sample mean and variance.

Anderson-Darling Test

The Anderson-Darling statistic and the Cramér-von Mises statistic belong to the quadratic class of EDF statistics. This class of statistics is based on the squared difference $(F_n(x) - F(x))^2$. Quadratic statistics have the following general form:

$$Q = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 \, \psi(x) dF(x)$$

The function $\psi(x)$ weights the squared difference $(F_n(x) - F(x))^2$.

The Anderson-Darling statistic (A^2) is defined as

$$A^2 = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 \left[F(x)\left(1 - F(x)\right)\right]^{-1} dF(x)$$

Here the weight function is $\psi(x) = \left[F(x)\left(1 - F(x)\right)\right]^{-1}$.

The Anderson-Darling statistic is computed as

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} \left[(2i-1) \log U_{(i)} + (2n + 1 - 2i) \log\left(\{1 - U_{(i)}\right)\right]$$

Cramér-von Mises Test

The Cramér-von Mises statistic (W^2) is defined as

$$W^2 = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 \, dF(x)$$

Here the weight function is $\psi(x) = 1$.

The Cramér-von Mises statistic is computed as

$$W^2 = \sum_{i=1}^{n} \left(U_{(i)} - \frac{2i-1}{2n}\right)^2 + \frac{1}{12n}$$

Percentile Computations

The CAPABILITY procedure automatically computes the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th percentiles (quantiles), as well as the minimum and maximum of each analysis variable. To compute percentiles other than these default percentiles, use the PCTLPTS= and PCTLPRE= options in the OUTPUT statement.

You can specify one of five definitions for computing the percentiles with the PCTLDEF= option. Let n be the number of nonmissing values for a variable, and let $x_1, x_2, \ldots, x_n$ represent the ordered values of the variable. Let the tth percentile be y, set $p = \frac{t}{100}$, and let

$$\begin{aligned} np &= j + g \quad \text{when PCTLDEF=1, 2, 3, or 5} \\ (n+1)p &= j + g \quad \text{when PCTLDEF=4} \end{aligned}$$

where j is the integer part of np, and g is the fractional part of np. Then the PCTLDEF= option defines the tth percentile, y, as described in the following table:

PCTLDEF=	Description	Formula
1	weighted average at x_{np}	$y = (1-g)x_j + gx_{j+1}$ where x_0 is taken to be x_1
2	observation numbered closest to np	$y = x_i$ if $g \neq \frac{1}{2}$ $y = x_j$ if $g = \frac{1}{2}$ and j is even $y = x_{j+1}$ if $g = \frac{1}{2}$ and j is odd where i is the integer part of $np + \frac{1}{2}$
3	empirical distribution function	$y = x_j$ if $g = 0$ $y = x_{j+1}$ if $g > 0$
4	weighted average aimed at $x_{(n+1)p}$	$y = (1-g)x_j + gx_{j+1}$ where x_{n+1} is taken to be x_n
5	empirical distribution function with averaging	$y = \frac{1}{2}(x_j + x_{j+1})$ if $g = 0$ $y = x_{j+1}$ if $g > 0$

Weighted Percentiles

When you use a WEIGHT statement, the percentiles are computed differently. The 100pth weighted percentile y is computed from the empirical distribution function with averaging

$$y = \begin{cases} x_1 & \text{if } w_1 > pW \\ \frac{1}{2}(x_i + x_{i+1}) & \text{if } \sum_{j=1}^{i} w_j = pW \\ x_{i+1} & \text{if } \sum_{j=1}^{i} w_j < pW < \sum_{j=1}^{i+1} w_j \end{cases}$$

where w_i is the weight associated with x_i, and where $W = \sum_{i=1}^{n} w_i$ is the sum of the weights.

Note that the PCTLDEF= option is not applicable when a WEIGHT statement is used. However, in this case, if all the weights are identical, the weighted percentiles are the same as the percentiles that would be computed without a WEIGHT statement and with PCTLDEF=5.

Confidence Limits for Percentiles

You can use the CIPCTLNORMAL option to request confidence limits for percentiles which assume the data are normally distributed. These limits are described in Section 4.4.1 of Hahn and Meeker (1991). When $0.0 < p < 0.5$, the two-sided $100(1-\alpha)\%$ confidence limits for the $100p$-th percentile are

$$\begin{aligned} \text{lower limit} &= \bar{X} - g'(\alpha/2; 1-p, n)s \\ \text{upper limit} &= \bar{X} - g'(1-\alpha/2; p, n)s \end{aligned}$$

where n is the sample size. When $0.5 \leq p < 1.0$, the two-sided $100(1-\alpha)\%$ confidence limits for the $100p$-th percentile are

$$\begin{aligned} \text{lower limit} &= \bar{X} + g'(\alpha/2; 1-p, n)s \\ \text{upper limit} &= \bar{X} + g'(1-\alpha/2; p, n)s \end{aligned}$$

One-sided $100(1-\alpha)\%$ confidence bounds are computed by replacing $\alpha/2$ by α in the appropriate preceding equation. The factor $g'(\gamma, p, n)$ is related to the noncentral t distribution and is described in Owen and Hua (1977) and Odeh and Owen (1980).

You can use the CIPCTLDF option to request confidence limits for percentiles which are distribution free (in particular, it is not necessary to assume that the data are normally distributed). These limits are described in Section 5.2 of Hahn and Meeker (1991). The two-sided $100(1-\alpha)\%$ confidence limits for the $100p$-th percentile are

$$\begin{aligned} \text{lower limit} &= X_{(l)} \\ \text{upper limit} &= X_{(u)} \end{aligned}$$

where $X_{(j)}$ is the jth order statistic when the data values are arranged in increasing order:

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$$

The lower rank l and upper rank u are integers that are symmetric (or nearly symmetric) around $\lfloor np \rfloor + 1$ where $\lfloor np \rfloor$ is the integer part of np, and where n is the sample size. Furthermore, l and u are chosen so that $X_{(l)}$ and $X_{(u)}$ are as close to $X_{\lfloor np \rfloor + 1}$ as possible while satisfying the coverage probability requirement

$$Q(u-1; n, p) - Q(l-1; n, p) \geq 1-\alpha$$

where $Q(k; n, p)$ is the cumulative binomial probability

$$Q(k; n, p) = \sum_{i=0}^{k} \left(\begin{array}{c} n \\ i \end{array} \right) p^i (1-p)^{n-i}$$

In some cases, the coverage requirement cannot be met, particularly when n is small and p is near 0 or 1. To relax the requirement of symmetry, you can specify CIPCTLDF(TYPE = ASYMMETRIC). This option requests symmetric limits when the coverage requirement can be met, and asymmetric limits otherwise.

If you specify CIPCTLDF(TYPE = LOWER), a one-sided $100(1-\alpha)\%$ lower confidence bound is computed as X_l, where l is the largest integer that satisfies the inequality

$$1 - Q(l-1; n, p) \geq 1-\alpha$$

with $0 < l \leq n$. If you specify CIPCTLDF(TYPE = UPPER), a one-sided $100(1-\alpha)\%$ upper confidence bound is computed as X_u, where u is the smallest integer that satisfies the inequality

$$Q(u-1; n, p) \geq 1-\alpha$$

where $0 < u \leq n$.

Note that confidence limits for percentiles are not computed when a WEIGHT statement is specified.

Robust Estimators

The CAPABILITY procedure provides several methods for computing robust estimates of location and scale, which are insensitive to outliers in the data.

Winsorized Means

The k-times Winsorized mean is a robust estimator of location which is computed as

$$\bar{x}_{wk} = \frac{1}{n}\left((k+1)x_{(k+1)} + \sum_{i=k+2}^{n-k-1} x_{(i)} + (k+1)x_{(n-k)}\right)$$

where n is the number of observations, and $x_{(i)}$ is the ith order statistic when the observations are arranged in increasing order:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

The Winsorized mean is the mean computed after replacing the k smallest observations with the $(k+1)$st smallest observation, and the k largest observations with the $(k+1)$st largest observation.

For data from a symmetric distribution, the Winsorized mean is an unbiased estimate of the population mean. However, the Winsorized mean does not have a normal distribution even if the data are normally distributed.

The Winsorized sum of squared deviations is defined as

$$s^2_{wk} = (k+1)(x_{(k+1)} - \bar{x}_{wk})^2 + \sum_{i=k+2}^{n-k-1}(x_{(i)} - \bar{x}_{wk})^2 + (k+1)(x_{(n-k)} - \bar{x}_{wk})^2$$

A Winsorized t test is given by

$$t_{wk} = \frac{\bar{x}_{wk} - \mu_0}{\mathrm{STDERR}(\bar{x}_{wk})}$$

where the standard error of the Winsorized mean is

$$\mathrm{STDERR}(\bar{x}_{wk}) = \frac{n-1}{n-2k-1}\frac{s_{wk}}{\sqrt{n(n-1)}}$$

When the data are from a symmetric distribution, the distribution of t_{wk} is approximated by a Student's t distribution with $n-2k-1$ degrees of freedom. Refer to Tukey and McLaughlin (1963) and Dixon and Tukey (1968).

A $100(1-\alpha)\%$ Winsorized confidence interval for the mean has upper and lower limits

$$\bar{x}_{wk} \pm t_{1-\alpha/2}\mathrm{STDERR}(\bar{x}_{wk})$$

where $t_{1-\alpha/2}$ is the $(1-\alpha/2)100$th percentile of the Student's t distribution with $n-2k-1$ degrees of freedom.

Trimmed Means

The k-times trimmed mean is a robust estimator of location which is computed as

$$\bar{x}_{tk} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}$$

where n is the number of observations, and $x_{(i)}$ is the ith order statistic when the observations are arranged in increasing order:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

The trimmed mean is the mean computed after the k smallest observations and the k largest observations in the sample are deleted.

For data from a symmetric distribution, the trimmed mean is an unbiased estimate of the population mean. However, the trimmed mean does not have a normal distribution even if the data are normally distributed.

A robust estimate of the variance of the trimmed mean t_{tk} can be obtained from the Winsorized sum of squared deviations; refer to Tukey and McLaughlin (1963). the corresponding trimmed t test is given by

$$t_{tk} = \frac{\bar{x}_{tk} - \mu_0}{\mathrm{STDERR}(\bar{x}_{tk})}$$

where the standard error of the trimmed mean is

$$\mathrm{STDERR}(\bar{x}_{tk}) = \frac{s_{tk}}{\sqrt{(n-2k)(n-2k-1)}}$$

and s_{wk} is the square root of the Winsorized sum of squared deviations.

When the data are from a symmetric distribution, the distribution of t_{tk} is approximated by a Student's t distribution with $n-2k-1$ degrees of freedom. Refer to Tukey and McLaughlin (1963) and Dixon and Tukey (1968).

A $100(1-\alpha)\%$ trimmed confidence interval for the mean has upper and lower limits

$$\bar{x}_{tk} \pm t_{1-\alpha/2}\mathrm{STDERR}(\bar{x}_{tk})$$

where $t_{1-\alpha/2}$ is the $(1-\alpha/2)100$th percentile of the Student's t distribution with $n-2k-1$ degrees of freedom.

Robust Estimates of Scale

The sample standard deviation, which is the most commonly used estimator of scale, is sensitive to outliers. Robust scale estimators, on the other hand, remain bounded when a single data value is replaced by an arbitrarily large or small value. The CAPABILITY procedure computes several robust measures of scale, including the interquartile range Gini's mean difference G, the median absolute deviation about the median (MAD), Q_n, and S_n. In addition, the procedure computes estimates of the normal standard deviation σ derived from each of these measures.

The interquartile range (IQR) is simply the difference between the upper and lower quartiles. For a normal population, σ can be estimated as IQR/1.34898.

Gini's mean difference is computed as

$$G = \frac{1}{\binom{n}{2}} \sum_{i<j} |x_i - x_j|$$

For a normal population, the expected value of G is $2\sigma/\sqrt{\pi}$. Thus $G\sqrt{\pi}/2$ is a robust estimator of σ when the data are from a normal sample. For the normal distribution, this estimator has high efficiency relative to the usual sample standard deviation, and it is also less sensitive to the presence of outliers.

A very robust scale estimator is the MAD, the median absolute deviation from the median (Hampel 1974), which is computed as

$$\text{MAD} = \text{med}_i(|x_i - \text{med}_j(x_j)|)$$

where the inner median, $\text{med}_j(x_j)$, is the median of the n observations, and the outer median (taken over i) is the median of the n absolute values of the deviations about the inner median. For a normal population, 1.4826MAD is an estimator of σ.

The MAD has low efficiency for normal distributions, and it may not always be appropriate for symmetric distributions. Rousseeuw and Croux (1993) proposed two statistics as alternatives to the MAD. The first is

$$S_n = 1.1926\ \text{med}_i(\text{med}_j(|x_i - x_j|))$$

where the outer median (taken over i) is the median of the n medians of $|x_i - x_j|$, $j = 1, 2, \ldots, n$. To reduce small-sample bias, $c_{sn}S_n$ is used to estimate σ, where c_{sn} is a correction factor; refer to Croux and Rousseeuw (1992).

The second statistic is

$$Q_n = 2.2219\ \{|x_i - x_j|; i < j\}_{(k)}$$

where

$$k = \binom{h}{2}$$

and $h = [n/2] + 1$. In other words, Q_n is 2.2219 times the kth order statistic of the $\binom{n}{2}$ distances between the data points. The bias-corrected statistic $c_{qn}Q_n$ is used to estimate σ, where c_{qn} is a correction factor; refer to Croux and Rousseeuw (1992).

Computing the Mode

The mode is the value that occurs most often in a set of observations. The CAPABILITY procedure counts repetitions of the actual values (or the rounded values, if you specify the ROUND= option). If a tie occurs for the most frequent value, the procedure reports the lowest mode in the table labeled *Basic Statistical Measures*. To list all possible modes, specify the MODES option in the PROC CAPABILITY statement. When no repetitions occur in the data, the procedure does not report the mode. The WEIGHT statement has no effect on the mode.

Assumptions and Terminology for Capability Indices

One of the fundamental assumptions in process capability analysis is that the process must be in statistical control. Without statistical control, the process is not predictable, the concept of a process distribution does not apply, and quantities related to the distribution, such as probabilities, percentiles, and capability indices, cannot be meaningfully estimated. Additionally, all of the standard process capability indices described in the next section require that the process distribution be normal, or at least approximately normal.

In many industries, statistical control is routinely checked with a Shewhart chart (such as an $\bar{X}$ and R chart) before capability indices such as

$$C_{pk} = \min\left(\frac{\mathrm{USL}-\mu}{3\sigma}, \frac{\mathrm{LSL}-\mu}{3\sigma}\right)$$

are computed. The control chart analysis yields estimates for the process mean μ and standard deviation σ, which are based on subgrouped data and can be used to estimate C_{pk}. In particular, σ can be estimated by

$$s_R = \bar{R}/d_2$$

rather than the ungrouped sample standard deviation

$$s = \frac{1}{n-1}\sqrt{\sum_{i=1} n(x_i - \bar{x})^2}$$

You can use the SHEWHART procedure to carry out the control chart analysis and to compute capability indices based on s_R. On the other hand, the CAPABILITY procedure computes indices based on s.

Some industry manuals distinguish these two approaches. For instance, the ASQC/AIAG manual *Fundamental Process Control* uses the notation C_{pk} for the estimate based on s_R, and it uses the notation P_{pk} for the estimate based on s. However, assuming that the process is in control and only common cause variation is present, both s_R and s are estimates of the same parameter σ, and so there is fundamentally no difference in the two approaches[2].

Once control has been established, attention should focus on the distribution of the process measurements, and at this point there is no practical or statistical advantage to working with subgrouped measurements. In fact, the use of s is closely associated with a wide variety of methods that are highly useful for process capability analysis, including tests for normality, graphical displays such as histograms and probability plots, and confidence intervals for parameters and capability indices.

Standard Capability Indices

This section provides computational details for the standard process capability indices computed by the CAPABILITY procedure: C_p, CPL, CPU, C_{pk}, and C_{pm}.

The Index Cp

The process capability index C_p, sometimes called the "process potential index," the "process capability ratio," or the "inherent capability index," is estimated as

$$\widehat{C}_p = \frac{USL - LSL}{6s}$$

[2]Statistically, s is a more efficient estimator of σ than s_R.

where *USL* is the upper specification limit, *LSL* is the lower specification limit, and *s* is the sample standard deviation. If you do not specify both the upper and the lower specification limits in the SPEC statement or the SPEC= data set, then C_p is assigned a missing value.

The interpretation of C_p can depend on the application, on past experience, and on local practice. However, broad guidelines for interpretation have been proposed by several authors. Ekvall and Juran (1974) classify C_p values as

- "not adequate" if $C_p < 1$
- "adequate" if $1 \leq C_p \leq 1.33$, but requiring close control as C_p approaches 1
- "more than adequate" if $C_p > 1.33$

Montgomery (1996) recommends minimum values of C_p as

- 1.33 for existing processes
- 1.50 for new processes or for existing processes when the variable is critical (for example, related to safety or strength)
- 1.67 for new processes when the variable is critical

Exact $100(1-\alpha)\%$ lower and upper confidence limits for C_p (denoted by LCL and UCL) are computed using percentiles of the chi-square distribution, as indicated by the following equations:

$$\begin{aligned} \text{lower limit} &= \hat{C}_p \sqrt{\chi^2_{\alpha/2,n-1}/(n-1)} \\ \text{upper limit} &= \hat{C}_p \sqrt{\chi^2_{1-\alpha/2,n-1}/(n-1)} \end{aligned}$$

Here, $\chi^2_{\alpha,\nu}$ denotes the lower 100αth percentile of the chi-square distribution with ν degrees of freedom. Refer to Chou, Owen, and Borrego (1990) and Kushler and Hurley (1992).

You can specify α with the ALPHA= option in the PROC CAPABILITY statement or with the CIINDICES(ALPHA=*value*) in the PROC CAPABILITY statement. The default value is 0.05. You can save these limits in the OUT= data set by specifying the keywords CPLCL and CPUCL in the OUTPUT statement. In addition, you can display these limits on plots produced by the CAPABILITY procedure by specifying the keywords in the INSET statement.

The Index CPL

The process capability index *CPL* is estimated as

$$\widehat{\mathrm{CPL}} = \frac{\bar{X} - LSL}{3s}$$

where $\bar{X}$ is the sample mean, *LSL* is the lower specification limit, and *s* is the sample standard deviation. If you do not specify the lower specification limit in the SPEC statement or the SPEC= data set, then *CPL* is assigned a missing value.

Montgomery (1996) refers to *CPL* as the "process capability ratio" in the case of one-sided lower specifications and recommends minimum values as follows:

- 1.25 for existing processes
- 1.45 for new processes or for existing processes when the variable is critical
- 1.60 for new processes when the variable is critical

Exact $100(1-\alpha)\%$ lower and upper confidence limits for *CPL* are computed using a generalization of the method of Chou, Owen, and Borrego (1990), who point out that the $100(1-\alpha)$ lower confidence limit for *CPL* (denoted by CPLLCL)satisfies the equation

$$\Pr\{T_{n-1}(\delta = 3\sqrt{n})\ \text{CPLLCL}\ \leq 3\text{CPL}\sqrt{n}\} = 1-\alpha$$

where $T_{n-1}(\delta)$ has a non-central *t* distribution with *n* – 1 degrees of freedom and noncentrality parameter δ. You can specify α with the ALPHA= option in the PROC CAPABILITY statement. The default value is 0.05. The confidence limits can be saved in an output data set by specifying the keywords CPLLCL and CPLUCL in the OUTPUT statement. In addition, you can display these limits on plots produced by the CAPABILITY procedure by specifying these keywords in the INSET statement.

The Index CPU

The process capability index *CPU* is estimated as

$$\widehat{\text{CPU}} = \frac{USL - \bar{X}}{3s}$$

where *USL* is the upper specification limit, $\bar{X}$ is the sample mean, and *s* is the sample standard deviation. If you do not specify the upper specification limit in the SPEC statement or the SPEC= data set, then *CPU* is assigned a missing value.

Montgomery (1996) refers to *CPU* as the "process capability ratio" in the case of one-sided upper specifications and recommends minimum values that are the same as those specified previously for *CPL*.

Exact $100(1-\alpha)\%$ lower and upper confidence limits for *CPU* are computed using a generalization of the method of Chou, Owen, and Borrego (1990), who point out that the $100(1-\alpha)$ lower confidence limit for *CPU* (denoted by CPULCL)satisfies the equation

$$\Pr\{T_{n-1}(\delta = 3\sqrt{n}\ \text{CPULCL}\ \geq 3\text{CPU}\sqrt{n}\} = 1-\alpha$$

where $T_{n-1}(\delta)$ has a non-central *t* distribution with *n* – 1 degrees of freedom and noncentrality parameter δ. You can specify α with the ALPHA= option in the PROC CAPABILITY statement. The default value is 0.05. The confidence limits can be saved in an output data set by specifying the keywords CPULCL and CPUUCL in the OUTPUT statement. In addition, you can display these limits on plots produced by the CAPABILITY procedure by specifying these keywords in the INSET statement.

The Index Cpk

The process capability index C_{pk} is defined as

$$C_{pk} = \frac{1}{3\sigma}\min(USL-\mu,\ \mu-LSL) = \min(CPU,\ CPL)$$

Note that the indices C_{pk}, C_p, and k are related as $C_{pk} = C_p(1-k)$. The CAPABILITY procedure estimates C_{pk} as

$$\widehat{C}_{pk} = \frac{1}{3s} \times \min(USL - \bar{X},\ \bar{X} - LSL) = \min(CPU,\ CPL)$$

where *USL* is the upper specification limit, *LSL* is the lower specification limit, $\bar{X}$ is the sample mean, and s is the sample standard deviation.

If you specify only the upper limit in the SPEC statement or the SPEC= data set, then C_{pk} is computed as *CPU*, and if you specify only the lower limit in the SPEC statement or the SPEC= data set, then C_{pk} is computed as *CPL*.

Bissell (1990) derived approximate two-sided 95% confidence limits for C_{pk} by assuming that the distribution of $\widehat{C}_{pk}$ is normal. Using Bissell's approach, $100(1-\alpha)\%$ lower and upper confidence limits can be computed as

$$\text{lower limit} = \widehat{C}_{pk}\left[1 - \Phi^{-1}(1-\alpha/2)\sqrt{\frac{1}{9n\widehat{C}_{pk}^2} + \frac{1}{2(n-1)}}\ \right]$$

$$\text{upper limit} = \widehat{C}_{pk}\left[1 + \Phi^{-1}(1-\alpha/2)\sqrt{\frac{1}{9n\widehat{C}_{pk}^2} + \frac{1}{2(n-1)}}\ \right]$$

where Φ denotes the cumulative standard normal distribution function. Kushler and Hurley (1992) concluded that Bissell's method gives reasonably accurate results.

You can specify α with the ALPHA= option in the PROC CAPABILITY statement. The default value is 0.05. These limits can be saved in an output data set by specifying the keywords CPKLCL and CPKUCL in the OUTPUT statement. In addition, you can display these limits on plots produced by the CAPABILITY procedure by specifying these same keywords in the INSET statement.

The Index Cpm

The process capability index C_{pm} is intended to account for deviation from the target T in addition to variability from the mean. This index is often defined as

$$C_{pm} = \frac{\text{USL} - \text{LSL}}{6\sqrt{\sigma^2 + (\mu - T)^2}}$$

A closely related version of C_{pm} is the index

$$C_{pm}^* = \frac{\min(\text{USL} - T, T - \text{LSL})}{3\sqrt{\sigma^2 + (\mu - T)^2}} = \frac{d - |T - m|}{3\sqrt{\sigma^2 + (\mu - T)^2}}$$

where $d = (\text{USL} - \text{LSL})/2$ and $m = (\text{USL} + \text{LSL})/2$. If $T = m$, then $C_{pm} = C_{pm}^*$. However, if $T \neq m$, then both indices suffer from problems of interpretation, as pointed out by Kotz and Johnson (1993), and their use should be avoided in this case.

The CAPABILITY procedure computes an estimator of C_{pm} as

$$\widehat{C}_{pm} = \frac{\min(USL - T,\ T - LSL)}{3\sqrt{s^2 + (\bar{X} - T)^2}}$$

where s is the sample standard deviation.

If you specify only a single specification limit SL in the SPEC statement or the SPEC= data set, then C_{pm} is estimated as

$$\widehat{C}_{pm} = \frac{|T - SL|}{3\sqrt{s^2 + (\bar{X} - T)^2}}$$

Boyles (1991) proposed a slightly modified point estimate for C_{pm} computed as

$$\widetilde{C}_{pm} = \frac{(\mathrm{USL} - \mathrm{LSL})/2}{3\sqrt{(\frac{n-1}{n})s^2 + (\bar{X} - T)^2}}$$

Boyles also suggested approximate two-sided $100(1-\alpha)\%$ confidence limits for C_{pm}, which are computed as

$$\begin{aligned} \text{lower limit} &= \widetilde{C}_{pm}\sqrt{\chi^2_{\alpha/2,\nu}/\nu} \\ \text{upper limit} &= \widetilde{C}_{pm}\sqrt{\chi^2_{1-\alpha/2,\nu}/\nu} \end{aligned}$$

Here $\chi^2_{\alpha,\nu}$ denotes the lower 100αth percentile of the chi-square distribution with ν degrees of freedom, where ν equals

$$\frac{n(1 + (\frac{\bar{X}-T}{s})^2)}{1 + 2(\frac{\bar{X}-T}{s})^2}$$

You can specify α with the ALPHA= option in the PROC CAPABILITY statement. The default value is 0.05. These confidence limits can be saved in an output data set by specifying the keywords CPMLCL and CPMUCL in the OUTPUT statement. In addition, you can display these limits on plots produced by the CAPABILITY procedure by specifying these keywords in the INSET statement.

Specialized Capability Indices

This section describes a number of specialized capability indices which you can request with the SPECIALINDICES option in the PROC CAPABILITY statement.

The Index k

The process capability index k (also denoted by K) is computed as

$$k = \frac{2|m - \bar{X}|}{USL - LSL}$$

where $m = \frac{1}{2}(USL + LSL)$ is the midpoint of the specification limits, $\bar{X}$ is the sample mean, *USL* is the upper specification limit, and *LSL* is the lower specification limit.

The formula for k used here is given by Kane (1986). Note that k is sometimes computed without taking the absolute value of $m - \bar{X}$ in the numerator. See Wadsworth, Stephens, and Godfrey (1986).

If you do not specify the upper and lower limits in the SPEC statement or the SPEC= data set, then k is assigned a missing value.

Boyles' Index C_{pm}^{+}

Boyles (1992) proposed the process capability index C_{pm}^{+} which is defined as

$$C_{pm}^{+} = \frac{1}{3}\left[\frac{E_{X<T}\left[(X-T)^2\right]}{(T-\mathrm{LSL})^2} + \frac{E_{X>T}\left[(X-T)^2\right]}{(\mathrm{USL}-T)^2}\right]^{-1/2}$$

He proposed this index as a modification of C_{pm} for use when $\mu \neq T$. The quantities

$$E_{X<T}\left[(X-T)^2\right] = E\left[(X-T)^2 | X < T\right] Pr\left[X < T\right]$$

and

$$E_{X>T}\left[(X-T)^2\right] = E\left[(X-T)^2 | X > T\right] Pr\left[X > T\right]$$

are referred to as semivariances. Kotz and Johnson (1993) point out that if $T = (\mathrm{LSL} + \mathrm{USL})/2$, then $C_{pm}^{+} = C_{pm}$.

Kotz and Johnson (1993) suggest that a natural estimator for C_{pm}^{+} is

$$\widehat{C}_{pm}^{+} = \frac{1}{3}\left[\frac{1}{n}\left\{\frac{\sum_{X_i<T}(X_i-T)^2}{(T-\mathrm{LSL})^2} + \frac{\sum_{X_i>T}(X_i-T)^2}{(\mathrm{USL}-T)^2}\right\}^{-1/2}\right]$$

Note that this index is not defined when either of the specification limits is equal to the target T. Refer to Section 3.5 of Kotz and Johnson (1993) for further details.

The Index C_{jkp}

Johnson, Kotz, and Pearn (1994) introduced a so-called "flexible" process capability index which takes into account possible differences in variability above and below the target T. They defined this index as

$$C_{jkp} = \frac{1}{3\sqrt{2}} \min\left(\frac{\mathrm{USL}-T}{\sqrt{E_{X>T}[(X-T)^2]}}, \frac{T-\mathrm{LSL}}{\sqrt{E_{X<T}[(X-T)^2]}}\right)$$

where $d = (\mathrm{USL} - \mathrm{LSL})/2$.

A natural estimator of this index is

$$\widehat{C}_{jkp} = \frac{1}{3\sqrt{2}} \min\left(\frac{\text{USL} - T}{\sqrt{\sum_{X_i > T}(X_i - T)^2/n}}, \frac{T - \text{LSL}}{\sqrt{\sum_{X_i < T}(X_i - T)^2/n}} \right)$$

For further details, refer to Section 4.4 of Kotz and Johnson (1993).

The Indices $C_{pm}(a)$

The class of capability indices $C_{pm}(a)$, indexed by the parameter a ($a > 0$) allows flexibility in choosing between the relative importance of variability and deviation of the mean from the target value T.

The class defined as

$$C_{pm}(a) = (1 - a\zeta^2)C_p$$

where $\zeta = (\mu - T)/\sigma$. The motivation for this definition is that if $|\zeta|$ is small, then

$$C_{pm} \approx (1 - \frac{1}{2}\zeta^2)C_p$$

A natural estimator of $C_{pm}(a)$ is

$$\frac{d}{3s}\widehat{C}_{pm}(a) = \left\{ 1 - a\left(\frac{\bar{X} - T}{s} \right)^2 \right\}$$

where $d = (\text{USL} - \text{LSL})/2$. You can specify the value of a with the SPECIALINDICES(CPMA=) option in the PROC CAPABILITY statement. By default, $a = 0.5$.

This index is not recommended for situation in which the target T is not equal to the midpoint of the specification limits.

For additional details, refer to Section 3.7 of Kotz and Johnson (1993).

The Index $C_{p(5.15)}$

Johnson *et al.* (1992) suggest the class of process capability indices defined as

$$C_{p(\theta)} = \frac{\text{USL} - \text{LSL}}{\theta\sigma}$$

where θ is chosen so that the proportion of conforming items is robust with respect to the shape of the process distribution. In particular, Kotz and Johnson (1993) recommend use of

$$C_{p(5.15)} = \frac{\text{USL} - \text{LSL}}{5.15\sigma}$$

which is estimated as

$$\widehat{C}_{p(5.15)} = \frac{\text{USL} - \text{LSL}}{5.15s}$$

For details, refer to Section 4.3.2 of Kotz and Johnson (1993).

The Index $C_{pk(5.15)}$

Similarly, Kotz and Johnson (1993) recommend use of the robust capability index

$$C_{pk(5.15)} = \frac{d - |\mu - (\mathrm{USL} + \mathrm{LSL})/2|}{2.575\sigma}$$

where $d = (\mathrm{USL} - \mathrm{LSL})/2$. This index is estimated as

$$\widehat{C}_{pk(5.15)} = \frac{d - |\bar{X} - (\mathrm{USL} + \mathrm{LSL})/2|}{2.575s}$$

For details, refer to Section 4.3.2 of Kotz and Johnson (1993).

The Index C_{pmk}

Pearn, Kotz, and Johnson (1992) proposed the index C_{pmk}

$$C_{pmk} = \frac{(\mathrm{USL} - \mathrm{LSL})/2 - |\mu - m|}{3\sqrt{\sigma^2 + (\mu - T)^2}}$$

where $m = (\mathrm{LCL} + \mathrm{UCL})/2$. A natural estimator for C_{pmk} is

$$\widehat{C}_{pmk} = \frac{(\mathrm{USL} - \mathrm{LSL})/2 - |\bar{X} - m|}{3\sqrt{(\frac{n-1}{n})s^2 + (\bar{X} - T)^2}}$$

where $m = (\mathrm{USL} + \mathrm{LSL})/2$.

For further details, refer to Section 3.6 of Kotz and Johnson (1993).

Wright's Index C_s

Wright (1995) defines the capability index

$$C_s = \frac{\min(\mathrm{USL} - \mu, \mu - \mathrm{LSL})}{3\sqrt{\sigma^2 + (\mu - T)^2 + \mu_3/\sigma}}$$

where $\mu_3 = E(X - \mu)^3$.

A natural estimator of C_s is

$$\widehat{C}_s = \frac{(\mathrm{USL} - \mathrm{LSL})/2 - |\bar{X} - m|}{3\sqrt{\left(\frac{n-1}{n}\right)s^2 + (\bar{X} - T)^2 + |c_4 s^2 b_3|}}$$

where c_4 is an unbiasing constant for the sample standard deviation, and b_3 is a measure of skewness. Wright (1995) shows that C_s compares favorably with C_{pmk} even when skewness is not present, and he advocates the use of C_s for monitoring near-normal processes when loss of capability typically leads to asymmetry.

Chen and Kotz (1996) proposed a modification to Wright's C_s index which introduces a multiplier, $\gamma > 0$, and is estimated as

$$\widehat{C}_s = \frac{(\mathrm{USL} - \mathrm{LSL})/2 - |\bar{X} - m|}{3\sqrt{\left(\frac{n-1}{n}\right)s^2 + (\bar{X} - T)^2 + \gamma|c_4 s^2 b_3|}}$$

If you specify a value for γ with the SPECIALINDICES(CSGAMMA=) option, the index C_s is computed with this modification. Otherwise it is computed using Wright's original definition.

The Index S_{jkp}

Boyles (1994) proposed a smooth version of C_{jkp} defined as

$$S_{jkp} = S\left(\frac{\text{USL} - T}{\sqrt{2E_{X>T}[(X-T)^2]}}, \frac{T - \text{LSL}}{\sqrt{2E_{X<T}[(X-T)^2]}}\right)$$

The CAPABILITY procedure estimates S_{jkp} as

$$\widehat{S}_{jkp} = S\left(\frac{\text{USL} - T}{\sqrt{2\sum_{X_i>T}(X_i-T)^2/n}}, \frac{T - \text{LSL}}{\sqrt{2\sum_{X_i<T}(X_i-T)^2/n}}\right)$$

where $S(x, y) = \Phi^{-1}[\{\Phi(x) + \Phi(y)\}/2]/3$.

The Index C_{pp}

Chen (1998) devised a process incapability index based on the C^*_{pm} index. The first term measures *inaccuracy* and the second measures *imprecision*. The C_{pp} index is estimated as

$$\widehat{C}_{pp} = \left(\frac{\bar{X} - T}{d^*/3}\right)^2 + \left(\frac{s}{d^*/3}\right)^2$$

where $d^* = \min(\text{USL} - T, T - \text{LSL})$.

The Index C''_{pp}

The index C_{pp} does not handle asymmetric tolerances well, as discussed by Kotz and Lovelace (1998). To address that shortcoming, Chen (1998) defined the index C''_{pp}, which is estimated by

$$\widehat{C}''_{pp} = \left(\frac{\widehat{A}}{d^*/3}\right)^2 + \left(\frac{s}{d^*/3}\right)$$

where

$$\widehat{A} = \max\left\{\frac{(\bar{X} - T)d}{T - \text{LSL}}, \frac{(T - \bar{X})d}{\text{USL} - T}\right\}$$

and $d = (USL - LSL)/2$.

The Index C_{pg}

Marcucci and Beazley (1988) defined the index

$$C_{pg} = \frac{1}{C_{pm}^2}$$

which is estimated as

$$\widehat{C}_{pg} = \frac{1}{\widehat{C}_{pm}^2}$$

The Index C_{pq}

Gupta and Kotz (1997) introduced the index C_{pq}, which is estimated by

$$\widehat{C}_{pq} = \widehat{C}_p \left[1 - \frac{1}{2} \left(\frac{\bar{X} - T}{s} \right)^2 \right]$$

The Index C_p^W

Bai and Choi (1997) defined the index

$$C_p^W = \frac{C_p}{\sqrt{1 + |1 - 2P_x|}}$$

where $P_x = \Pr(X \leq \mu)$. It is estimated by

$$\widehat{C}_p^W = \frac{\widehat{C}_p}{\sqrt{1 + |1 - 2\widehat{P}_x|}}$$

where $\widehat{P}_x$ is the fraction of observations less than or equal to $\bar{X}$. For more information about C_p^W, see Kotz and Lovelace (1998).

The Index C_{pk}^W

Bai and Choi (1997) also proposed the index

$$C_{pk}^W = \min \left\{ \frac{\text{USL} - \mu}{3\sigma\sqrt{2P_x}}, \frac{\mu - \text{LSL}}{3\sigma\sqrt{2(1 - P_x)}} \right\}$$

It is estimated by

$$\widehat{C}_{pk}^{W} = \min\left\{\frac{\text{USL} - \bar{X}}{3s\sqrt{2\widehat{P}_x}}, \frac{\bar{X} - \text{LSL}}{3s\sqrt{2(1-\widehat{P}_x)}}\right\}$$

where $\widehat{P}_x$ is the fraction of observations less than or equal to $\bar{X}$. For more information about C_{pk}^{W}, see Kotz and Lovelace (1998).

The Index C_{pm}^{W}

The index C_{pm}^{W}, also introduced by Bai and Choi (1997), is defined as

$$C_{pm}^{W} = \frac{C_{pm}}{\sqrt{1 + |1 - 2P_T|}}$$

where $P_T = \Pr(X \leq T)$. It is estimated by

$$\widehat{C}_{pm}^{W} = \frac{\widehat{C}_{pm}}{\sqrt{1 + |1 - 2\widehat{P}_T|}}$$

where $\widehat{P}_T$ is the fraction of observations less than or equal to T. For more information about C_{pm}^{W}, see Kotz and Lovelace (1998).

The Index C_{pc}

Luceño (1996) proposed the index

$$C_{pc} = \frac{\text{USL} - \text{LSL}}{6\sqrt{\frac{\pi}{2}E|X - M|}}$$

where $M = (\text{USL} + \text{LSL})/2$. It is estimated by

$$\widehat{C}_{pc} = \frac{\text{USL} - \text{LSL}}{6\sqrt{\frac{\pi}{2}c}}$$

where

$$c = \frac{1}{n}\sum_{i=1}^{n}|X_i - M|$$

Vännmann's Index $C_p(u, v)$

Vännmann (1995) introduced the generalized index $C_p(u, v)$, which reduces to the following capability indices given appropriate choices of u and v:

- $C_p(0, 0) = C_p$
- $C_p(0, 1) = C_{pk}$
- $C_p(1, 0) = C_{pm}$
- $C_p(1, 1) = C_{pmk}$

$C_p(u, v)$ is defined as

$$C_p(u, v) = \frac{d - u|\,\mu - M\,|}{3\sqrt{\sigma^2 + v(\,\mu - T)^2}}$$

and estimated by

$$\widehat{C}_p(u, v) = \frac{d - u|\bar{X} - M\,|}{3\sqrt{(\frac{n-1}{n})s^2 + v(\bar{X} - T)^2}}$$

You can specify u with the SPECIALINDICES(CPU=) option and v with the SPECIALINDICES(CPV=) option. By default, $u = 0$ and $v = 4$.

Vännmann's Index $C_p(v)$

Vännmann (1997) also proposed the index $C_p(v)$, which is equivalent to $C_p(u, v)$ with $u = 1$. It is estimated as

$$\widehat{C}_p(v) = \frac{d - |\bar{X} - M\,|}{3\sqrt{(\frac{n-1}{n})s^2 + v(\bar{X} - T)^2}}$$

You can specify v with the SPECIALINDICES(CPV=) option. By default, $v = 4$.

Missing Values

If a variable for which statistics are calculated has a missing value, that value is ignored in the calculation of statistics, and the missing values are tabulated separately. A missing value for one such variable does not affect the treatment of other variables in the same observation.

If the WEIGHT variable has a missing value, the observation is excluded from the analysis. If the FREQ variable has a missing value, the observation is excluded from the analysis. If a variable in a BY or ID statement has a missing value, the procedure treats it as it would treat any other value of a BY or ID variable.

ODS Tables

This section describes the ODS tables produced by the CAPABILITY procedure.

Table 6.5 summarizes the ODS tables that you can request with options in the PROC CAPABILITY statement.

Table 6.5 ODS Tables Produced with the PROC CAPABILITY Statement

Table Name	Description	Option
BasicIntervals	confidence intervals for mean, standard deviation, variance	CIBASIC
BasicMeasures	measures of location and variability	default
ExtremeObs	extreme observations	default
ExtremeValues	extreme values	NEXTRVAL=
Frequencies	frequencies	FREQ
LocationCounts	counts used for sign test and signed rank test	LOCCOUNTS
MissingValues	missing values	default
Modes	modes	MODES
Moments	sample moments	default
Quantiles	quantiles	default
RobustScale	robust measures of scale	ROBUSTSCALE
TestsForLocation	tests for location	default
TestsForNormality	tests for normality	NORMALTEST
TrimmedMeans	trimmed means	TRIMMED=
WinsorizedMeans	Winsorized means	WINSORIZED=

Table 6.6 summarizes the ODS tables related to capability indices that you can request with options in the PROC CAPABILITY statement when you provide specification limits with a SPEC statement or with a SPEC= data set.

Table 6.6 ODS Tables Related to Specification Limits

Table Name	Description	Option
CIProbExSpecs	confidence limits for probabilities of exceeding specifications	CIPROBEX
Indices	standard capability indices	default
SpecialIndices	specialized capability indices	SPECIALINDICES
Specifications	percents outside specification limits based on empirical	default

Table 6.7 summarizes the ODS tables related to fitted distributions that you can request with options in the HISTOGRAM statement.

Table 6.7 ODS Tables Produced with the HISTOGRAM Statement

Table Name	Description	Option
Bins	histogram bins	MIDPERCENTS suboption with any distribution option, such as NORMAL(MIDPERCENTS)
FitIndices	capability indices computed from fitted distribution	INDICES suboption with any distribution option, such as LOGNORMAL(INDICES)
FitQuantiles	quantiles of fitted distribution	any distribution option such as NORMAL
GoodnessOfFit	goodness-of-fit tests for fitted distribution	any distribution option such as NORMAL
ParameterEstimates	parameter estimates for fitted distribution	any distribution option such as NORMAL
Specifications	percents outside specification limits based on empirical and fitted distributions	any distribution option such as NORMAL

The following table summarizes the ODS tables that you can request with options in the INTERVALS statement.

Table 6.8 ODS Tables Produced with the INTERVALS Statement

Table Name	Description	Option
Intervals1	prediction interval for future observations	METHODS=1
Intervals2	prediction interval for mean	METHODS=2
Intervals3	tolerance interval for proportion of population	METHODS=3
Intervals4	confidence limits for mean	METHODS=4
Intervals5	prediction interval for standard deviation	METHODS=5
Intervals6	confidence limits for standard deviation	METHODS=6

Examples: CAPABILITY Procedure

This section provides a more advanced example of the PROC CAPABILITY statement.

Example 6.1: Reading Specification Limits

NOTE: See *Reading Spec Limits from an Input Data Set* in the SAS/QC Sample Library.

You can specify specification limits either in the SPEC statement or in a SPEC= data set. In "Computing Capability Indices" on page 199, limits were specified in a SPEC statement. This example illustrates how

to create a SPEC= data set to read specification limits with the SPEC= option in the PROC CAPABILITY statement.

Consider the drink can data presented in "Computing Descriptive Statistics" on page 197. Suppose, in addition to the fluid weight of each drink can, the weight of the can itself is stored in a variable named Cweight, and both variables are saved in a data set called Can2. A partial listing of Can2 follows:

```
proc print data=Can2(obs=5);
run;
```

Output 6.1.1 The Data Set Can2

Process Capability Analysis of Fluid Weight

Obs	Weight	Cweight
1	12.07	1.07
2	12.02	0.86
3	12.00	1.06
4	12.01	1.08
5	11.98	1.02

The following DATA step creates a data set named Limits containing specification limits for the fluid weight and the can weight. Limits has 4 variables (_VAR_, _LSL_, _USL_, and _TARGET_) and 2 observations. The first observation contains the specification limit information for the variable Weight, and the second contains the specification limit information for the variable Cweight.

```
data Limits;
   length _var_ $8;
   _var_    = 'Weight';
   _lsl_    = 11.95;
   _target_ = 12;
   _usl_    = 12.05;
   output;
   _var_    = 'Cweight';
   _lsl_    = 0.90;
   _target_ = 1;
   _usl_    = 1.10;
   output;
run;
```

The following statements read the specification information from the Limits data set into the CAPABILITY procedure by using the SPEC= option. These statements print summary statistics, capability indices, and specification limit information for Weight and Cweight. Figure 6.1 and Figure 6.2 display the output for Weight. Output 6.1.2 displays the output for Cweight.

```
title 'Process Capability Analysis of Drink Can Data';
proc capability data=Can2 specs=Limits;
   var Cweight;
run;
```

Output 6.1.2 Printed Output for Variable Cweight

Process Capability Analysis of Drink Can Data

The CAPABILITY Procedure
Variable: Cweight (Can Weight (ounces))

Moments			
N	100	**Sum Weights**	100
Mean	1.004	**Sum Observations**	100.4
Std Deviation	0.06330941	**Variance**	0.00400808
Skewness	-0.074821	**Kurtosis**	-0.5433858
Uncorrected SS	101.1984	**Corrected SS**	0.3968
Coeff Variation	6.30571767	**Std Error Mean**	0.00633094

Basic Statistical Measures			
Location		**Variability**	
Mean	1.004000	**Std Deviation**	0.06331
Median	1.000000	**Variance**	0.00401
Mode	1.040000	**Range**	0.29000
		Interquartile Range	0.08500

Note: The mode displayed is the smallest of 2 modes with a count of 8.

Tests for Location: Mu0=0				
Test		**Statistic**	**p Value**	
Student's t	**t**	158.5862	**Pr > |t|**	<.0001
Sign	**M**	50	**Pr >= |M|**	<.0001
Signed Rank	**S**	2525	**Pr >= |S|**	<.0001

Tests for Normality				
Test		**Statistic**	**p Value**	
Shapiro-Wilk	**W**	0.987310	**Pr < W**	0.4588
Kolmogorov-Smirnov	**D**	0.061410	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.048175	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.361939	**Pr > A-Sq**	>0.2500

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	1.150
99%	1.140
95%	1.105
90%	1.080
75% Q3	1.045
50% Median	1.000
25% Q1	0.960
10%	0.910
5%	0.900
1%	0.870
0% Min	0.860

Output 6.1.2 *continued*

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0.86	2	1.11	42
0.88	89	1.12	28
0.88	64	1.12	34
0.90	68	1.13	48
0.90	59	1.15	52

Specification Limits			
Limit		Percent	
Lower (LSL)	0.900000	% < LSL	3.00000
Target	1.000000	% Between	92.00000
Upper (USL)	1.100000	% > USL	5.00000

Process Capability Indices			
Index	Value	95% Confidence Limits	
Cp	0.526515	0.453237	0.599670
CPL	0.547575	0.446607	0.647299
CPU	0.505454	0.408856	0.600808
Cpk	0.505454	0.409407	0.601501
Cpm	0.525467	0.454973	0.601113

Example 6.2: Enhancing Reference Lines

NOTE: See *Controlling the Appearance of Spec Limits* in the SAS/QC Sample Library.

A telecommunications company manufactures amplifiers to be used in telephones. Each amplifier is designed to boost the input signal by 5 decibels (dB). Because it is difficult to make every amplifier's boosting power exactly 5 decibels, the company decides that amplifiers that boost the input signal between 4 and 6 decibels are acceptable. Therefore, the target value is 5 decibels, and the lower and upper specification limits are 4 and 6 decibels, respectively. The following data set contains the boosting powers of a sample of 75 amplifiers:

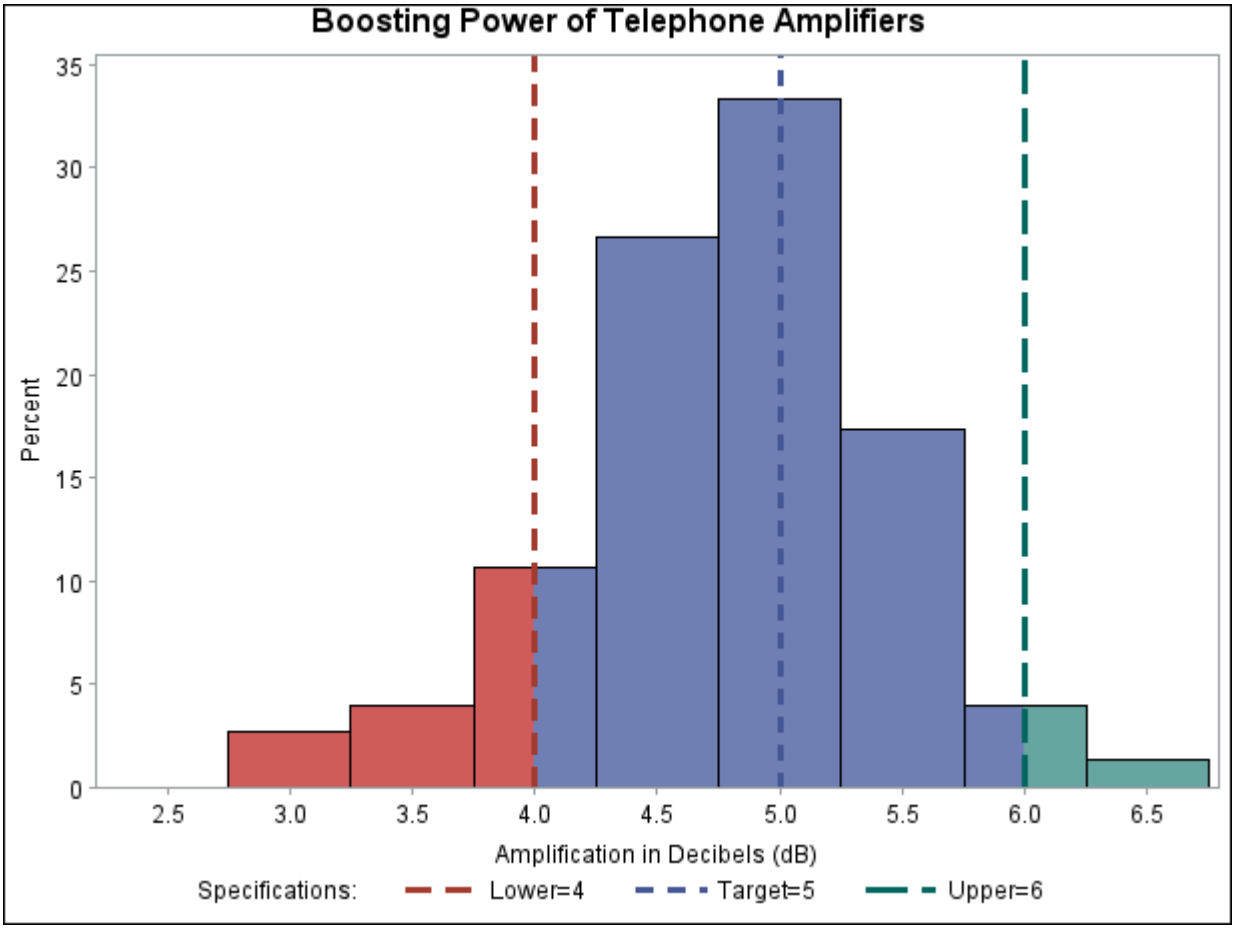
```
data Amps;
   label Decibels = 'Amplification in Decibels (dB)';
   input Decibels @@;
   datalines;
4.54 4.87 4.66 4.90 4.68 5.22 4.43 5.14 3.07 4.22
5.09 3.41 5.75 5.16 3.96 5.37 5.70 4.11 4.83 4.51
4.57 4.16 5.73 3.64 5.48 4.95 4.57 4.46 4.75 5.38
5.19 4.35 4.98 4.87 3.53 4.46 4.57 4.69 5.27 4.67
5.03 4.50 5.35 4.55 4.05 6.63 5.32 5.24 5.73 5.08
5.07 5.42 5.05 5.70 4.79 4.34 5.06 4.64 4.82 3.24
4.79 4.46 3.84 5.05 5.46 4.64 6.13 4.31 4.81 4.98
4.95 5.57 4.11 4.15 5.95
;
```

The SPEC statement provides several options to control the appearance of reference lines for the specification limits and the target value. The following statements use the data set Amps to create a histogram that demonstrates some of these options:

```
ods graphics off;
legend2 FRAME CFRAME=ligr CBORDER=black POSITION=center;
title 'Boosting Power of Telephone Amplifiers';
proc capability data=Amps;
   spec  target = 5      lsl = 4         usl = 6
         ltarget = 2     llsl = 3        lusl = 4
         wtarget = 2     wlsl = 2        wusl = 2
         cleft           cright;
   histogram Decibels / cbarline = black;
run;
```

The resulting histogram is shown in Output 6.2.1. The LTARGET=, LLSL=, and LUSL= options control the line type of the reference lines for the target, lower specification limit, and upper specification limit, respectively. Likewise, the WTARGET=, WLSL=, and WUSL= options control the line widths. The CLEFT= option controls the color used to fill the area to the left of the lower specification limit. Similarly, the CRIGHT= option controls the color used to fill the area to the right of the upper specification limit.

Output 6.2.1 Controlling the Appearance of Specification Limits



Example 6.3: Displaying a Confidence Interval for Cpk

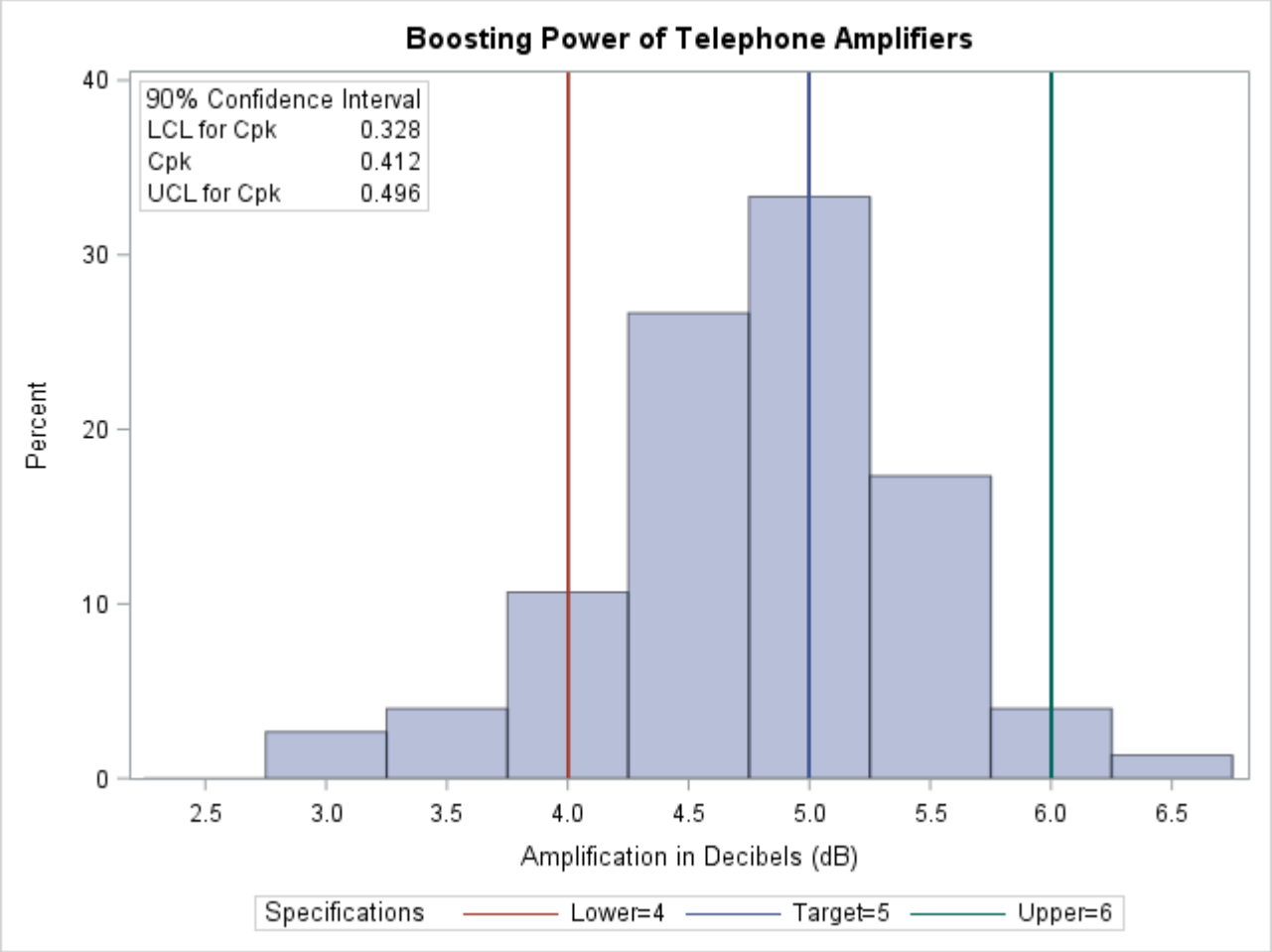
NOTE: See *Displaying a Confidence Interval for Cpm* in the SAS/QC Sample Library.

In this example, the capability index C_{pk} is computed for the amplification data in Amps. To examine the accuracy of this estimate, the following statements calculate a 90% confidence interval for C_{pk}, then display the interval on a histogram (shown in Output 6.3.1) with the INSET statement:

```
title 'Boosting Power of Telephone Amplifiers';
proc capability data=Amps noprint alpha=0.10;
   var Decibels;
   spec  target = 5  lsl = 4  usl = 6
        ltarget = 2 llsl = 3 lusl = 4;
   histogram Decibels / odstitle = title;;
   inset cpklcl cpk cpkucl / header = '90% Confidence Interval'
                             format = 6.3;
run;
```

The ALPHA= option in the PROC CAPABILITY statement controls the level of the confidence interval. In this case, the 90% confidence interval on C_{pk} is wide (from 0.328 to 0.496), indicating that the process may need adjustments in order to improve process variability. Confidence limits for capability indices can be displayed using the INSET statement (as shown in Output 6.3.1) or saved in an output data set by using the OUTPUT statement. For formulas and details about capability indices, see the section "Specialized Capability Indices" on page 239. For more information about the INSET statement, see "INSET Statement: CAPABILITY Procedure" on page 384.

Output 6.3.1 Confidence Interval on C_{pk}



The following statements can be used to produce a table of process capability indices including the index C_{pk}:

```
ods select indices;
proc capability data=Amps alpha=0.10;
   spec  target = 5  lsl = 4  usl = 6
        ltarget = 2 llsl = 3 lusl = 4;
   var Decibels;
run;
```

Output 6.3.2 Process Capability Indices

Boosting Power of Telephone Amplifiers

The CAPABILITY Procedure
Variable: Decibels (Amplification in Decibels (dB))

Process Capability Indices			
Index	**Value**	**90% Confidence Limits**	
Cp	0.508962	0.439538	0.576922
CPL	0.411920	0.326620	0.495136
CPU	0.606004	0.501261	0.708127
Cpk	0.411920	0.327599	0.496241
Cpm	0.488674	0.425292	0.556732

CDFPLOT Statement: CAPABILITY Procedure

Overview: CDFPLOT Statement

The CDFPLOT statement plots the observed cumulative distribution function (*cdf*) of a variable, defined as

$$\begin{aligned} F_N(x) &= \text{percent of nonmissing values} \leq x \\ &= \frac{\text{number of values} \leq x}{N} \times 100\% \end{aligned}$$

where N is the number of nonmissing observations. The cdf is an increasing step function that has a vertical jump of $\frac{1}{N}$ at each value of x equal to an observed value. The cdf is also referred to as the empirical cumulative distribution function (ecdf).

You can use options in the CDFPLOT statement to do the following:

- superimpose specification limits
- superimpose fitted theoretical distributions
- specify graphical enhancements (such as color or text height)

You can also create a comparative cdf plot by using the CDFPLOT statement in conjunction with a CLASS statement.

You have three alternatives for producing cdf plots with the CDFPLOT statement:

- ODS Graphics output is produced if ODS Graphics is enabled, for example by specifying the ODS GRAPHICS ON statement prior to the PROC statement.
- Otherwise, traditional graphics are produced by default if SAS/GRAPH is licensed.
- Legacy line printer charts are produced when you specify the LINEPRINTER option in the PROC statement.

See Chapter 4, "SAS/QC Graphics," for more information about producing these different kinds of graphs.

Getting Started: CDFPLOT Statement

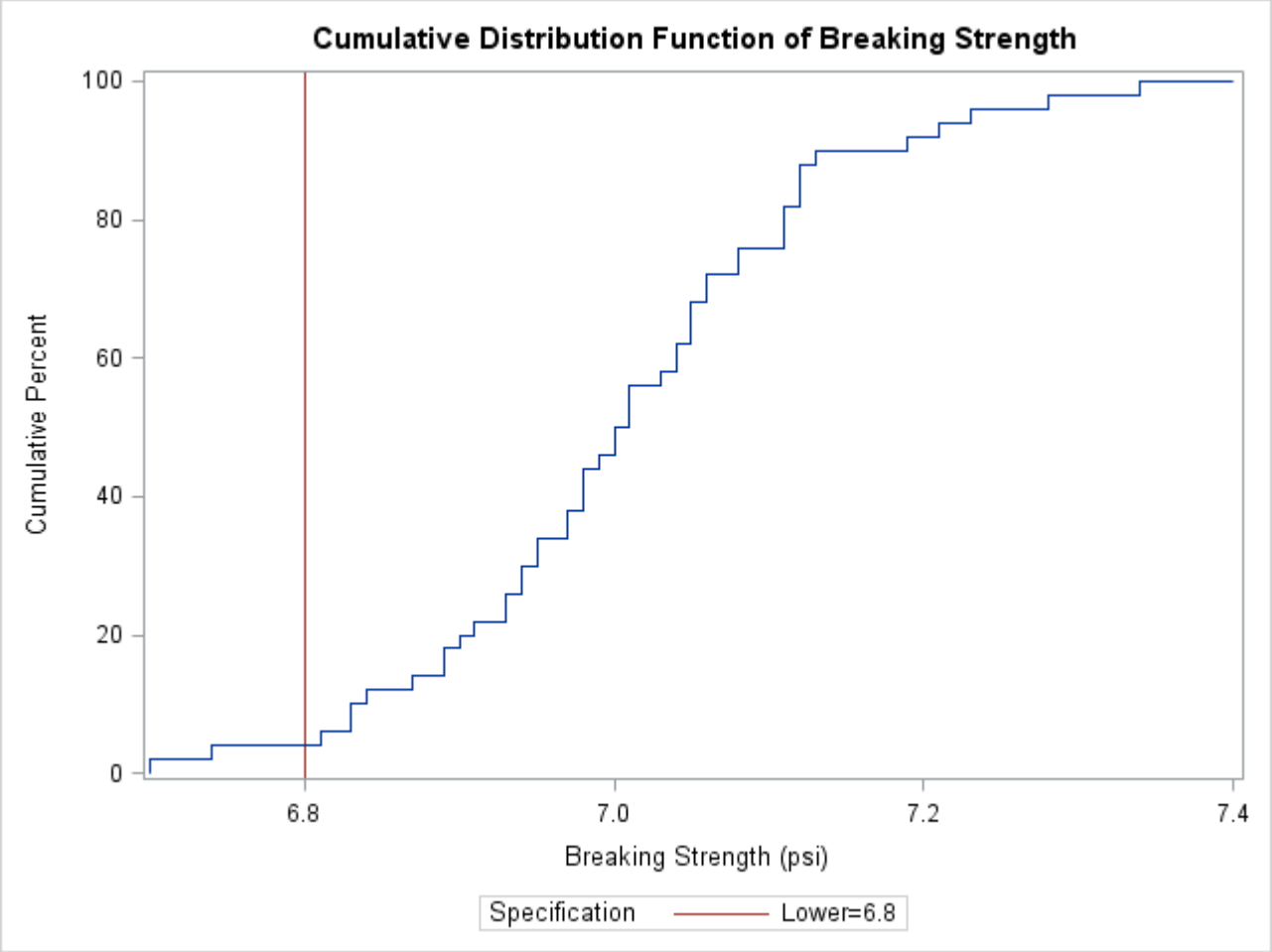
Creating a Cumulative Distribution Plot

NOTE: See *CDF Plot with Superimposed Normal Curve* in the SAS/QC Sample Library.

This section introduces the CDFPLOT statement with a simple example. A company that produces fiber optic cord is interested in the breaking strength of the cord. The following statements create a data set named Cord, which contains 50 breaking strengths measured in pounds per square inch (psi), and they display the cdf plot in Figure 6.5. The plot shows a symmetric distribution with observations concentrated 6.9 and 7.1. The plot also shows that only a small percentage (< 5%) of the observations are below the lower specification limit of 6.8.

```
data Cord;
   label Strength="Breaking Strength (psi)";
   input Strength @@;
   datalines;
6.94 6.97 7.11 6.95 7.12 6.70 7.13 7.34 6.90 6.83
7.06 6.89 7.28 6.93 7.05 7.00 7.04 7.21 7.08 7.01
7.05 7.11 7.03 6.98 7.04 7.08 6.87 6.81 7.11 6.74
6.95 7.05 6.98 6.94 7.06 7.12 7.19 7.12 7.01 6.84
6.91 6.89 7.23 6.98 6.93 6.83 6.99 7.00 6.97 7.01
;

title 'Cumulative Distribution Function of Breaking Strength';
proc capability data=Cord noprint;
   spec lsl=6.8;
   cdf Strength / odstitle=title;
run;
```

Figure 6.5 Cumulative Distribution Function



Syntax: CDFPLOT Statement

The syntax for the CDFPLOT statement is as follows:

CDFPLOT < *variables* > < / *options* > **;**

You can specify the keyword CDF as an alias for CDFPLOT. You can specify any number of CDFPLOT statements after a PROC CAPABILITY statement. The components of the CDFPLOT statement are described as follows:

variables

specify variables for which to create cdf plots. If you specify a VAR statement, the variables must also be listed in the VAR statement. Otherwise, the variables can be any numeric variables in the input data set. If you do not specify variables in a CDFPLOT statement, then a cdf plot is created for each variable listed in the VAR statement, or for each numeric variable in the input data set if you do not use a VAR statement.

For example, suppose a data set named steel contains exactly three numeric variables, length, width and height. The following statements create a cdf plot for each of the three variables:

```
proc capability data=steel;
   cdfplot;
run;
```

The following statements create a cdf plot for length and a cdf plot for width:

```
proc capability data=steel;
   var length width;
   cdfplot;
run;
```

The following statements create a cdf plot for width:

```
proc capability data=steel;
   var length width;
   cdfplot width;
run;
```

By default, the horizontal axis of a cdf plot is labeled with the variable name. If you specify a label for a variable, however, the label is used. The default vertical axis label is *Cumulative Percent*, and the axis is scaled in percent of observations.

If you specify a SPEC statement or a SPEC= data set in addition to the CDFPLOT statement, then the specification limits for each variable are displayed as reference lines and are identified in a legend.

options

add features to plots. All options appear after the slash (/) in the CDFPLOT statement. In the following example, the NORMAL option superimposes a normal cdf on the plot, and the CTEXT= option specifies the color of the text.

```
proc capability data=steel;
   cdfplot length / normal ctext=yellow;
run;
```

Summary of Options

The following tables list all options by function. The section "Dictionary of Options" on page 263 describes each option in detail.

Distribution Options

You can use the options listed in Table 6.9 to superimpose a fitted theoretical distribution function on your cdf plot.

Table 6.9 Options for Specifying a Theoretical Distribution

Option	Description
BETA(*beta-options*)	plots beta distribution with threshold parameter θ, scale parameter σ, and shape parameters α and β
EXPONENTIAL(*exponential-options*)	plots exponential distribution with threshold parameter θ and scale parameter σ
GAMMA(*gamma-options*)	plots gamma distribution with threshold parameter θ, scale parameter σ, and shape parameter α
GUMBEL(*Gumbel-options*)	plots Gumbel distribution with location parameter μ and scale parameter σ
IGAUSS(*iGauss-options*)	plots inverse Gaussian distribution with mean μ and shape parameter λ
LOGNORMAL(*lognormal-options*)	plots lognormal distribution with threshold parameter θ, scale parameter ζ, and shape parameter σ,
NORMAL(*normal-options*)	plots normal distribution with mean μ and standard deviation σ
PARETO(*Pareto-options*)	plots generalized Pareto distribution with threshold parameter θ, scale parameter σ, and shape parameter α
POWER(*power-options*)	plots power function distribution with threshold parameter θ, scale parameter σ, and shape parameter α
RAYLEIGH(*Rayleigh-options*)	plots Rayleigh distribution with threshold parameter θ and scale parameter σ
WEIBULL(*Weibull-options*)	plots Weibull distribution function with threshold parameter θ, scale parameter σ, and shape parameter c

Table 6.10 summarizes options that specify distribution parameters and control the display of the theoretical distribution curve. You can specify these options in parentheses after the distribution option. For example, the following statements use the NORMAL option to superimpose a normal distribution:

```
proc capability;
   cdfplot / normal(mu=10 sigma=0.5 color=red);
run;
```

The COLOR= option specifies the color for the curve, and the *normal-options* MU= and SIGMA= specify the parameters $\mu = 10$ and $\sigma = 0.5$ for the distribution function. If you do not specify these parameters, maximum likelihood estimates are computed.

Table 6.10 Distribution Options

Option	Description
Options Used with All Distributions	
COLOR=	specifies color of theoretical distribution function
L=	specifies line type of theoretical distribution function
SYMBOL=	specifies *character* used to plot theoretical distribution function on line printer plots
W=	specifies width of theoretical distribution function
Beta-Options	
ALPHA=	specifies first shape parameter α for beta distribution function
BETA=	specifies second shape parameter β for beta distribution function
SIGMA=	specifies scale parameter σ for beta distribution function
THETA=	specifies lower threshold parameter θ for beta distribution function
Exponential-Options	
SIGMA=	specifies scale parameter σ for exponential distribution function
THETA=	specifies threshold parameter θ for exponential distribution function
Gamma-Options	
ALPHA=	specifies shape parameter α for gamma distribution function
ALPHADELTA=	specifies change in successive estimates of α at which the Newton-Raphson approximation of $\hat{\alpha}$ terminates
ALPHAINITIAL=	specifies initial value for α in the Newton-Raphson approximation of $\hat{\alpha}$
MAXITER=	specifies maximum number of iterations in the Newton-Raphson approximation of $\hat{\alpha}$
SIGMA=	specifies scale parameter σ for gamma distribution function
THETA=	specifies threshold parameter θ for gamma distribution function
Gumbel-Options	
MU=	specifies location parameter μ for Gumbel distribution function
SIGMA=	specifies scale parameter σ for Gumbel distribution function
IGauss-Options	
LAMBDA=	specifies shape parameter λ for inverse Gaussian distribution function
MU=	specifies mean μ for inverse Gaussian distribution function
Lognormal-Options	
SIGMA=	specifies shape parameter σ for lognormal distribution function
THETA=	specifies threshold parameter θ for lognormal distribution function
ZETA=	specifies scale parameter ζ for lognormal distribution function
Normal-Options	
MU=	specifies mean μ for normal distribution function
SIGMA=	specifies standard deviation σ for normal distribution function
Pareto-Options	
ALPHA=	specifies shape parameter α for generalized Pareto distribution function
SIGMA=	specifies scale parameter σ for generalized Pareto distribution function

Table 6.10 (*continued*)

Option	Description
THETA=	specifies threshold parameter θ for generalized Pareto distribution function
Power-Options	
ALPHA=	specifies shape parameter α for power function distribution
SIGMA=	specifies scale parameter σ for power function distribution
THETA=	specifies threshold parameter θ for power function distribution
Rayleigh-Options	
SIGMA=	specifies scale parameter σ for Rayleigh distribution function
THETA=	specifies threshold parameter θ for Rayleigh distribution function
Weibull-Options	
C=	specifies shape parameter c for Weibull distribution function
CDELTA=	specifies change in successive estimates of c at which the Newton-Raphson approximation of $\hat{c}$ terminates
CINITIAL=	specifies initial value for c in the Newton-Raphson approximation of $\hat{c}$
MAXITER=	specifies maximum number of iterations in the Newton-Raphson approximation of $\hat{c}$
SIGMA=	specifies scale parameter σ for Weibull distribution function
THETA=	specifies threshold parameter θ for Weibull distribution function

General Options

Table 6.11 General CDFPLOT Statement Options

Option	Description
General Plot Layout Options	
CONTENTS=	specifies table of contents entry for cdf plot grouping
HREF=	specifies reference lines perpendicular to the horizontal axis
HREFLABELS=	specifies labels for HREF= lines
NOCDFLEGEND	suppresses legend for superimposed theoretical cdf
NOECDF	suppresses plot of empirical (observed) distribution function
NOFRAME	suppresses frame around plotting area
NOLEGEND	suppresses legend
NOSPECLEGEND	suppresses specifications legend
VREF=	specifies reference lines perpendicular to the vertical axis
VREFLABELS=	specifies labels for VREF= lines
VSCALE=	specifies scale for vertical axis
Graphics Options	
ANNOTATE=	specifies annotate data set
CAXIS=	specifies color for axis
CFRAME=	specifies color for frame
CHREF=	specifies colors for HREF= lines
CSTATREF=	specifies colors for STATREF= lines

Table 6.11 (*continued*)

Option	Description
CTEXT=	specifies color for text
CVREF=	specifies colors for VREF= lines
DESCRIPTION=	specifies description for graphics catalog member
FONT=	specifies text font
HAXIS=	specifies AXIS statement for horizontal axis
HEIGHT=	specifies height of text used outside framed areas
HMINOR=	specifies number of horizontal axis minor tick marks
HREFLABPOS=	specifies position for HREF= line labels
INFONT=	specifies software font for text inside framed areas
INHEIGHT=	specifies height of text inside framed areas
LHREF=	specifies line styles for HREF= lines
LSTATREF=	specifies line styles for STATREF= lines
LVREF=	specifies line styles for VREF= lines
NAME=	specifies name for plot in graphics catalog
NOHLABEL	suppresses label for horizontal axis
NOVLABEL	suppresses label for vertical axis
NOVTICK	suppresses tick marks and tick mark labels for vertical axis
STATREF=	specifies reference lines at values of summary statistics
STATREFLABELS=	specifies labels for STATREF= lines
STATREFSUBCHAR=	specifies substitution character for displaying statistic values in STATREFLABELS= labels
TURNVLABELS	turns and vertically strings out characters in labels for vertical axis
VAXIS=	specifies AXIS statement for vertical axis
VAXISLABEL=	specifies label for vertical axis
VMINOR=	specifies number of vertical axis minor tick marks
VREFLABPOS=	specifies position for VREF= line labels
WAXIS=	specifies line thickness for axes and frame
Options for ODS Graphics Output	
ODSFOOTNOTE=	specifies footnote displayed on cdf plot
ODSFOOTNOTE2=	specifies secondary footnote displayed on cdf plot
ODSTITLE=	specifies title displayed on cdf plot
ODSTITLE2=	specifies secondary title displayed on cdf plot
Options for Comparative Plots	
ANNOKEY	applies annotation requested in ANNOTATE= data set to key cell only
CFRAMESIDE=	specifies color for filling row label frames
CFRAMETOP=	specifies color for filling column label frames
CPROP=	specifies color for proportion of frequency bar
CTEXTSIDE=	specifies color for row labels
CTEXTTOP=	specifies color for column labels
INTERTILE=	specifies distance between tiles in comparative plot
NCOLS=	specifies number of columns in comparative plot
NROWS=	specifies number of rows in comparative plot
OVERLAY	overlays plots for different class levels (ODS Graphics only)

Table 6.11 (*continued*)

Option	Description
Options for Line Printer Charts	
CDFSYMBOL=	specifies character for plotted points
HREFCHAR=	specifies line character for HREF= lines
VREFCHAR=	specifies line character for VREF= lines

Dictionary of Options

The following entries provide detailed descriptions of the options specific to the CDFPLOT statement. See "Dictionary of Common Options: CAPABILITY Procedure" on page 533 for detailed descriptions of options common to all the plot statements.

ALPHA=*value*

specifies the shape parameter α for distribution functions requested with the BETA, GAMMA, PARETO, and POWER options. Enclose the ALPHA= option in parentheses after the distribution keyword. If you do not specify a value for α, the procedure calculates a maximum likelihood estimate. For examples, see the entries for the distribution options.

BETA<(*beta-options***)>**

displays a fitted beta distribution function on the cdf plot. The equation of the fitted cdf is

$$F(x) = \begin{cases} 0 & \text{for } x \le \theta \\ I_{\frac{x-\theta}{\sigma}}(\alpha, \beta) & \text{for } \theta < x < \theta + \sigma \\ 1 & \text{for } x \ge \sigma + \theta \end{cases}$$

where $I_y(\alpha, \beta)$ is the incomplete beta function, and

θ = lower threshold parameter (lower endpoint)
σ = scale parameter ($\sigma > 0$)
α = shape parameter ($\alpha > 0$)
β = shape parameter ($\beta > 0$)

The beta distribution is bounded below by the parameter θ and above by the value $\theta + \sigma$. You can specify θ and σ by using the THETA= and SIGMA= *beta-options*, as illustrated in the following statements, which fit a beta distribution bounded between 50 and 75. The default values for θ and σ are 0 and 1, respectively.

```
proc capability;
   cdfplot / beta(theta=50 sigma=25);
run;
```

The beta distribution has two shape parameters, α and β. If these parameters are known, you can specify their values with the ALPHA= and BETA= *beta-options*. If you do not specify values for α and β, the procedure calculates maximum likelihood estimates.

The BETA option can appear only once in a CDFPLOT statement. See Table 6.10 for a list of secondary options you can specify with the BETA distribution option.

BETA=*value*

B=*value*

specifies the second shape parameter β for beta distribution functions requested by the BETA option. Enclose the BETA= option in parentheses after the BETA keyword. If you do not specify a value for β, the procedure calculates a maximum likelihood estimate. For examples, see the preceding entry for the BETA option.

C=*value*

specifies the shape parameter c for Weibull distribution functions requested with the WEIBULL option. Enclose the C= option in parentheses after the WEIBULL keyword. If you do not specify a value for c, the procedure calculates a maximum likelihood estimate. You can specify the SHAPE= option as an alias for the C= option.

CDFSYMBOL='*character*'

specifies the character used to plot the points on legacy line printer cdf plots. The default is the plus sign (+). This option is ignored unless you specify the LINEPRINTER option in the PROC CAPABILITY statement. Use the SYMBOL statement to control the plotting symbol in traditional graphics output.

EXPONENTIAL< (*exponential-options***) >**

EXP< (*exponential-options* **) >**

displays a fitted exponential distribution function on the cdf plot. The equation of the fitted cdf is

$$F(x) = \begin{cases} 0 & \text{for } x \le \theta \\ 1 - \exp\left(-\frac{x-\theta}{\sigma}\right) & \text{for } x > \theta \end{cases}$$

where
θ = threshold parameter
σ = scale parameter ($\sigma > 0$)

The parameter θ must be less than or equal to the minimum data value. You can specify θ with the THETA= *exponential-option*. The default value for θ is 0. You can specify σ with the SIGMA= *exponential-option*. By default, a maximum likelihood estimate is computed for σ. For example, the following statements fit an exponential distribution with $\theta = 10$ and a maximum likelihood estimate for σ:

```
proc capability;
   cdfplot / exponential(theta=10 l=2 color=green);
run;
```

The exponential curve is green and has a line type of 2.

The EXPONENTIAL option can appear only once in a CDFPLOT statement. See Table 6.10 for a list of secondary options you can specify with the EXPONENTIAL option.

GAMMA< (*gamma-options***) >**

displays a fitted gamma distribution function on the cdf plot. The equation of the fitted cdf is

$$F(x) = \begin{cases} 0 & \text{for } x \le \theta \\ \frac{1}{\Gamma(\alpha)\sigma} \int_\theta^x \left(\frac{t-\theta}{\sigma}\right)^{\alpha-1} \exp\left(-\frac{t-\theta}{\sigma}\right) dt & \text{for } x > \theta \end{cases}$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)
α = shape parameter ($\alpha > 0$)

The parameter θ for the gamma distribution must be less than the minimum data value. You can specify θ with the THETA= *gamma-option*. The default value for θ is 0. In addition, the gamma distribution has a shape parameter α and a scale parameter σ. You can specify these parameters with the ALPHA= and SIGMA= *gamma-options*. By default, maximum likelihood estimates are computed for α and σ. For example, the following statements fit a gamma distribution function with $\theta = 4$ and maximum likelihood estimates for α and σ:

```
proc capability;
   cdfplot / gamma(theta=4);
run;
```

Note that the maximum likelihood estimate of α is calculated iteratively using the Newton-Raphson approximation. The *gamma-options* ALPHADELTA=, ALPHAINITIAL=, and MAXITER= control the approximation.

The GAMMA option can appear only once in a CDFPLOT statement. See Table 6.10 for a list of secondary options you can specify with the GAMMA option.

GUMBEL< (*Gumbel-options*) >
displays a fitted Gumbel distribution (also known as Type 1 extreme value distribution) function on the cdf plot. The equation of the fitted cdf is

$$F(x) = \exp\left(-e^{-(x-\mu)/\sigma}\right)$$

where

μ = location parameter
σ = scale parameter ($\sigma > 0$)

You can specify known values for μ and σ with the MU= and SIGMA= *Gumbel-options*. By default, maximum likelihood estimates are computed for μ and σ.

The GUMBEL option can appear only once in a CDFPLOT statement. See Table 6.10 for a list of secondary options you can specify with the GUMBEL option.

IGAUSS< (*iGauss-options*) >
displays a fitted inverse Gaussian distribution function on the cdf plot. The equation of the fitted cdf is

$$F(x) = \Phi\left\{\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\mu} - 1\right)\right\} + e^{2\lambda/\mu}\Phi\left\{-\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\mu} + 1\right)\right\}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function, and

μ = mean parameter ($\mu > 0$)
λ = shape parameter ($\lambda > 0$)

You can specify known values for μ and λ with the MU= and LAMBDA= *iGauss-options*. By default, maximum likelihood estimates are computed for μ and λ.

The IGAUSS option can appear only once in a CDFPLOT statement. See Table 6.10 for a list of secondary options you can specify with the IGAUSS option.

LAMBDA=*value*

specifies the shape parameter λ for distribution functions requested with the IGAUSS option. Enclose the LAMBDA= option in parentheses after the IGAUSS distribution keyword. If you do not specify a value for λ, the procedure calculates a maximum likelihood estimate.

LEGEND=*name* | **NONE**

specifies the name of a LEGEND statement describing the legend for specification limit reference lines and superimposed distribution functions. Specifying LEGEND=NONE, which suppresses all legend information, is equivalent to specifying the NOLEGEND option. This option is ignored unless you are producing traditional graphics.

LOGNORMAL< (*lognormal-options***) >**

displays a fitted lognormal distribution function on the cdf plot. The equation of the fitted cdf is

$$F(x) = \begin{cases} 0 & \text{for } x \le \theta \\ \Phi\left(\frac{\log(x-\theta)-\zeta}{\sigma}\right) & \text{for } x > \theta \end{cases}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function, and

θ = threshold parameter
ζ = scale parameter
σ = shape parameter ($\sigma > 0$)

The parameter θ for the lognormal distribution must be less than the minimum data value. You can specify θ with the THETA= *lognormal-option*. The default value for θ is 0. In addition, the lognormal distribution has a shape parameter σ and a scale parameter ζ. You can specify these parameters with the SIGMA= and ZETA= *lognormal-options*. By default, maximum likelihood estimates are computed for σ and ζ. For example, the following statements fit a lognormal distribution function with $\theta = 10$ and maximum likelihood estimates for σ and ζ:

```
proc capability;
   cdfplot / lognormal(theta = 10);
run;
```

The LOGNORMAL option can appear only once in a CDFPLOT statement. See Table 6.10 for a list of secondary options you can specify with the LOGNORMAL option.

MU=*value*

specifies the parameter μ for distribution functions requested with the GUMBEL, IGAUSS, and NORMAL options. Enclose the MU= option in parentheses after the distribution keyword. For the normal and inverse Gaussian distributions, the default value of μ is the sample mean. If you do not specify a value for μ for the Gumbel distribution, the procedure calculates a maximum likelihood estimate.

NOCDFLEGEND
suppresses the legend for the superimposed theoretical cumulative distribution function.

NOECDF
suppresses the observed distribution function (the empirical cumulative distribution function) of the variable, which is drawn by default. This option enables you to create theoretical cdf plots without displaying the data distribution. The NOECDF option can be used only with a theoretical distribution (such as the NORMAL option).

NOLEGEND
suppresses legends for specification limits, theoretical distribution functions, and hidden observations. Specifying the NOLEGEND option is equivalent to specifying LEGEND=NONE.

NORMAL< (*normal-options*) >
displays a fitted normal distribution function on the cdf plot. The equation of the fitted cdf is

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \quad \text{for } -\infty < x < \infty$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function, and

μ = mean
σ = standard deviation ($\sigma > 0$)

You can specify known values for μ and σ with the MU= and SIGMA= *normal-options*, as shown in the following statements:

```
proc capability;
   cdfplot / normal(mu=14 sigma=.05);
run;
```

By default, the sample mean and sample standard deviation are calculated for μ and σ. The NORMAL option can appear only once in a CDFPLOT statement. For an example, see Output 6.4.1. See Table 6.10 for a list of secondary options you can specify with the NORMAL option.

NOSPECLEGEND
NOSPECL
suppresses the portion of the legend for specification limit reference lines.

PARETO< (*Pareto-options*) >
displays a fitted generalized Pareto distribution function on the cdf plot. The equation of the fitted cdf is

$$F(x) = 1 - \left(1 - \frac{\alpha(x-\theta)}{\sigma}\right)^{\frac{1}{\alpha}}$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)
α = shape parameter

The parameter θ for the generalized Pareto distribution must be less than the minimum data value. You can specify θ with the THETA= *Pareto-option*. The default value for θ is 0. In addition, the generalized Pareto distribution has a shape parameter α and a scale parameter σ. You can specify these parameters with the ALPHA= and SIGMA= *Pareto-options*. By default, maximum likelihood estimates are computed for α and σ.

The PARETO option can appear only once in a CDFPLOT statement. See Table 6.10 for a list of secondary options you can specify with the PARETO option.

POWER< (*power-options*) >
displays a fitted power function distribution on the cdf plot. The equation of the fitted cdf is

$$F(x) = \begin{cases} 0 & \text{for } x \leq \theta \\ \left(\frac{x-\theta}{\sigma}\right)^{\alpha} & \text{for } \theta < x < \theta + \sigma \\ 1 & \text{for } x \geq \theta + \sigma \end{cases}$$

where

θ = lower threshold parameter (lower endpoint)
σ = scale parameter ($\sigma > 0$)
α = shape parameter ($\alpha > 0$)

The power function distribution is bounded below by the parameter θ and above by the value $\theta + \sigma$. You can specify θ and σ by using the THETA= and SIGMA= *power-options*. The default values for θ and σ are 0 and 1, respectively.

You can specify a value for the shape parameter, α, with the ALPHA= *power-option*. If you do not specify a value for α, the procedure calculates a maximum likelihood estimate.

The power function distribution is a special case of the beta distribution with its second shape parameter, $\beta = 1$.

The POWER option can appear only once in a CDFPLOT statement. See Table 6.10 for a list of secondary options you can specify with the POWER option.

RAYLEIGH< (*Rayleigh-options*) >
displays a fitted Rayleigh distribution function on the cdf plot. The equation of the fitted cdf is

$$F(x) = 1 - e^{-(x-\theta)^2/(2\sigma^2)}$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)

The parameter θ for the Rayleigh distribution must be less than the minimum data value. You can specify θ with the THETA= *Rayleigh-option*. The default value for θ is 0. You can specify σ with the SIGMA= *Rayleigh-option*. By default, a maximum likelihood estimate is computed for σ.

The RAYLEIGH option can appear only once in a CDFPLOT statement. See Table 6.10 for a list of secondary options you can specify with the RAYLEIGH option.

SIGMA=*value*

specifies the parameter σ for distribution functions requested by the BETA, EXPONENTIAL, GAMMA, GUMBEL, LOGNORMAL, NORMAL, PARETO, POWER, RAYLEIGH, and WEIBULL options. Enclose the SIGMA= option in parentheses after the distribution keyword. The following table summarizes the use of the SIGMA= option:

Distribution Option	SIGMA= Specifies	Default Value	Alias
BETA POWER	scale parameter σ	1	SCALE=
EXPONENTIAL GAMMA WEIBULL	scale parameter σ	maximum likelihood estimate	SCALE=
GUMBEL PARETO RAYLEIGH	scale parameter σ	maximum likelihood estimate	
LOGNORMAL	shape parameter σ	maximum likelihood estimate	SHAPE=
NORMAL	scale parameter σ	standard deviation	

SYMBOL='*character***'**

specifies the *character* used to plot the theoretical distribution function on legacy line printer plots. Enclose the SYMBOL= option in parentheses after the distribution option. The default character is the first letter of the distribution option keyword. This option is ignored unless you specify the LINEPRINTER option in the PROC CAPABILITY statement.

THETA=*value*

THRESHOLD=*value*

specifies the lower threshold parameter θ for theoretical cumulative distribution functions requested with the BETA, EXPONENTIAL, GAMMA, LOGNORMAL, PARETO, POWER, RAYLEIGH, and WEIBULL options. Enclose the THETA= option in parentheses after the distribution keyword. The default *value* is 0.

VSCALE=PERCENT | PROPORTION

specifies the scale of the vertical axis. 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.

WEIBULL< (*Weibull-options***) >**

displays a fitted Weibull distribution function on the cdf plot. The equation of the fitted cdf is

$$F(x) = \begin{cases} 0 & \text{for } x \le \theta \\ 1 - \exp\left(-\left(\frac{x-\theta}{\sigma}\right)^c\right) & \text{for } x > \theta \end{cases}$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)
c = shape parameter ($c > 0$)

The parameter θ must be less than the minimum data value. You can specify θ with the THETA= *Weibull-option*. The default value for θ is 0. In addition, the Weibull distribution has a shape parameter c and a scale parameter σ. You can specify these parameters with the SIGMA= and C= *Weibull-options*. By default, maximum likelihood estimates are computed for c and σ. For example, the following statements fit a Weibull distribution function with $\theta = 15$ and maximum likelihood estimates for σ and c:

```
proc capability;
   cdfplot / weibull(theta=15);
run;
```

Note that the maximum likelihood estimate of c is calculated iteratively using the Newton-Raphson approximation. The *Weibull-options* CDELTA=, CINITIAL=, and MAXITER= control the approximation.

The WEIBULL option can appear only once in a CDFPLOT statement. See Table 6.10 for a list of secondary options you can specify with the WEIBULL option.

ZETA=*value*

specifies a value for the scale parameter ζ for a lognormal distribution function requested with the LOGNORMAL option. Enclose the ZETA= option in parentheses after the LOGNORMAL keyword. If you do not specify a *value* for ζ, a maximum likelihood estimate is computed. You can specify the SCALE= option as an alias for the ZETA= option.

Details: CDFPLOT Statement

ODS Graphics

Before you create ODS Graphics output, ODS Graphics must be enabled (for example, by using the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" (Chapter 21, *SAS/STAT User's Guide*).

The appearance of a graph produced with ODS Graphics is determined by the style associated with the ODS destination where the graph is produced. CDFPLOT options used to control the appearance of traditional graphics are ignored for ODS Graphics output.

When ODS Graphics is in effect, the CDFPLOT statement assigns a name to the graph it creates. You can use this name to reference the graph when using ODS. The name is listed in Table 6.12.

Table 6.12 ODS Graphics Produced by the CDFPLOT Statement

ODS Graph Name	Plot Description
CDFPlot	cumulative distribution function plot

See Chapter 4, "SAS/QC Graphics," for more information about ODS Graphics and other methods for producing charts.

Examples: CDFPLOT Statement

This section illustrates how to display a fitted distribution function, inset tables, and display reference lines on your cdf plot.

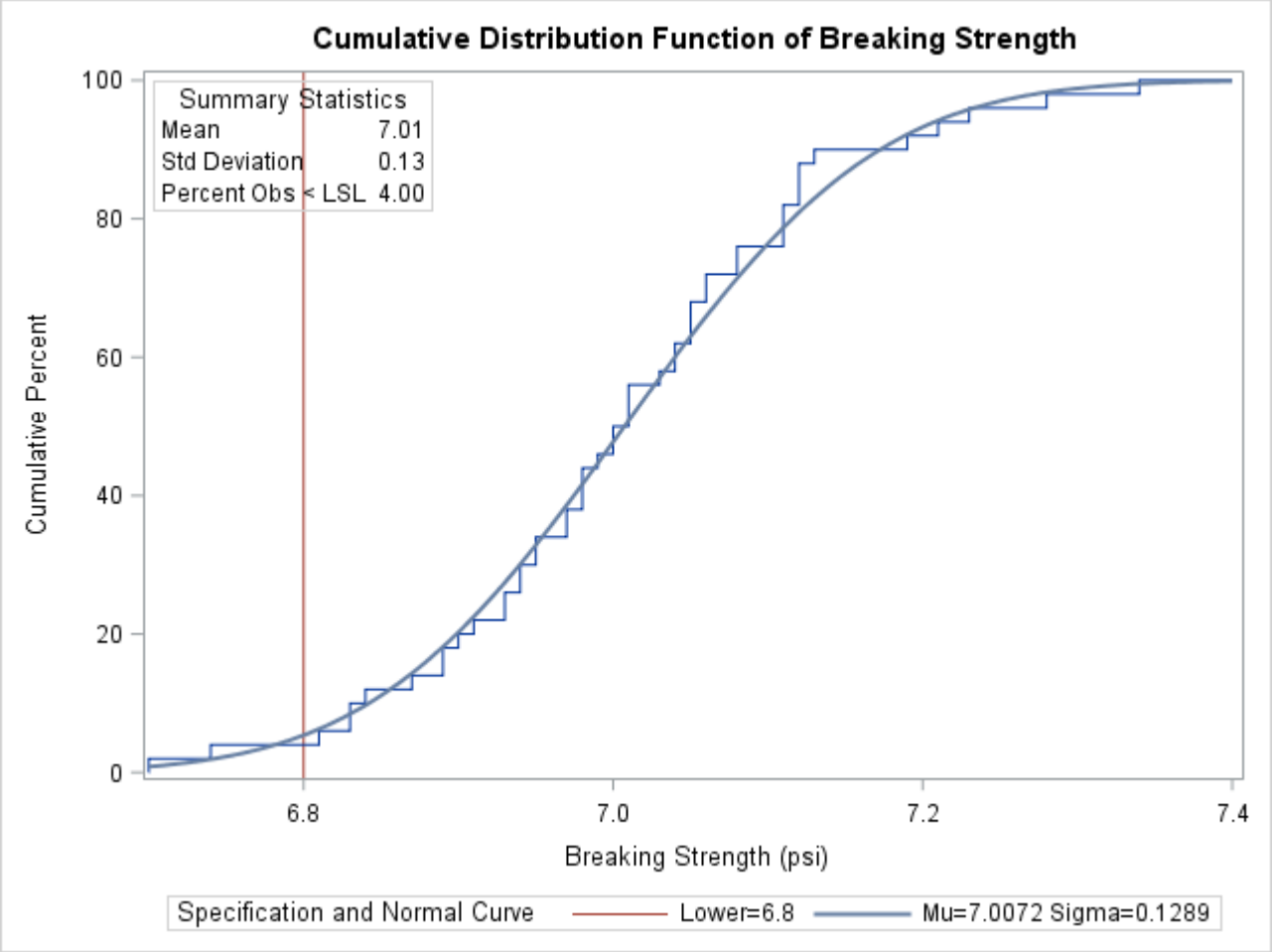
Example 6.4: Fitting a Normal Distribution

NOTE: See *CDF Plot with Superimposed Normal Curve* in the SAS/QC Sample Library.

You can use the CDFPLOT statement to fit any of eleven theoretical distributions (beta, exponential, gamma, Gumbel, inverse Gaussian, lognormal, normal, generalized Pareto, power function, Rayleigh, and Weibull) and superimpose them on the cdf plot. The following statements use the NORMAL option to display a fitted normal distribution function on a cdf plot of breaking strengths. The data set Cord is given in Figure 6.5, and the plot is shown in Output 6.4.1.

```
title 'Cumulative Distribution Function of Breaking Strength';
proc capability data=Cord noprint;
   spec lsl=6.8;
   cdf Strength / normal
                  odstitle=title;
   inset mean std pctlss / format = 5.2
                           header = "Summary Statistics";
run;
```

Output 6.4.1 Superimposed Normal Distribution Function



The NORMAL option requests the fitted curve. The INSET statement requests an inset containing the mean, the standard deviation, and the percent of observations below the lower specification limit. For more information about the INSET statement, see "INSET Statement: CAPABILITY Procedure" on page 384. The SPEC statement requests a lower specification limit at 6.8. For more information about the SPEC statement, see "SPEC Statement" on page 214.

The agreement between the empirical and the normal distribution functions in Output 6.4.1 is evidence that the normal distribution is an appropriate model for the distribution of breaking strengths.

The CAPABILITY procedure provides a variety of other tools for assessing goodness of fit. Goodness-of-fit tests (see "Printed Output" on page 348) provide a quantitative assessment of a proposed distribution. Probability and Q-Q plots, created with the PROBPLOT ("PROBPLOT Statement: CAPABILITY Procedure" on page 460), QQPLOT ("QQPLOT Statement: CAPABILITY Procedure" on page 492), and PPPLOT ("PPPLOT Statement: CAPABILITY Procedure" on page 438) statements, provide effective graphical diagnostics.

Example 6.5: Using Reference Lines with CDF Plots

NOTE: See *CDF Plot with Superimposed Normal Curve* in the SAS/QC Sample Library.

Customer requirements dictate that the breaking strengths in the previous example have upper and lower specification limits of 7.2 and 6.8 psi, respectively. Moreover, less than 5% of the cords can have breaking strengths outside the limits.

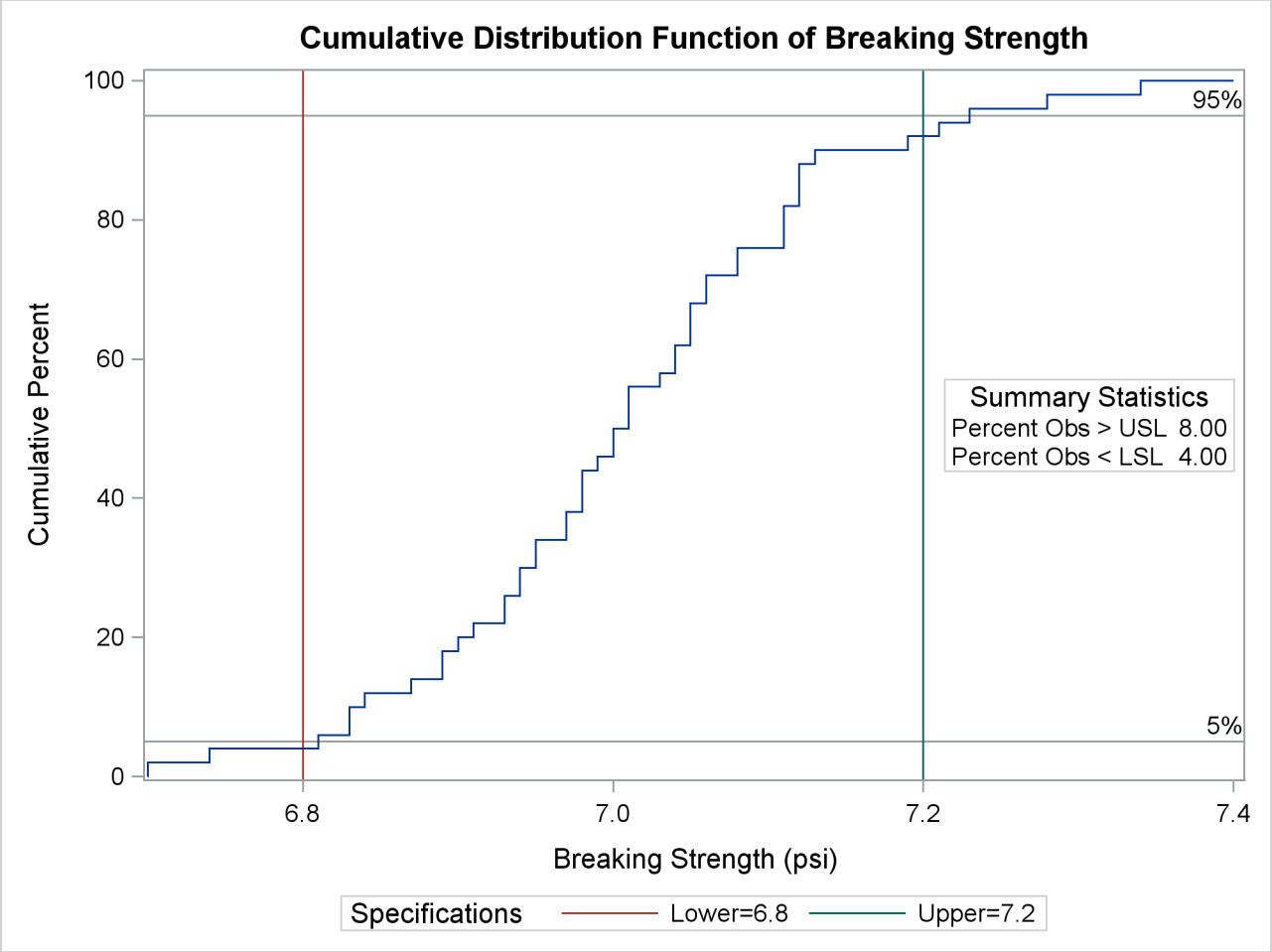
The following statements create a cdf plot with reference lines at the 5% and 95% cumulative percent levels:

```
proc capability data=Cord noprint;
   spec lsl=6.8 usl=7.2;
   cdf Strength / vref         = 5 95
                  vreflabels   = '5%' '95%'
                  odstitle     = title;
   inset pctgtr pctlss / format = 5.2
                         pos    = e
                         header = "Summary Statistics";
run;
```

The INSET statement requests an inset with the percentages of measurements above the upper limit and below the lower limit. For more information about the INSET statement, see "INSET Statement: CAPABILITY Procedure" on page 384.

In Output 6.5.1, the empirical cdf is below the intersection between the lower specification limit line and the 5% line, so less than 5% of the measurements are below the lower limit. The ecdf, however, is *also* below the intersection between the upper specification limit line and the 95% line, implying that *more* than 5% of the measurements are greater than the upper limit. Thus, the goal of having less than 5% of the measurements above the upper specification limit has not been met.

Output 6.5.1 Reference Lines with a Cumulative Distribution Function Plot



COMPHISTOGRAM Statement: CAPABILITY Procedure

Overview: COMPHISTOGRAM Statement

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 or endpoints 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

You have two alternatives for producing comparative histograms with the COMPHISTOGRAM statement:

- ODS Graphics output is produced if ODS Graphics is enabled, for example by specifying the ODS GRAPHICS ON statement prior to the PROC statement.
- Otherwise, traditional graphics are produced if SAS/GRAPH is licensed.

See Chapter 4, "SAS/QC Graphics," for more information about producing these different kinds of graphs.

NOTE: You cannot use the COMPHISTOGRAM statement together with the CLASS statement.

Getting Started: COMPHISTOGRAM Statement

This section introduces the COMPHISTOGRAM statement with examples that illustrate commonly used options. Complete syntax for the COMPHISTOGRAM statement is presented in the section "Syntax: COMPHISTOGRAM Statement" on page 278, and advanced examples are given in the section "Examples: COMPHISTOGRAM Statement" on page 294.

Creating a One-Way Comparative Histogram

NOTE: See *Comparative Histograms with Normal Curves* in the SAS/QC Sample Library.

The effective channel length (in microns) is measured for 1225 field effect transistors. The channel lengths are saved as values of the variable Length in a SAS data set named Channel:

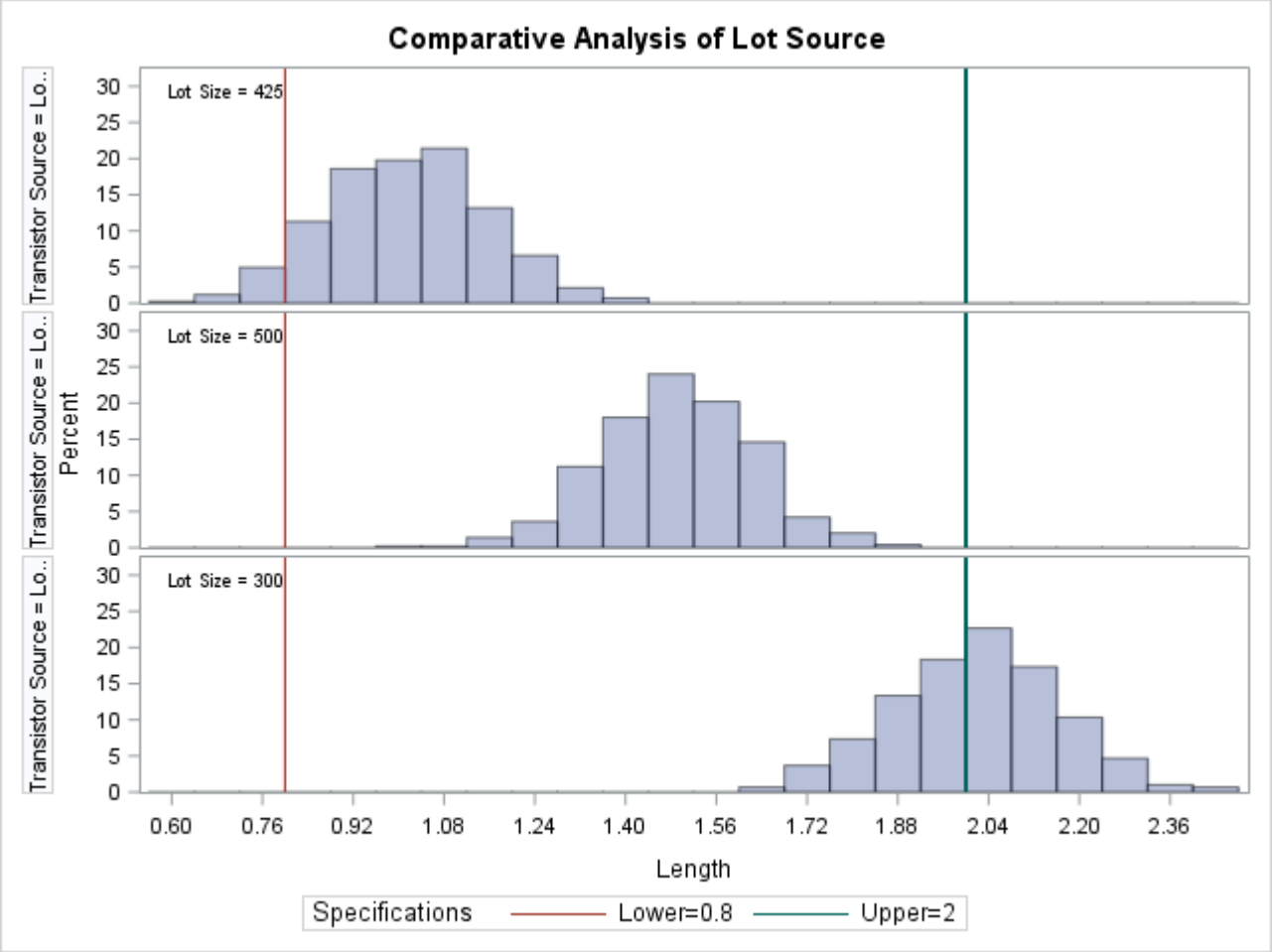
```
data Channel;
   length Lot $ 16;
   input Length @@;
   select;
      when (_n_ <= 425) Lot='Lot 1';
      when (_n_ >= 926) Lot='Lot 3';
      otherwise Lot='Lot 2';
   end;
   datalines;
0.91 1.01 0.95 1.13 1.12 0.86 0.96 1.17 1.36 1.10
0.98 1.27 1.13 0.92 1.15 1.26 1.14 0.88 1.03 1.00
0.98 0.94 1.09 0.92 1.10 0.95 1.05 1.05 1.11 1.15
1.11 0.98 0.78 1.09 0.94 1.05 0.89 1.16 0.88 1.19
1.01 1.08 1.19 0.94 0.92 1.27 0.90 0.88 1.38 1.02

   ... more lines ...

2.13 2.05 1.90 2.07 2.15 1.96 2.15 1.89 2.15 2.04
1.95 1.93 2.22 1.74 1.91
;
```

The data set Channel is also used in Example 6.12, where a kernel density estimate is superimposed on the histogram of channel lengths. The display in Output 6.12.1 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 that uses Lot as a classification variable. The following statements create the comparative histogram shown in Figure 6.6:

```
title "Comparative Analysis of Lot Source";
proc capability data=Channel noprint;
   specs lsl = 0.8 usl = 2.0;
   comphist Length / class      = Lot
                     nrows      = 3
                     nlegend    = 'Lot Size'
                     nlegendpos = nw
                     odstitle   = title;
   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 the section "Dictionary of Options" on page 284 for descriptions of these options, and see the section "SPEC Statement" on page 214 for details of the SPEC statement.

Figure 6.6 Comparison by Lot Source



Adding Fitted Normal Curves to a Comparative Histogram

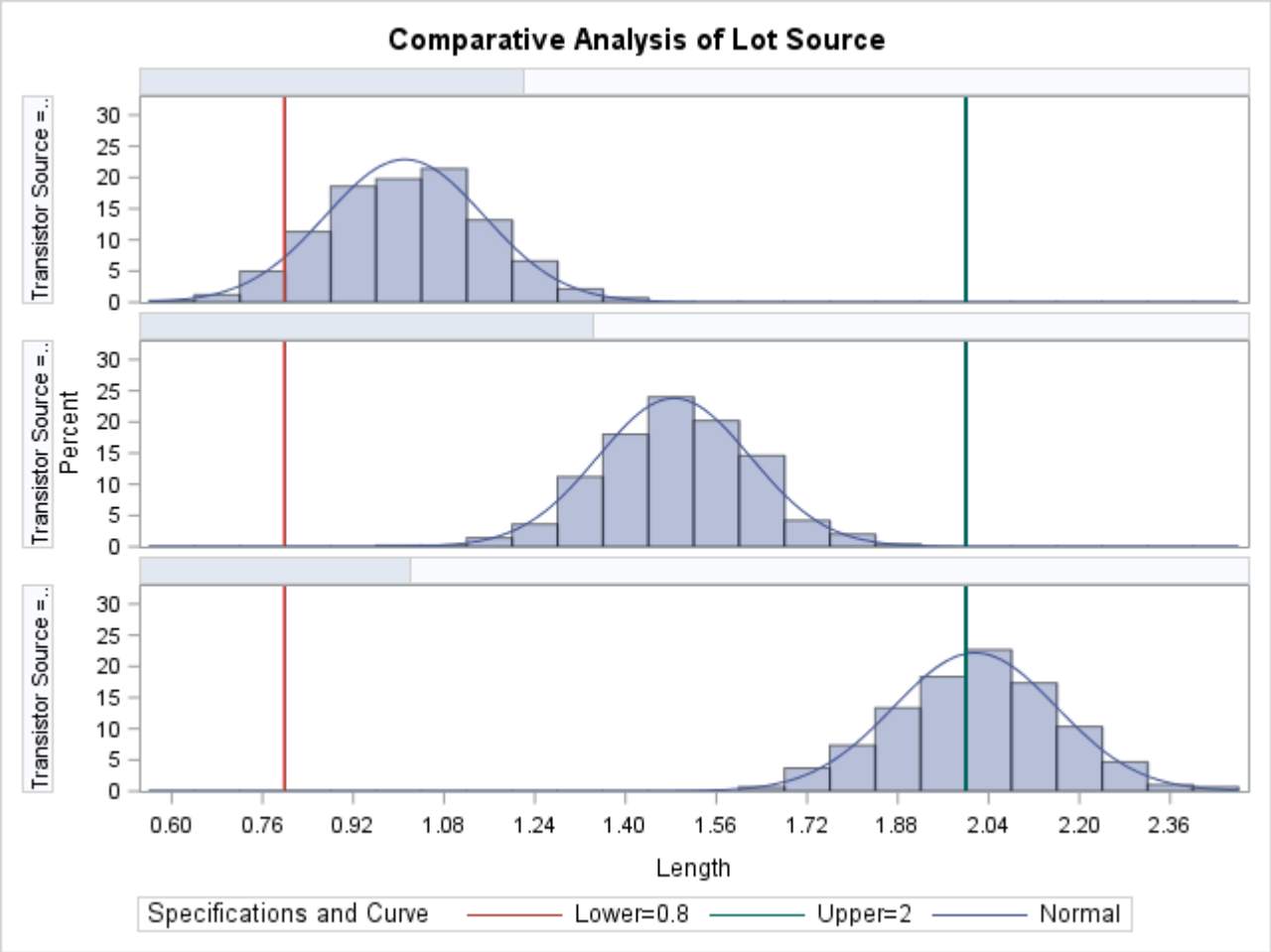
NOTE: See *Comparative Histograms with Normal Curves* in the SAS/QC Sample Library.

In Figure 6.6, 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.

```
title "Comparative Analysis of Lot Source";
proc capability data=Channel noprint;
   specs lsl = 0.8 usl = 2.0;
   comphist Length / class     = Lot
                     nrows     = 3
                     intertile = 1
                     odstitle  = title
                     cprop
                     normal;
   label Lot = 'Transistor Source';
run;
```

The comparative histogram is displayed in Figure 6.7.

Figure 6.7 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 284 for details concerning these options.

Syntax: COMPHISTOGRAM Statement

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 of the values of the *variable* for each level of the *class-variables*. 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):

```
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 by 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 6.6.1 and Output 6.7.1 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 284.

Distribution Options

Table 6.13 lists the options for requesting that a fitted normal distribution or a kernel density estimate be overlaid on the comparative histogram.

Table 6.13 Density Estimation Options

Option	Description
KERNEL(*kernel-options*)	fits kernel density estimates
NORMAL(*normal-options*)	fits normal distribution with mean μ and standard deviation σ

You can specify the secondary options listed in Table 6.14 in parentheses after the KERNEL option to control features of kernel density estimates.

Table 6.14 Kernel-Options

Option	Description
C=	specifies standardized bandwidth parameter c for kernel density estimate
COLOR=	specifies color of the kernel density curve
FILL	fills area under kernel density curve
K=	specifies NORMAL, TRIANGULAR, or QUADRATIC kernel
L=	specifies line type used for kernel density curve
LOWER=	specifies lower bound for kernel density curve
UPPER=	specifies upper bound for kernel density curve
W=	specifies line width for kernel density curve

You can specify the secondary options listed in Table 6.15 in parentheses after the NORMAL option to control features of fitted normal distributions.

Table 6.15 Normal-Options

Option	Description
COLOR=	specifies color of normal curve
FILL	fills area under normal curve
L=	specifies line type of normal curve
MU=	specifies mean μ for fitted normal curve
SIGMA=	specifies standard deviation σ for fitted normal curve
W=	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 because these parameters are not specified.

General Options

Table 6.16 General COMPHISTOGRAM Statement Options

Option	Description
Classification Options	
CLASS=	specifies classification variables
CLASSKEY=	specifies key cell

Table 6.16 (*continued*)

Option	Description
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
ORDER1=	specifies display order for values of the first CLASS= variable
ORDER2=	specifies display order for values of the second CLASS= variable
Layout Options	
BARLABEL=	produces labels above histogram bars
BARWIDTH=	specifies width for the bars
CLIPSPEC=	clips histogram bars at specification limits if there are no observations beyond the limits
ENDPOINTS=	labels interval endpoints and specifies how they are determined
HOFFSET=	specifies offset for horizontal axis
INTERTILE=	specifies distance between tiles
MAXNBIN=	specifies maximum number of bins displayed
MAXSIGMAS=	limits number of bins displayed to range of *value* standard deviations above and below mean of data in key cell
MIDPOINTS=	specifies how midpoints are determined
NCOLS=	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=	specifies number of rows in comparative histogram
RTINCLUDE	includes right endpoint in interval
WBARLINE=	specifies line thickness for bar outlines
Axis and Legend Options	
GRID	adds grid corresponding to vertical axis
LGRID=	specifies line style for grid requested with GRID option
NLEGEND	specifies form of the legend displayed inside tiles
NLEGENDPOS=	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
TILELEGLABEL=	specifies label displayed when _CTILE_ and _TILELG_ variables are provided in the CLASSSPEC= data set
TURNVLABELS	turns and strings out vertically characters in vertical axis labels
VAXIS=	specifies tick mark values for vertical axis
VAXISLABEL=	specifies label for vertical axis
VOFFSET=	specifies length of offset at upper end of vertical axis
VSCALE=	specifies scale for vertical axis
WAXIS=	specifies line thickness for axes and frame
WGRID=	specifies line thickness for grid

Table 6.16 (*continued*)

Option	Description
Reference Line Options	
FRONTREF	draws reference lines in front of histogram bars
HREF=	specifies reference lines perpendicular to horizontal axis
HREFLABELS=	specifies labels for HREF= lines
HREFLABPOS=	specifies vertical position of labels for HREF= lines
LHREF=	specifies line style for HREF= lines
LVREF=	specifies line style for VREF= lines
VREF=	specifies reference lines perpendicular to vertical axis
VREFLABELS=	specifies labels for VREF= lines
VREFLABPOS=	specifies horizontal position of labels for VREF= lines
Text Enhancement Options	
FONT=	specifies software font for text
HEIGHT=	specifies height of text used outside framed areas
INFONT=	specifies software font for text inside framed areas
INHEIGHT=	specifies height of text inside framed areas
Color and Pattern Options	
CAXIS=	specifies color for axis
CBARLINE=	specifies color for outline of the bars
CFILL=	specifies color for filling bars
CFRAME=	specifies color for frame
CFRAMENLEG=	specifies the color for the frame requested by the NLEGEND option
CFRAMESIDE=	specifies color for filling frame for row labels
CFRAMETOP=	specifies color for filling frame for column labels
CGRID=	specifies color for grid lines
CHREF=	specifies color for HREF= lines
CPROP=	specifies color for proportion of frequency bar
CTEXT=	specifies color for text
CTEXTSIDE=	specifies color for row labels
CTEXTTOP=	specifies color for column labels
CVREF=	specifies color for VREF= lines
PFILL=	specifies pattern used to fill bars
Input and Output Data Set Options	
ANNOKEY	applies annotation requested in ANNOTATE= data set to key cell only
ANNOTATE=	annotate data set
CLASSSPEC=	data set with specification limit information for each cell
OUTHISTOGRAM=	information on histogram intervals
Graphics Catalog Options	
DESCRIPTION=	specifies description for graphics catalog member
NAME=	specifies name for plot in graphics catalog

Dictionary of Options

The following sections describe in detail the options specific to the COMPHISTOGRAM statement. See "Dictionary of Common Options: CAPABILITY Procedure" on page 533 for detailed descriptions of options common to all the plot statements.

General Options

You can specify the following options whether you are producing ODS Graphics output or traditional graphics:

BARLABEL=COUNT | PERCENT | PROPORTION

displays labels above the histogram bars. If you specify BARLABEL=COUNT, the label shows the number of observations associated with a given bar. BARLABEL=PERCENT shows the percent of observations represented by that bar. If you specify BARLABEL=PROPORTION, the label displays the proportion of observations associated with the bar.

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 347):

```
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.

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 6.6.

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 6.7.1.

Classification variables can be numeric or character. 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 6.6.

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*. 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 6.7.1. 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*

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 6.6 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

ENDPOINTS=*value-list* **| KEY | UNIFORM**

specifies that histogram interval endpoints, rather than midpoints, are aligned with horizontal axis tick marks, and specifies how the endpoints are determined. The method you specify is used for all process variables analyzed with the COMPHISTOGRAM statement.

If you specify ENDPOINTS=*value-list*, the *values* must be listed in increasing order and must be evenly spaced. The difference between consecutive endpoints is used as the width of the histogram bars. The first value is the lower bound of the first histogram bin and the last value is the upper bound of the last bin. Thus, the number of values in the list is one greater than the number of bins it specifies. 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.

If you specify ENDPOINTS=KEY, the procedure first determines the endpoints for the data in the key cell. The initial number of endpoints is based on the number of observations in the key cell by using the method of Terrell and Scott (1985). The endpoint list for the key cell is then extended in either direction as necessary until it spans the data in the remaining cells. If the key cell contains no observations, the method of determining bins reverts to ENDPOINTS=UNIFORM.

If you specify ENDPOINTS=UNIFORM, the procedure determines the endpoints by using all the observations as if there were no cells. In other words, the number of endpoints is computed from the total sample size by using the method of Terrell and Scott (1985).

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	COMPHIST	CFILL=*color*
limits	COMPHIST	PFILL=*pattern*
left of lower	SPEC	CLEFT=*color*
specification limit	SPEC	PLEFT=*pattern*
right of upper	SPEC	CRIGHT=*color*
specification limit	SPEC	PRIGHT=*pattern*

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 6.6.1. Refer to *SAS/GRAPH: Help* 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.

GRID

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

INTERTILE=*value*

specifies the distance in horizontal percent screen units between tiles. For an example, see Figure 6.7. 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:

Option	Description
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
LOWER=	specifies the lower bound for the curve
UPPER=	specifies the upper bound for the curve

See Output 6.6.1 for an example. By default, the estimate is based on the AMISE method. For more information, see "Kernel Density Estimates" on page 347.

LOWER=*value*

specifies the lower bound for a kernel density estimate curve. Enclose the LOWER= option in parentheses after the KERNEL option. You can specify a single lower bound or a list of lower bounds. By default, a kernel density estimate curve has no lower bound.

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 6.6 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 by 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 computed from the total sample size by 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.

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.

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 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

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

where

μ = mean
σ = standard deviation ($\sigma > 0$)
h = width of histogram interval
v = vertical scaling factor
and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

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 6.7 for another illustration. Table 6.15 lists options that can be specified in parentheses after the NORMAL option.

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 or endpoints 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. By default, interval midpoint values are saved in the variable _MIDPT_. If the ENDPOINTS= option is specified, intervals are identified by endpoint values instead. If RTINCLUDE is specified, the _MAXPT_ variable contains upper endpoint values. Otherwise, lower endpoint values are saved in the _MINPT_ variable.

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.

UPPER=*value*
specifies the upper bound for a kernel density estimate curve. Enclose the UPPER= option in parentheses after the KERNEL option. You can specify a single upper bound or a list of upper bounds. By default, a kernel density estimate curve has no upper bound.

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.

Options for Traditional Graphics

You can specify the following options if you are producing traditional graphics:

BARWIDTH=*value*
specifies the width of the histogram bars in screen percent units.

CBARLINE=*color*
specifies the color of the outline of the histogram bars. This option overrides the C= option in the SYMBOL1 statement.

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 6.6.1 and Example 6.7 for examples. Refer to *SAS/GRAPH: Help* for a list of colors. By default, bars and curve areas are not filled.

CFRAMENLEG=*color* | **EMPTY**
CFRAMENLEG
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.

CGRID=*color*
specifies the color for grid lines requested with the GRID option. By default, grid lines are the same color as the axes. If you use CGRID=, you do not need to specify the GRID option.

CLIPSPEC=CLIP | **NOFILL**
specifies that histogram bars are clipped at the upper and lower specification limit lines when there are no observations outside the specification limits. The bar intersecting the lower specification limit is clipped if there are no observations less than the lower limit; the bar intersecting the upper specification limit is clipped if there are no observations greater than the upper limit. If you specify CLIPSPEC=CLIP, the histogram bar is truncated at the specification limit. If you specify CLIPSPEC=NOFILL, the portion of a filled histogram bar outside the specification limit is left unfilled. Specifying CLIPSPEC=NOFILL when histogram bars are not filled has no effect.

FRONTREF
draws reference lines requested with the HREF= and VREF= options in front of the histogram bars. By default, reference lines are drawn behind the histogram bars and can be obscured by them.

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

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.

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 6.6 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 6.6 for an illustration. The default is NE.

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: Help* for a list of pattern values. By default, the bars and curve areas are not filled.

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:*.

VOFFSET=*value*
specifies the offset in percent screen units at the upper end of the vertical axis.

WBARLINE=*n*
specifies the width of bar outlines. By default, $n = 1$.

WGRID=*n*
specifies the width of the grid lines requested with the GRID option. By default, grid lines are the same width as the axes. If you use the WGRID= option, you do not need to specify the GRID option.

Details: COMPHISTOGRAM Statement

ODS Graphics

Before you create ODS Graphics output, ODS Graphics must be enabled (for example, by using the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" (Chapter 21, *SAS/STAT User's Guide*).

The appearance of a graph produced with ODS Graphics is determined by the style associated with the ODS destination where the graph is produced. COMPHISTOGRAM options used to control the appearance of traditional graphics are ignored for ODS Graphics output.

When ODS Graphics is in effect, the COMPHISTOGRAM statement assigns a name to the graph it creates. You can use this name to reference the graph when using ODS. The name is listed in Table 6.17.

Table 6.17 ODS Graphics Produced by the COMPHISTOGRAM Statement

ODS Graph Name	Plot Description
Histogram	comparative histogram

See Chapter 4, "SAS/QC Graphics," for more information about ODS Graphics and other methods for producing charts.

Examples: COMPHISTOGRAM Statement

This section provides advanced examples of comparative histograms.

Example 6.6: Adding Insets with Descriptive Statistics

NOTE: See *Machine Study with Comparative Histogram* in the SAS/QC Sample Library.

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';
   datalines;
-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
 0.01  -0.51   0.07  -0.16  -0.32  -0.42  -0.42  -0.34  -0.34  -0.35
-0.49   0.11  -0.42   0.76   0.02  -0.59  -0.28   1.12  -0.02  -0.60
-0.64   0.13  -0.32  -0.77  -0.02  -0.07  -0.49  -0.53  -0.22   0.61
-0.23   0.02   0.53   0.23  -0.44  -0.05   0.37  -0.42   0.70  -0.35

   ... more lines ...

 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.

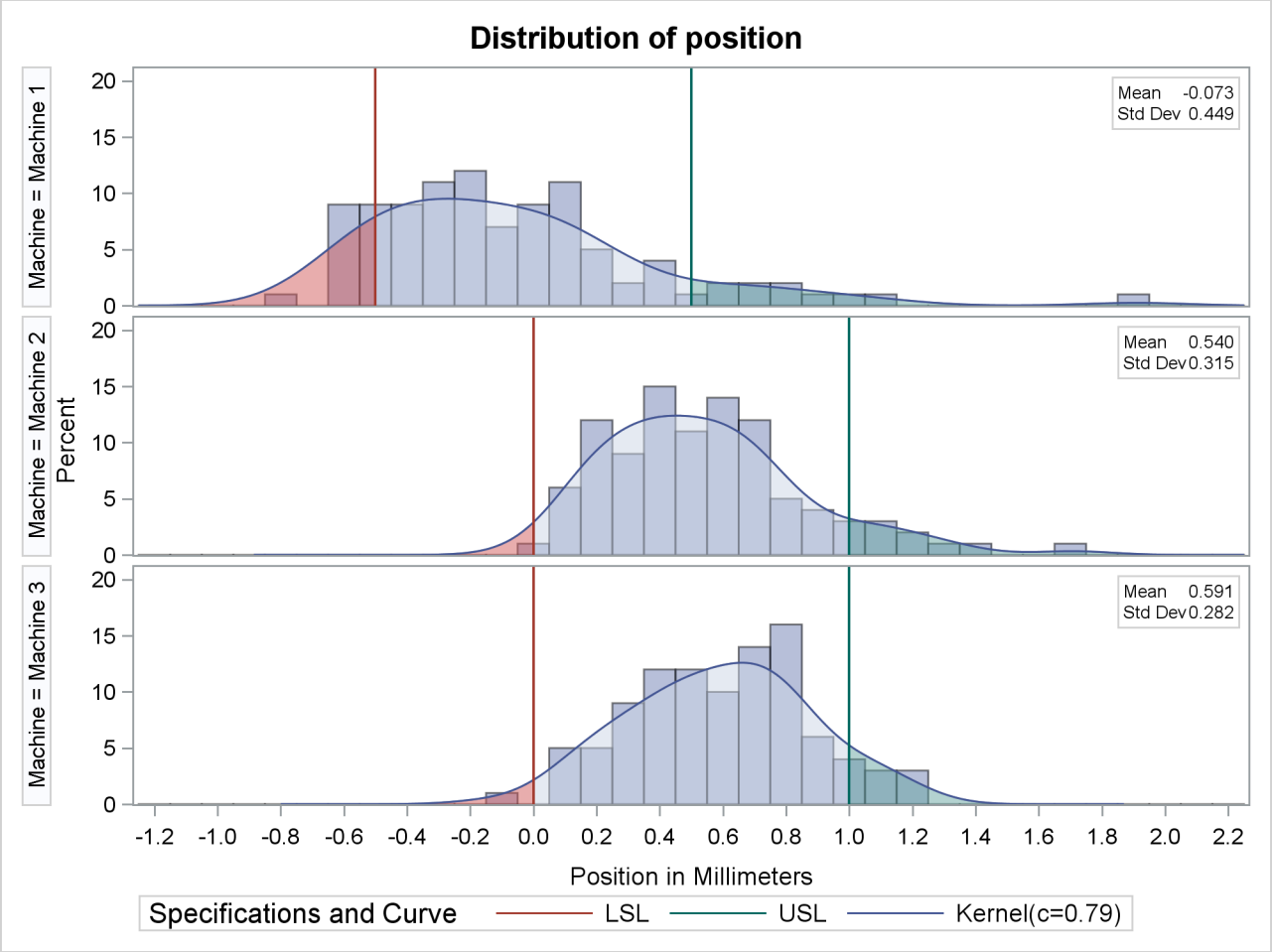
```
data speclims;
   input Machine $9. _lsl_ _usl_;
   _var_ = 'position';
   datalines;
Machine 1 -0.5 0.5
Machine 2  0.0 1.0
Machine 3  0.0 1.0
;
```

The following statements create a comparative histogram for the measurements in Machines that displays the specification limits in speclims.

```
proc capability data=Machines noprint;
   spec cleft cright;
   comphist position / class      = Machine
                       nrows      = 3
                       intertile  = 1
                       midpoints  = -1.2 to 2.2 by 0.1
                       kernel(fill)
                       classspecs = speclims;
   inset mean std="Std Dev" / pos = ne format = 6.3;
run;
```

The display is shown in Output 6.6.1.

Output 6.6.1 Comparative Histogram



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 6.7: Creating a Two-Way Comparative Histogram

NOTE: See *Two-Way Comparative Histogram* in the SAS/QC Sample Library.

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.

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.

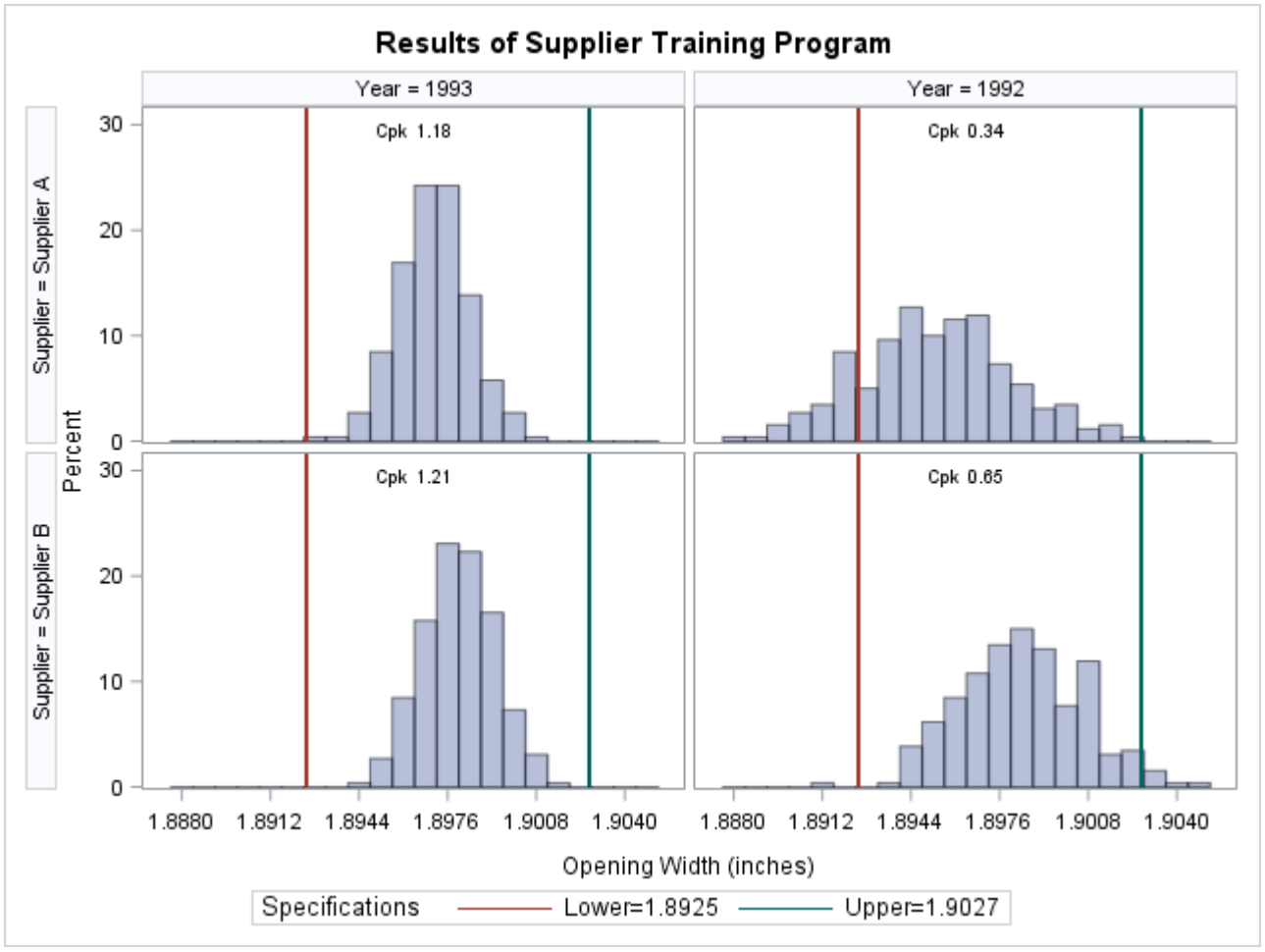
```
proc format ;
   value mytime  1 = '1992'
                 2 = '1993' ;

data disk;
   input @1 supplier $10. year width;
   label width = 'Opening Width (inches)';
   format year mytime.;
   datalines;
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 6.7.1:

```
* Define a format for time periods;
proc format ;
   value mytime
      1 = '1992'
      2 = '1993'
      3 = 'Target for 1994'
   ;
* Simulate the data;
data Disk;
   keep Supplier Year Width;
   length Supplier $ 16;
   label  Width = 'Opening Width (inches)';
   format Year mytime. ;
   Year = 1;
   Supplier = 'Supplier A';
   avg      = 1.895 ;
   std      = 0.0027 ;
   do i = 1 to 260;
      Width = avg + std * rannor(15535); output;
   end;
   Supplier = 'Supplier B';
   avg      = 1.8983 ;
   std      = 0.0024 ;
   do i = 1 to 260;
      Width = avg + std * rannor(15535); output;
   end;
   Year = 2;
   Supplier = 'Supplier A';
   avg      = 1.8970 ;
   std      = 0.0013 ;
   do i = 1 to 260;
```

```
      Width = avg + std * rannor(15535); output;
   end;
   Supplier = 'Supplier B';
   avg      = 1.8980 ;
   std      = 0.0013 ;
   do i = 1 to 260;
      Width = avg + std * rannor(15535); output;
   end;
run;

title "Results of Supplier Training Program";
proc capability data=Disk noprint;
   specs  lsl = 1.8925
          usl = 1.9027;
   comphist Width / class      = ( Supplier Year )
                    classkey   = ('Supplier A' '1993')
                    intertile  = 1.0
                    vaxis      = 0 10 20 30
                    ncols      = 2
                    nrows      = 2
                    odstitle   = title;
   inset cpk (4.2) / noframe pos = n;
run;
```

Output 6.7.1 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 (because 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 285 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 INSET statement is used to display the capability index C_{pk} for each cell. Output 6.7.1 provides evidence that both suppliers have reduced variability from 1992 to 1993.

HISTOGRAM Statement: CAPABILITY Procedure

Overview: HISTOGRAM Statement

Histograms are typically used in process capability analysis to compare the distribution of measurements from an in-control process with its specification limits. In addition to creating histograms, you can use the HISTOGRAM statement to do the following:

- specify the midpoints or endpoints for histogram intervals
- display specification limits on histograms
- display density curves for fitted theoretical distributions on histograms
- request goodness-of-fit tests for fitted distributions
- display kernel density estimates on histograms
- inset summary statistics and process capability indices on histograms
- save histogram intervals and parameters of fitted distributions in output data sets
- create hanging histograms
- request graphical enhancements
- create comparative histograms by using the HISTOGRAM statement together with a CLASS statement

You have three alternatives for producing histograms with the HISTOGRAM statement:

- ODS Graphics output is produced if ODS Graphics is enabled, for example by specifying the ODS GRAPHICS ON statement prior to the PROC statement.
- Otherwise, traditional graphics are produced by default if SAS/GRAPH is licensed.

- Legacy line printer charts are produced when you specify the LINEPRINTER option in the PROC statement.

See Chapter 4, "SAS/QC Graphics," for more information about producing these different kinds of graphs.

Getting Started: HISTOGRAM Statement

This section introduces the HISTOGRAM statement with examples that illustrate commonly used options. Complete syntax for the HISTOGRAM statement is presented in the section "Syntax: HISTOGRAM Statement" on page 305, and advanced examples are given in the section "Examples: HISTOGRAM Statement" on page 362.

Creating a Histogram with Specification Limits

NOTE: See *Histogram with Fitted Normal Curve* in the SAS/QC Sample Library.

A semiconductor manufacturer produces printed circuit boards that are sampled to determine whether the thickness of their copper plating lies between a lower specification limit of 3.45 mils and an upper specification limit of 3.55 mils. The plating process is assumed to be in statistical control. The plating thicknesses of 100 boards are saved in a data set named Trans, created by the following statements:

```
data Trans;
   input Thick @@;
   label Thick='Plating Thickness (mils)';
   datalines;
3.468 3.428 3.509 3.516 3.461 3.492 3.478 3.556 3.482 3.512
3.490 3.467 3.498 3.519 3.504 3.469 3.497 3.495 3.518 3.523
3.458 3.478 3.443 3.500 3.449 3.525 3.461 3.489 3.514 3.470
3.561 3.506 3.444 3.479 3.524 3.531 3.501 3.495 3.443 3.458
3.481 3.497 3.461 3.513 3.528 3.496 3.533 3.450 3.516 3.476
3.512 3.550 3.441 3.541 3.569 3.531 3.468 3.564 3.522 3.520
3.505 3.523 3.475 3.470 3.457 3.536 3.528 3.477 3.536 3.491
3.510 3.461 3.431 3.502 3.491 3.506 3.439 3.513 3.496 3.539
3.469 3.481 3.515 3.535 3.460 3.575 3.488 3.515 3.484 3.482
3.517 3.483 3.467 3.467 3.502 3.471 3.516 3.474 3.500 3.466
;
```

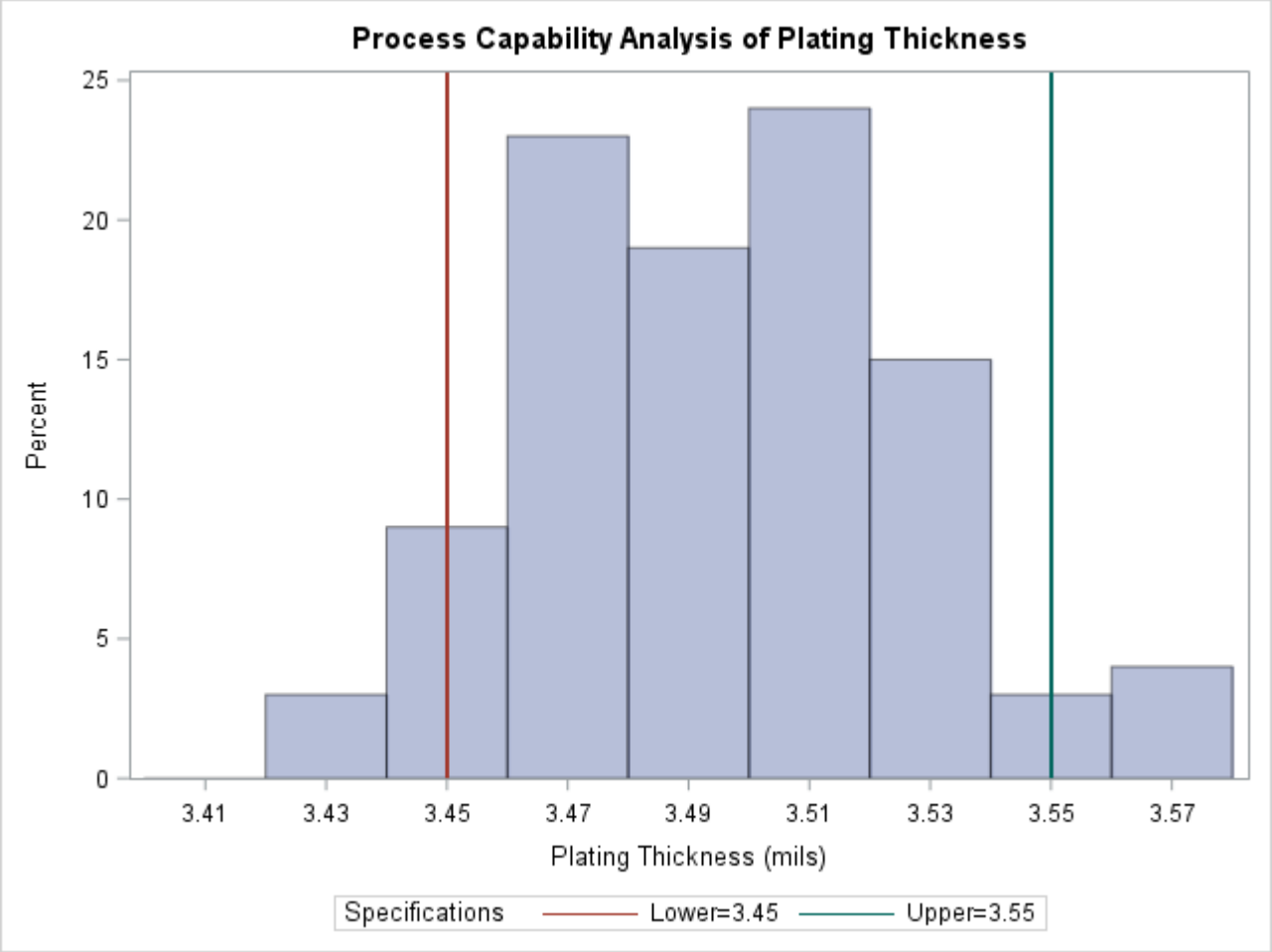
The following statements create the histogram shown in Figure 6.8:

```
title 'Process Capability Analysis of Plating Thickness';
proc capability data=Trans noprint;
   spec lsl = 3.45 usl = 3.55;
   histogram Thick / odstitle = title;
run;
```

A histogram is created for each variable listed after the keyword HISTOGRAM. The SPEC statement, which is optional, provides the specification limits that are displayed on the histogram. For more information about the SPEC statement, see "SPEC Statement" on page 214.

The NOPRINT option suppresses printed output with summary statistics for the variable Thick that would be displayed by default. See "Computing Descriptive Statistics" on page 197 for an example of this output.

Figure 6.8 Histogram Created with Traditional Graphics



Adding a Normal Curve to the Histogram

NOTE: See *Histogram with Fitted Normal Curve* in the SAS/QC Sample Library.

This example is a continuation of the preceding example.

The following statements fit a normal distribution from the thickness measurements and superimpose the fitted density curve on the histogram:

```
proc capability data=Trans;
   spec lsl = 3.45 usl = 3.55;
   histogram / normal;
run;
```

The ODS GRAPHICS ON statement specified before the PROC CAPABILITY statement enables ODS Graphics, so the histogram is created using ODS Graphics instead of traditional graphics.

The NORMAL option summarizes the fitted distribution in the printed output shown in Figure 6.9, and it specifies that the normal curve be displayed on the histogram shown in Figure 6.10.

Figure 6.9 Summary for Fitted Normal Distribution

Process Capability Analysis of Plating Thickness

The CAPABILITY Procedure
Fitted Normal Distribution for Thick (Plating Thickness (mils))

Parameters for Normal Distribution

Parameter	Symbol	Estimate
Mean	Mu	3.49533
Std Dev	Sigma	0.032117

Goodness-of-Fit Tests for Normal Distribution

Test		Statistic	DF		p Value
Kolmogorov-Smirnov	**D**	0.05563823		**Pr > D**	>0.150
Cramer-von Mises	**W-Sq**	0.04307548		**Pr > W-Sq**	>0.250
Anderson-Darling	**A-Sq**	0.27840748		**Pr > A-Sq**	>0.250
Chi-Square	**Chi-Sq**	6.96953022	5	**Pr > Chi-Sq**	0.223

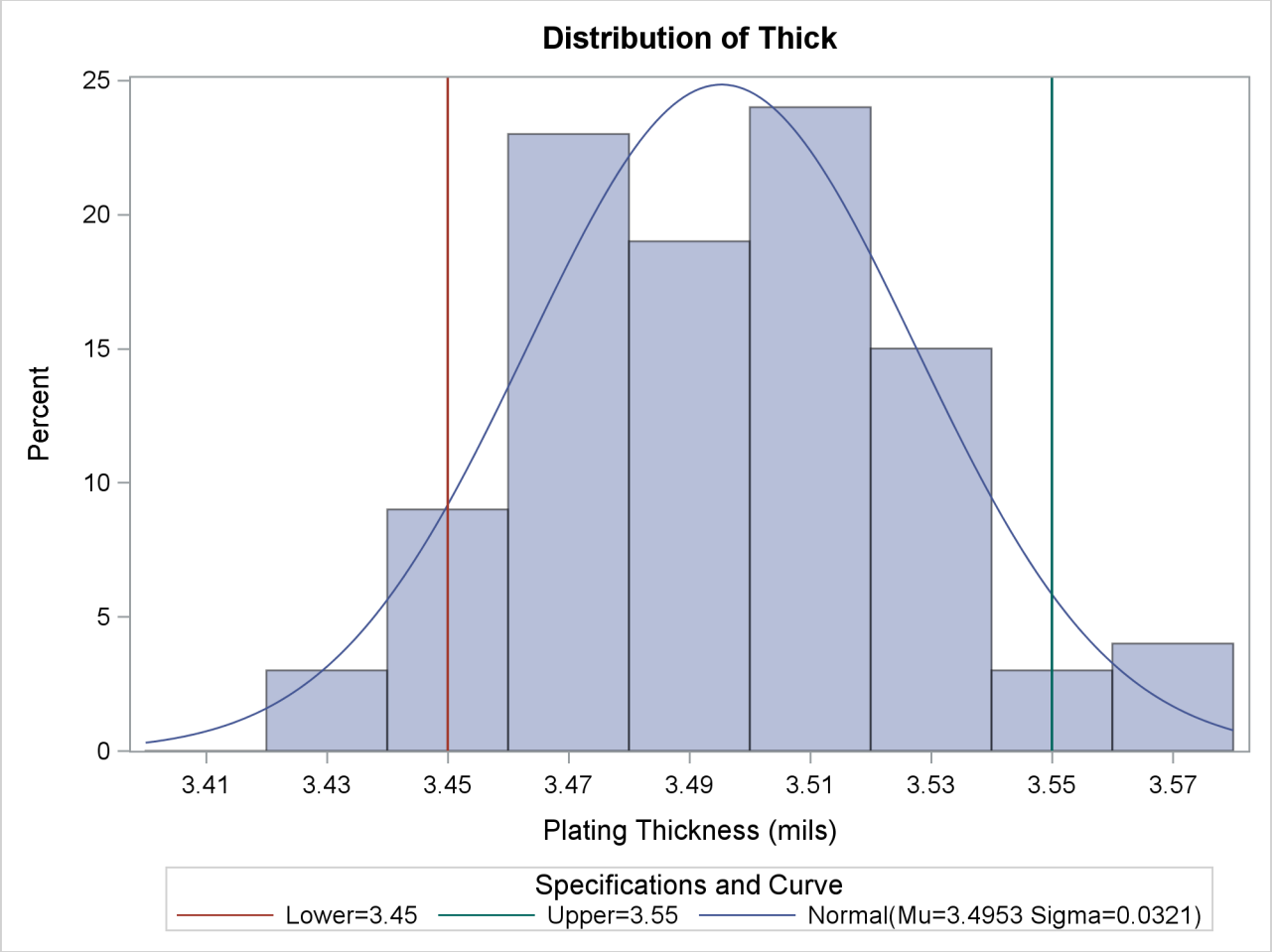
Percent Outside Specifications for Normal Distribution

Lower Limit		Upper Limit	
LSL	3.450000	**USL**	3.550000
Obs Pct < LSL	8.000000	**Obs Pct > USL**	5.000000
Est Pct < LSL	7.906248	**Est Pct > USL**	4.435722

Quantiles for Normal Distribution

	Quantile	
Percent	Observed	Estimated
1.0	3.42950	3.42061
5.0	3.44300	3.44250
10.0	3.45750	3.45417
25.0	3.46950	3.47367
50.0	3.49600	3.49533
75.0	3.51650	3.51699
90.0	3.53550	3.53649
95.0	3.55300	3.54816
99.0	3.57200	3.57005

Figure 6.10 Histogram Superimposed with Normal Curve



The printed output includes the following:

- parameters for the normal curve. The normal parameters μ and σ are estimated by the sample mean ($\hat{\mu} = 3.49533$) and the sample standard deviation ($\hat{\sigma} = 0.032117$).
- goodness-of-fit tests based on the empirical distribution function (EDF): the Anderson-Darling, Cramer-von Mises, and Kolmogorov-Smirnov tests. The p-values for these tests are greater than the usual cutoff values of 0.05 and 0.10, indicating that the thicknesses are normally distributed.
- a chi-square goodness-of-fit test. The p-value of 0.223 for this test indicates that the thicknesses are normally distributed. In general EDF tests (when available) are preferable to chi-square tests. See the section "EDF Goodness-of-Fit Tests" on page 350 for details.
- observed and estimated percentages outside the specification limits
- observed and estimated quantiles

For details, including formulas for the goodness-of-fit tests, see "Printed Output" on page 348. Note that the NOPRINT option in the PROC CAPABILITY statement suppresses only the printed output with summary statistics for the variable Thick. To suppress the printed output in Figure 6.9, specify the NOPRINT option enclosed in parentheses after the NORMAL option as in "Customizing a Histogram" on page 304.

The NORMAL option is one of many options that you can specify in the HISTOGRAM statement. See the section "Syntax: HISTOGRAM Statement" on page 305 for a complete list of options or the section "Dictionary of Options" on page 314 for detailed descriptions of options.

Customizing a Histogram

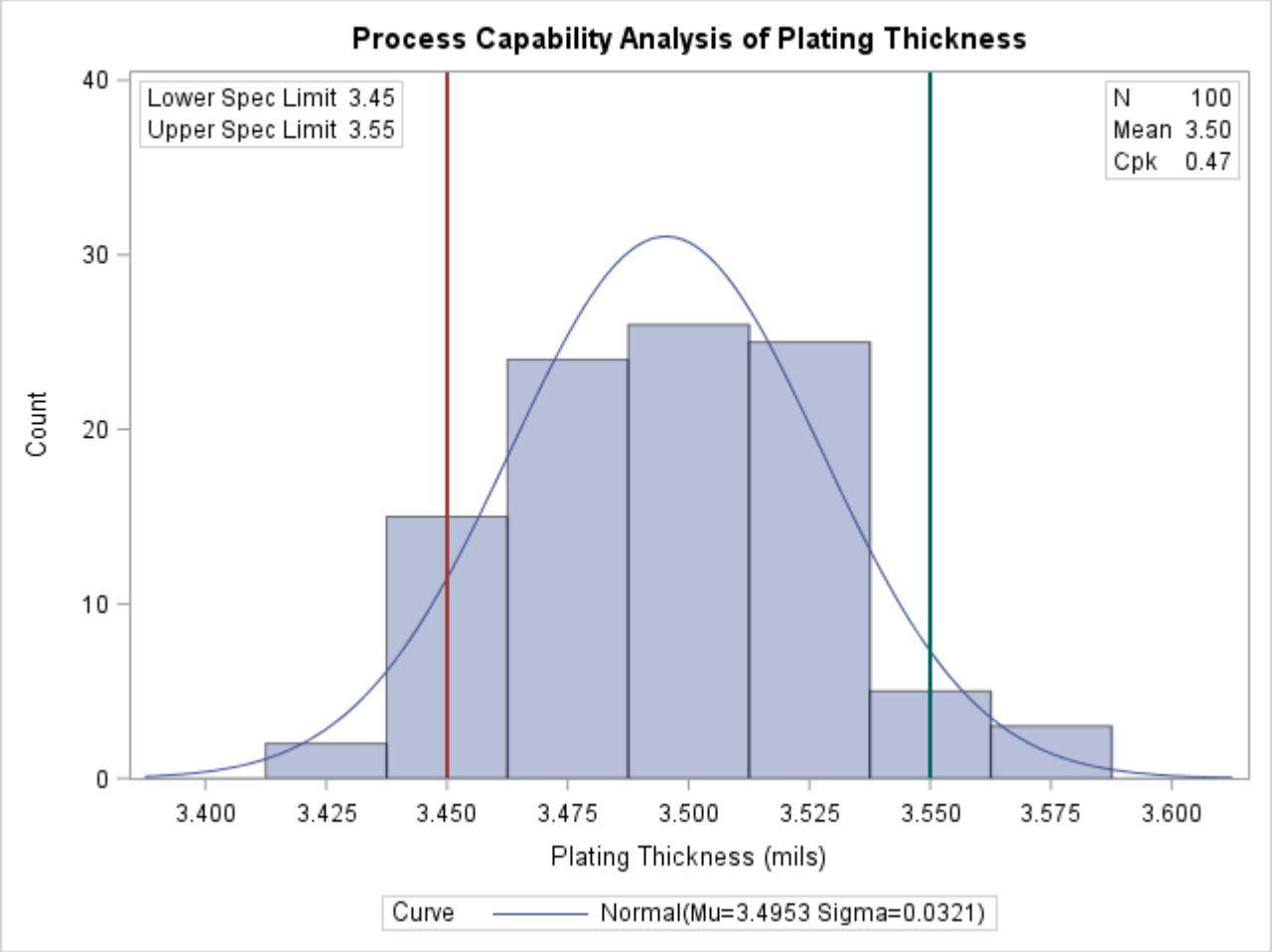
NOTE: See *Histogram with Fitted Normal Curve* in the SAS/QC Sample Library.

This example is a continuation of the preceding example. The following statements show how you can use HISTOGRAM statement options and INSET statements to customize a histogram:

```
title 'Process Capability Analysis of Plating Thickness';
proc capability data=Trans noprint;
   spec lsl = 3.45 usl = 3.55;
   histogram Thick / normal
                     midpoints = 3.4 to 3.6 by 0.025
                     vscale    = count
                     odstitle  = title
                     nospeclegend;
   inset lsl usl;
   inset n mean (5.2) cpk (5.2);
run;
```

The histogram is displayed in Figure 6.11.

Figure 6.11 Customizing the Appearance of the Histogram



The MIDPOINTS= option specifies a list of values to use as bin midpoints. The VSCALE=COUNT option requests a vertical axis scaled in counts rather than percents. The INSET statements inset the specification limits and summary statistics. The NOSPECLEGEND option suppress the default legend for the specification limits that is shown in Figure 6.8.

For more information about HISTOGRAM statement options, see the section "Dictionary of Options" on page 314. For details on the INSET statement, see "INSET Statement: CAPABILITY Procedure" on page 384.

Syntax: HISTOGRAM Statement

The syntax for the HISTOGRAM statement is as follows:

HISTOGRAM < *variables* > < / *options* > **;**

You can specify the keyword HIST as an alias for HISTOGRAM. You can use any number of HISTOGRAM statements after a PROC CAPABILITY statement. The components of the HISTOGRAM statement are described as follows.

variables

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

For example, suppose a data set named steel contains exactly two numeric variables named length and width. The following statements create two histograms, one for length and one for width:

```
proc capability data=steel;
   histogram;
run;
```

The following statements also create histograms for length and width:

```
proc capability data=steel;
   var length width;
   histogram;
run;
```

The following statements create a histogram for length only:

```
proc capability data=steel;
   var length width;
   histogram length;
run;
```

options

add features to the histogram. Specify all options after the slash (/) in the HISTOGRAM statement.

For example, in the following statements, the NORMAL option displays a fitted normal curve on the histogram, the MIDPOINTS= option specifies midpoints for the histogram, and the CTEXT= option specifies the color of the text:

```
proc capability data=steel;
   histogram length / normal
                      midpoints = 5.6 5.8 6.0 6.2 6.4
                      ctext     = yellow;
run;
```

Summary of Options

The following tables list the HISTOGRAM statement options by function. For detailed descriptions, see "Dictionary of Options" on page 314.

Parametric Density Estimation Options

Table 6.18 lists options that display a parametric density estimate on the histogram.

Table 6.18 Parametric Distribution Options

Option	Description
BETA(*beta-options*)	fits beta distribution with threshold parameter θ, scale parameter σ, and shape parameters α and β
EXPONENTIAL(*exponential-options*)	fits exponential distribution with threshold parameter θ and scale parameter σ
GAMMA(*gamma-options*)	fits gamma distribution with threshold parameter θ, scale parameter σ, and shape parameter α
GUMBEL(*Gumbel-options*)	plots Gumbel distribution with location parameter μ and scale parameter σ
IGAUSS(*iGauss-options*)	plots inverse Gaussian distribution with mean μ and shape parameter λ
LOGNORMAL(*lognormal-options*)	fits lognormal distribution with threshold parameter θ, scale parameter ζ, and shape parameter σ
NORMAL(*normal-options*)	fits normal distribution with mean μ and standard deviation σ
PARETO(*Pareto-options*)	plots Pareto distribution with threshold parameter θ, scale parameter σ, and shape parameter α
POWER(*power-options*)	plots power function distribution with threshold parameter θ, scale parameter σ, and shape parameter α
RAYLEIGH(*Rayleigh-options*)	plots Rayleigh distribution with threshold parameter θ and scale parameter σ
SB(*SB-options*)	fits Johnson S_B distribution with threshold parameter θ, scale parameter σ, and shape parameters δ and γ
SU(*SU-options*)	fits Johnson S_U distribution with location parameter θ, scale parameter σ, and shape parameters δ and γ
WEIBULL(*Weibull-options*)	fits Weibull distribution with threshold parameter θ, scale parameter σ, and shape parameter c

Table 6.19 lists secondary options that specify parameters for fitted parametric distributions and that control the display of fitted curves. Specify these secondary options in parentheses after the distribution keyword. For example, the following statements fit a normal curve by using the NORMAL option:

```
proc capability;
   histogram / normal(color=red mu=10 sigma=0.5);
run;
```

The COLOR= *normal-option* draws the curve in red, and the MU= and SIGMA= *normal-options* specify the parameters $\mu = 10$ and $\sigma = 0.5$ for the curve. Note that the sample mean and sample standard deviation are used to estimate μ and σ, respectively, when the MU= and SIGMA= options are not specified.

You can specify lists of values for distribution parameters to display more than one fitted curve from the same distribution family on a histogram. Option values are matched by list position. You can specify the value *EST* in a list of distribution parameter values to use an estimate of the parameter.

For example, the following code displays two normal curves on a histogram:

```
proc capability;
   histogram / normal(color=(red blue) mu=10 est sigma=0.5 est);
run;
```

The first curve is red, with $\mu = 10$ and $\sigma = 0.5$. The second curve is blue, with μ equal to the sample mean and σ equal to the sample standard deviation.

See the section "Formulas for Fitted Curves" on page 336 for detailed information about the families of parametric distributions that you can fit with the HISTOGRAM statement.

Table 6.19 Distribution Options

Option	Description
Options Used with All Parametric Distributions	
COLOR=	specifies color of fitted density curve
FILL	fills area under fitted density curve
INDICES	calculates capability indices based on fitted distribution
L=	specifies line type of fitted curve
MIDPERCENTS	prints table of midpoints of histogram intervals
NOPRINT	suppresses printed output summarizing fitted curve
PERCENTS=	lists percents for which quantiles calculated from data and quantiles estimated from fitted curve are tabulated
SYMBOL=	specifies character used for fitted density curve in line printer plots
W=	specifies width of fitted density curve
Beta-Options	
ALPHA=	specifies first shape parameter α for fitted beta curve
BETA=	specifies second shape parameter β for fitted beta curve
SIGMA=	specifies scale parameter σ for fitted beta curve
THETA=	specifies lower threshold parameter θ for fitted beta curve
Exponential-Options	
SIGMA=	specifies scale parameter σ for fitted exponential curve
THETA=	specifies threshold parameter θ for fitted exponential curve
Gamma-Options	
ALPHA=	specifies shape parameter α for fitted gamma curve
ALPHADELTA=	specifies change in successive estimates of α at which the Newton-Raphson approximation of $\hat{\alpha}$ terminates

Table 6.19 (*continued*)

Option	Description
ALPHAINITIAL=	specifies initial value for α in Newton-Raphson approximation of $\hat{\alpha}$
MAXITER=	specifies maximum number of iterations in Newton-Raphson approximation of $\hat{\alpha}$
SIGMA=	specifies scale parameter σ for fitted gamma curve
THETA=	specifies threshold parameter θ for fitted gamma curve
Gumbel-Options	
EDFNSAMPLES=	specifies number of samples for EDF goodness-of-fit simulation
EDFSEED=	specifies seed value for EDF goodness-of-fit simulation
MU=	specifies location parameter μ for fitted Gumbel curve
SIGMA=	specifies scale parameter σ for fitted Gumbel curve
IGauss-Options	
EDFNSAMPLES=	specifies number of samples for EDF goodness-of-fit simulation
EDFSEED=	specifies seed value for EDF goodness-of-fit simulation
LAMBDA=	specifies shape parameter λ for fitted inverse Gaussian curve
MU=	specifies mean μ for fitted inverse Gaussian curve
Lognormal-Options	
SIGMA=	specifies shape parameter σ for fitted lognormal curve
THETA=	specifies threshold parameter θ for fitted lognormal curve
ZETA=	specifies scale parameter ζ for fitted lognormal curve
Normal-Options	
MU=	specifies mean μ for fitted normal curve
SIGMA=	specifies standard deviation σ for fitted normal curve
Pareto-Options	
ALPHA=	specifies shape parameter α for fitted Pareto curve
EDFNSAMPLES=	specifies number of samples for EDF goodness-of-fit simulation
EDFSEED=	specifies seed value for EDF goodness-of-fit simulation
SIGMA=	specifies scale parameter σ for fitted Pareto curve
THETA=	specifies threshold parameter θ for fitted Pareto curve
Power-Options	
ALPHA=	specifies shape parameter α for fitted power function curve
SIGMA=	specifies scale parameter σ for fitted power function curve
THETA=	specifies threshold parameter θ for fitted power function curve
Rayleigh-Options	
EDFNSAMPLES=	specifies number of samples for EDF goodness-of-fit simulation
EDFSEED=	specifies seed value for EDF goodness-of-fit simulation
SIGMA=	specifies scale parameter σ for fitted Rayleigh curve
THETA=	specifies threshold parameter θ for fitted Rayleigh curve
S_B**-Options**	
DELTA=	specifies first shape parameter δ for fitted S_B curve
FITINTERVAL=	specifies z-value for method of percentiles
FITMETHOD=	specifies method of parameter estimation
FITTOLERANCE=	specifies tolerance for method of percentiles
GAMMA=	specifies second shape parameter γ for fitted S_B curve
SIGMA=	specifies scale parameter σ for fitted S_B curve

Table 6.19 (*continued*)

Option	Description
THETA=	specifies lower threshold parameter θ for fitted S_B curve
S_U-Options	
DELTA=	specifies first shape parameter δ for fitted S_U curve
FITINTERVAL=	specifies z-value for method of percentiles
FITMETHOD=	specifies method of parameter estimation
FITTOLERANCE=	specifies tolerance for method of percentiles
GAMMA=	specifies second shape parameter γ for fitted S_U curve
OPTBOUNDRANGE=	specifies the sampling range for parameter starting values in MLE optimization
OPTMAXITER=	specifies an iteration limit for MLE optimization
OPTMAXSTARTS=	specifies the maximum number of starting points to be used for MLE optimization
OPTPRINT	prints an iteration history for MLE optimization
OPTSEED=	specifies a seed value for MLE optimization
OPTTOLERANCE=	specifies the optimality tolerance for MLE optimization
SIGMA=	specifies scale parameter σ for fitted S_U curve
THETA=	specifies location parameter θ for fitted S_U curve
Weibull-Options	
C=	specifies shape parameter c for fitted Weibull curve
CDELTA=	specifies change in successive estimates of c at which the Newton-Raphson approximation of $\hat{c}$ terminates
CINITIAL=	specifies initial value for c in Newton-Raphson approximation of $\hat{c}$
MAXITER=	specifies maximum number of iterations in Newton-Raphson approximation of $\hat{c}$
SIGMA=	specifies scale parameter σ for fitted Weibull curve
THETA=	specifies threshold parameter θ for fitted Weibull curve

Nonparametric Density Estimation Options

Table 6.20 Kernel Density Estimation Options

Option	Description
KERNEL(*kernel-options*)	fits kernel density estimates

Specify the options listed in Table 6.21 in parentheses after the keyword KERNEL to control features of kernel density estimates requested with the KERNEL option.

Table 6.21 Kernel-Options

Option	Description
C=	specifies standardized bandwidth parameter c for fitted kernel density estimate
COLOR=	specifies color of the fitted kernel density curve
FILL	fills area under fitted kernel density curve
K=	specifies type of kernel function
L=	specifies line type used for fitted kernel density curve
LOWER=	specifies lower bound for fitted kernel density curve
SYMBOL=	specifies character used for fitted kernel density curve in line printer plots
UPPER=	specifies upper bound for fitted kernel density curve
W=	specifies line width for fitted kernel density curve

General Options

Table 6.22 summarizes general options for the HISTOGRAM statement, including options for enhancing charts and producing output data sets.

Table 6.22 General HISTOGRAM Statement Options

Option	Description
Options to Create Output Data Sets	
OUTFIT=	requests information about fitted curves
OUTHISTOGRAM=	requests information about histogram intervals
OUTKERNEL=	creates a data set containing kernel density estimates
General Histogram Layout Options	
CLIPCURVES	scales vertical axis without considering fitted curves
CONTENTS=	specifies table of contents entry for histogram grouping
CURVELEGEND=	specifies LEGEND statement for curves
ENDPOINTS=	lists endpoints for histogram intervals
HANGING	constructs hanging histogram
HREF=	specifies reference lines perpendicular to the horizontal axis
HREFLABELS=	specifies labels for HREF= lines
MIDPERCENTS	prints table of histogram intervals
MIDPOINTS=	lists midpoints for histogram intervals
NENDPOINTS=	specifies number of histogram interval endpoints
NMIDPOINTS=	specifies number of histogram interval midpoints
NOBARS	suppresses histogram bars
NOCURVELEGEND	suppresses legend for curves
NOFRAME	suppresses frame around plotting area
NOLEGEND	suppresses legend
NOPLOT	suppresses plot
NOSPECLEGEND	suppresses specifications legend
NOTABCONTENTS	suppresses table of contents entries for tables produced by HISTOGRAM statement

Table 6.22 (*continued*)

Option	Description
RTINCLUDE	includes right endpoint in interval
SPECLEGEND=	specifies LEGEND statement for specification limits
VREF=	specifies reference lines perpendicular to the vertical axis
VREFLABELS=	specifies labels for VREF= lines
VSCALE=	specifies scale for vertical axis
Options to Enhance Graphical Output	
ANNOTATE=	specifies annotate data set
BARLABEL=	produces labels above histogram bars
BARWIDTH=	specifies width for the bars
BMCFILL=	specifies fill color for box-and-whisker plot in bottom margin
BMCFRAME=	specifies fill color bottom margin plot frame
BMCOLOR=	specifies color for bottom margin plot
BMMARGIN=	specifies height of margin for bottom margin plot
BMPLOT=	requests a plot in bottom margin of histogram
CAXIS=	specifies color for axis
CBARLINE=	specifies color for outlines of histogram bars
CFILL=	specifies color for filling under curve
CFRAME=	specifies color for frame
CGRID=	specifies color for grid lines
CHREF=	specifies colors for HREF= lines
CLIPREF	draws reference lines behind histogram bars
CLIPSPEC=	clips histogram bars at specification limits
CSTATREF=	specifies colors for STATREF= lines
CTEXT=	specifies color for text
CVREF=	specifies colors for VREF= lines
DESCRIPTION=	specifies description for plot in graphics catalog
FONT=	specifies software font for text
FRONTREF	draws reference lines in front of histogram bars
GRID	creates a grid
HAXIS=	specifies AXIS statement for horizontal axis
HEIGHT=	specifies height of text used outside framed areas
HMINOR=	specifies number of horizontal minor tick marks
HOFFSET=	specifies offset for horizontal axis
HREFLABPOS=	specifies vertical position of labels for HREF= lines
INFONT=	specifies software font for text inside framed areas
INHEIGHT=	specifies height of text inside framed areas
INTERBAR=	specifies space between histogram bars
LEGEND=	identifies LEGEND statement
LGRID=	specifies a line type for grid lines
LHREF=	specifies line styles for HREF= lines
LSTATREF=	specifies line styles for STATREF= lines
LVREF=	specifies line styles for VREF= lines
MAXNBIN=	specifies maximum number of bins to display

Table 6.22 (*continued*)

Option	Description
MAXSIGMAS=	limits the number of bins that display to within a specified number of standard deviations above and below mean of data in key cell
MIDPOINTS=	specifies midpoints for histogram intervals
NAME=	specifies name for plot in graphics catalog
NOHLABEL	suppresses label for horizontal axis
NOVLABEL	suppresses label for vertical axis
NOVTICK	suppresses tick marks and tick mark labels for vertical axis
PFILL=	specifies pattern for filling under curve
STATREF=	specifies reference lines at values of summary statistics
STATREFLABELS=	specifies labels for STATREF= lines
STATREFSUBCHAR=	specifies substitution character for displaying statistic values in STATREFLABELS= labels
TURNVLABELS	turns and vertically strings out characters in labels for vertical axis
VAXIS=	specifies AXIS statement or values for vertical axis
VAXISLABEL=	specifies label for vertical axis
VMINOR=	specifies number of vertical minor tick marks
VOFFSET=	specifies length of offset at upper end of vertical axis
VREFLABPOS=	specifies horizontal position of labels for VREF= lines
WAXIS=	specifies line thickness for axes and frame
WBARLINE=	specifies line thickness for bar outlines
WGRID=	specifies line thickness for grid
Options for ODS Graphics Output	
ODSFOOTNOTE=	specifies footnote displayed on histogram
ODSFOOTNOTE2=	specifies secondary footnote displayed on histogram
ODSTITLE=	specifies title displayed on histogram
ODSTITLE2=	specifies secondary title displayed on histogram
Options for Comparative Plots	
ANNOKEY	applies annotation requested in ANNOTATE= data set to key cell only
CFRAMESIDE=	specifies color for filling frame for row labels
CFRAMETOP=	specifies color for filling frame for column labels
CPROP=	specifies color for proportion of frequency bar
CTEXTSIDE=	specifies color for row labels of comparative histograms
CTEXTTOP=	specifies color for column labels of comparative histograms
INTERTILE=	specifies distance between tiles
NCOLS=	specifies number of columns in comparative histogram
NROWS=	specifies number of rows in comparative histogram
OVERLAY	overlays plots for different class levels (ODS Graphics only)
Options to Enhance Line Printer Plots	
HREFCHAR=	specifies line character for HREF= lines
VREFCHAR=	specifies line character for VREF= lines

Dictionary of Options

The following sections provide detailed descriptions of options specific to the HISTOGRAM statement. See "Dictionary of Common Options: CAPABILITY Procedure" on page 533 for detailed descriptions of options common to all the plot statements.

General Options

ALPHA=*value-list*

specifies the shape parameter α for fitted curves requested with the BETA, GAMMA, PARETO, and POWER options. Enclose the ALPHA= option in parentheses after the distribution keyword. If you do not specify a value for α, the procedure calculates a maximum likelihood estimate. See Example 6.8. You can specify A= as an alias for ALPHA= if you use it as a *beta-option*. You can specify SHAPE= as an alias for ALPHA= if you use it as a *gamma-option*.

BARLABEL=COUNT | PERCENT | PROPORTION

displays labels above the histogram bars. If you specify BARLABEL=COUNT, the label shows the number of observations associated with a given bar. BARLABEL=PERCENT shows the percent of observations represented by that bar. If you specify BARLABEL=PROPORTION, the label displays the proportion of observations associated with the bar.

BETA< (*beta-options* **) >**

displays a fitted beta density curve on the histogram. The curve equation is

$$p(x) = \begin{cases} \frac{(x-\theta)^{\alpha-1}(\sigma+\theta-x)^{\beta-1}}{B(\alpha,\beta)\sigma^{(\alpha+\beta-1)}} hv & \text{for } \theta < x < \theta+\sigma \\ 0 & \text{for } x \le \theta \text{ or } x \ge \theta+\sigma \end{cases}$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and

θ = lower threshold parameter (lower endpoint parameter)
σ = scale parameter ($\sigma > 0$)
α = shape parameter ($\alpha > 0$)
β = shape parameter ($\beta > 0$)
h = width of histogram interval
v = vertical scaling factor
and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

The beta distribution is bounded below by the parameter θ and above by the value $\theta+\sigma$. You can specify θ and σ by using the THETA= and SIGMA= *beta-options*. The following statements fit a beta distribution bounded between 50 and 75 by using maximum likelihood estimates for α and β:

```
proc capability;
   histogram length / beta(theta=50 sigma=25);
run;
```

In general, the default values for THETA= and SIGMA= are 0 and 1, respectively. You can specify THETA=EST and SIGMA=EST to request maximum likelihood estimates for θ and σ.

The beta distribution has two shape parameters, α and β. If these parameters are known, you can specify their values with the ALPHA= and BETA= *beta-options*. If you do not specify values, the procedure calculates maximum likelihood estimates for α and β.

The BETA option can appear only once in a HISTOGRAM statement. Table 6.19 lists secondary options you can specify with the BETA option. See Example 6.8. Also see "Formulas for Fitted Curves" on page 336.

BETA=*value-list*

B=*value-list*

specifies the second shape parameter β for beta density curves requested with the BETA option. Enclose the BETA= option in parentheses after the BETA option. If you do not specify a value for β, the procedure calculates a maximum likelihood estimate. See Example 6.8.

BMPLOT=CARPET | DOTPLOT | SKELETAL | SCHEMATIC

produces a carpet plot, dot plot, or box-and-whisker plot along the bottom margin of a histogram. A carpet plot or dot plot shows the distribution of individual observations along the histogram's horizontal axis. A carpet plot represents each observation with a vertical line. A dot plot marks each observation with a symbol. A box-and-whisker plot gives a summary of the data distribution that a histogram alone does not provide. The left and right edges of the box are located at the first and third quartiles. A central vertical line is drawn at the median and a symbol is plotted inside the box at the mean. If you specify the SKELETAL keyword, a box-and-whisker plot is produced with whiskers extending to the minimum and maximum values. If you specify SCHEMATIC, a *schematic* box-and-whisker plot is produced. In a schematic box-and-whisker plot, the whiskers extend to the smallest value within the *lower fence* and the largest value within the *upper fence*. Fences are defined in terms of the interquartile range (IQR). The lower fence is 1.5 IQR below the first quartile and the upper fence is 1.5 IQR above the third quartile. Each observation outside the fences is plotted with a symbol.

C=*value-list*

specifies the shape parameter c for Weibull density curves requested with the WEIBULL option. Enclose the C= option in parentheses after the WEIBULL option. If you do not specify a value for c, the procedure calculates a maximum likelihood estimate. See Example 6.9. You can specify the SHAPE= option as an alias for the C= option.

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

specifies the standardized bandwidth parameter c for kernel density estimates requested with the KERNEL option. Enclose the C= option in parentheses after the KERNEL option. You can specify up to five values to request multiple estimates. You can also specify the C=MISE option, which produces the estimate with a bandwidth that minimizes the approximate mean integrated square error (MISE). For example, the following statements compute three density estimates:

```
proc capability;
   histogram length / kernel(c=0.5 1.0 mise);
run;
```

The first two estimates have standardized bandwidths of 0.5 and 1.0, respectively, and the third has a bandwidth that minimizes the approximate MISE.

You can also use the C= option with the K= option, which specifies the kernel function, to compute multiple estimates. If you specify more kernel functions than bandwidths, the last bandwidth in the list is repeated for the remaining estimates. Likewise, if you specify more bandwidths than kernel functions, the last kernel function is repeated for the remaining estimates. For example, the following statements compute three density estimates:

```
proc capability;
   histogram length / kernel(c=1 2 3 k=normal quadratic);
run;
```

The first uses a normal kernel and a bandwidth of 1, the second uses a quadratic kernel and a bandwidth of 2, and the third uses a quadratic kernel and a bandwidth of 3. See Example 6.12.

If you do not specify a value for c, the bandwidth that minimizes the approximate MISE is used for all the estimates.

CLIPCURVES

scales the vertical axis without taking fitted curves into consideration. Curves that extend above the tallest histogram bar may be clipped. You can use this option to avoid compression of the histogram bars due to extremely high fitted curve peaks.

DELTA=*value-list*

specifies the first shape parameter δ for Johnson S_B and Johnson S_U density curves requested with the SB and SU options. Enclose the DELTA= option in parentheses after the SB or SU option. If you do not specify a value for δ, the procedure calculates an estimate.

EDFNSAMPLES=*value*

specifies the number of simulation samples used to compute p-values for EDF goodness-of-fit statistics for density curves requested with the GUMBEL, IGAUSS, PARETO, and RAYLEIGH options. Enclose the EDFNSAMPLES= option in parentheses after the distribution option. The default value is 500.

EDFSEED=*value*

specifies an integer value used to start the pseudo-random number generator when creating simulation samples for computing EDF goodness-of-fit statistic p-values for density curves requested with the GUMBEL, IGAUSS, PARETO, and RAYLEIGH options. Enclose the EDFSEED= option in parentheses after the distribution option. By default, the procedure uses a random number seed generated from reading the time of day from the computer's clock.

ENDPOINTS

ENDPOINTS=*value-list*

specifies that histogram interval endpoints, rather than midpoints, are aligned with horizontal axis tick marks. If you specify ENDPOINTS, the number of histogram intervals is based on the number of observations by using the method of Terrell and Scott (1985). If you specify ENDPOINTS=*value-list*, the *values* must be listed in increasing order and must be evenly spaced. All observations in the input data set, as well as any specification limits, must lie between the first and last values specified. The same *value-list* is used for all variables.

EXPONENTIAL< (*exponential-options* **) >**

EXP< (*exponential-options* **) >**

displays a fitted exponential density curve on the histogram. The curve equation is

$$p(x) = \begin{cases} \frac{hv}{\sigma} \exp(-(\frac{x-\theta}{\sigma})) & \text{for } x \geq \theta \\ 0 & \text{for } x < \theta \end{cases}$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)
h = width of histogram interval
v = vertical scaling factor
and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

The parameter θ must be less than or equal to the minimum data value. You can specify θ with the THETA= *exponential-option*. The default value for θ is zero. If you specify THETA=EST, a maximum likelihood estimate is computed for θ. You can specify σ with the SIGMA= *exponential-option*. By default, a maximum likelihood estimate is computed for σ. For example, the following statements fit an exponential curve with $\theta = 10$ and with a maximum likelihood estimate for σ:

```
proc capability;
   histogram / exponential(theta=10 l=2 color=red);
run;
```

The curve is red and has a line type of 2. The EXPONENTIAL option can appear only once in a HISTOGRAM statement. Table 6.19 lists secondary options you can specify with the EXPONENTIAL option. See "Formulas for Fitted Curves" on page 336.

FILL

fills areas under a parametric density curve or kernel density estimate with colors and patterns. Enclose the FILL option in parentheses after a curve option or the KERNEL option, as in the following statements:

```
proc capability;
   histogram length / normal(fill) cfill=green pfill=solid;
run;
```

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 HISTOGRAM statement, as summarized in the following table:

Area Under Curve	Statement	Option
between specification	HISTOGRAM	CFILL=
limits	HISTOGRAM	PFILL=
left of lower	SPEC	CLEFT=
specification limit	SPEC	PLEFT=
right of upper	SPEC	CRIGHT=
specification limit	SPEC	PRIGHT=

If you do not display specification limits, the CFILL= and PFILL= options 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 6.8.1. Refer to *SAS/GRAPH: Help* for a list of available patterns and colors. If you do not specify the FILL option but specify the options in the preceding table, the colors and patterns are applied to the corresponding areas under the histogram.

FITINTERVAL=*value*

specifies the value of z for the method of percentiles when this method is used to fit a Johnson S_B or Johnson S_U distribution. The FITINTERVAL= option is specified in parentheses after the SB or SU option. The default of z is 0.524.

FITMETHOD=PERCENTILE | MLE | MOMENTS

specifies the method used to estimate the parameters of a Johnson S_B or Johnson S_U distribution. The FITMETHOD= option is specified in parentheses after the SB or SU option. By default, the method of percentiles is used. You can specify the MLE keyword to request maximum likelihood estimation. The OPTBOUNDRANGE=, OPTMAXITER=, OPTMAXSTARTS=, OPTPRINT, OPTSEED=, and OPTTOLERANCE= options control the optimizer that performs the maximum likelihood calculation.

FITTOLERANCE=*value*

specifies the tolerance value for the ratio criterion when the method of percentiles is used to fit a Johnson S_B or Johnson S_U distribution. The FITTOLERANCE= option is specified in parentheses after the SB or SU option. The default value is 0.01.

GAMMA< (*gamma-options***) >**

displays a fitted gamma density curve on the histogram. The curve equation is

$$p(x) = \begin{cases} \frac{hv}{\Gamma(\alpha)\sigma}\left(\frac{x-\theta}{\sigma}\right)^{\alpha-1}\exp\left(-\left(\frac{x-\theta}{\sigma}\right)\right) & \text{for } x > \theta \\ 0 & \text{for } x \le \theta \end{cases}$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)
α = shape parameter ($\alpha > 0$)
h = width of histogram interval
v = vertical scaling factor
and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

The parameter θ for the gamma distribution must be less than the minimum data value. You can specify θ with the THETA= *gamma-option*. The default value for θ is 0. If you specify THETA=EST, a maximum likelihood estimate is computed for θ. In addition, the gamma distribution has a shape parameter α and a scale parameter σ. You can specify these parameters with the ALPHA= and SIGMA= *gamma-options*. By default, maximum likelihood estimates are computed for α and σ. For example, the following statements fit a gamma curve with $\theta = 4$ and with maximum likelihood estimates for α and σ:

```
proc capability;
   histogram length / gamma(theta=4);
run;
```

Note that the maximum likelihood estimate of α is calculated iteratively using the Newton-Raphson approximation. The ALPHADELTA=, ALPHAINITIAL=, and MAXITER= *gamma-options* control the approximation.

The GAMMA option can appear only once in a HISTOGRAM statement. Table 6.19 lists secondary options you can specify with the GAMMA option. See Example 6.9 and "Formulas for Fitted Curves" on page 336.

GAMMA=*value-list*
specifies the second shape parameter γ for Johnson S_B and Johnson S_U density curves requested with the SB and SU options. Enclose the GAMMA= option in parentheses after the SB or SU option. If you do not specify a value for γ, the procedure calculates an estimate.

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

GUMBEL<(*Gumbel-options***)>**
displays a fitted Gumbel (also known as Type 1 extreme value distribution) density curve on the histogram. The curve equation is

$$p(x) = \frac{hv}{\sigma} e^{-(x-\mu)/\sigma} \exp\left(-e^{-(x-\mu)/\sigma}\right)$$

where

μ = location parameter
σ = scale parameter ($\sigma > 0$)
h = width of histogram interval
v = vertical scaling factor
and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

You can specify values for μ and σ with the MU= and SIGMA= *Gumbel-options*. By default, maximum likelihood estimates are computed for μ and σ.

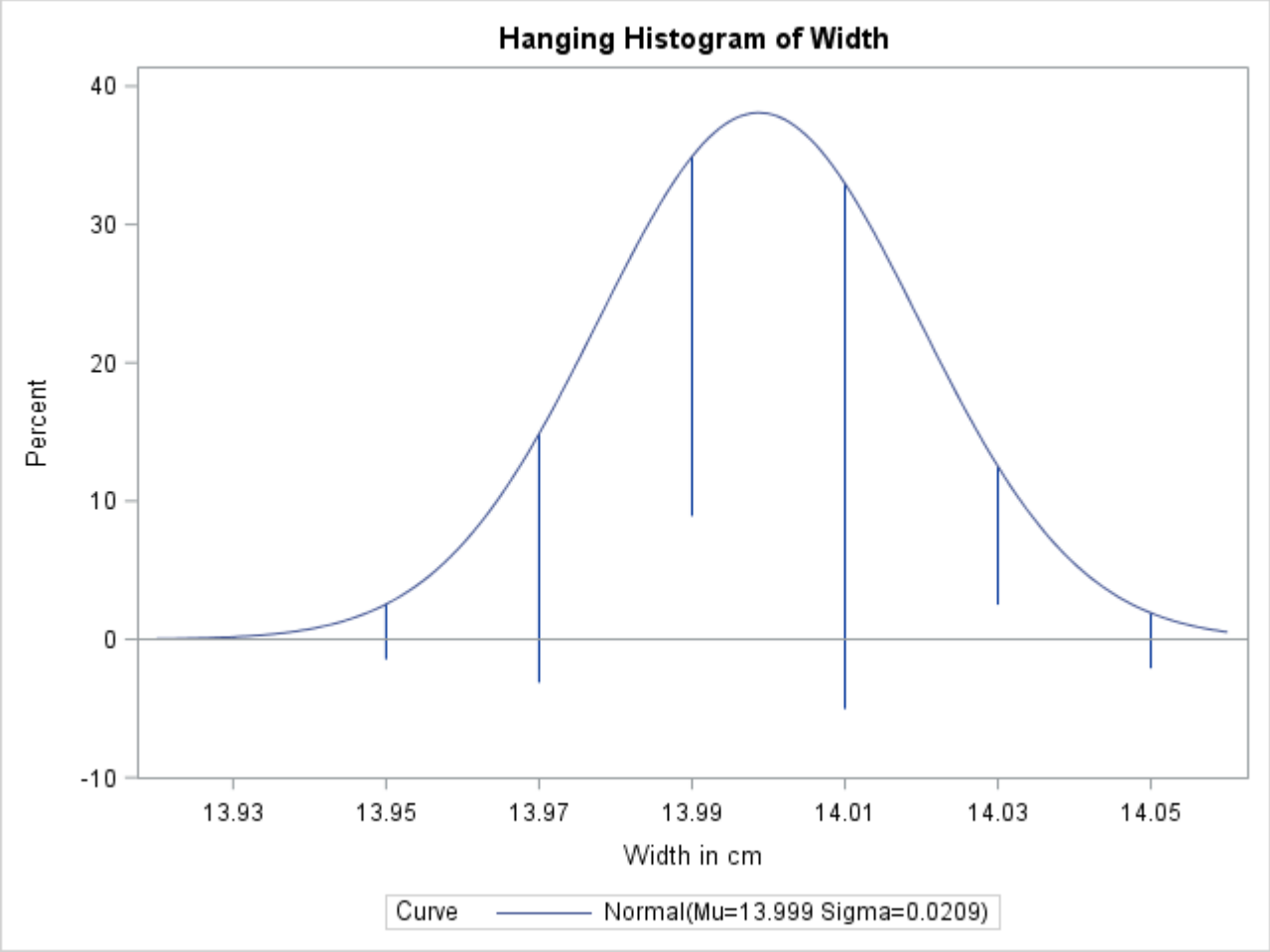
The GUMBEL option can appear only once in a HISTOGRAM statement. Table 6.19 lists secondary options you can specify with the GUMBEL option. See "Formulas for Fitted Curves" on page 336.

HANGING

HANG

requests a hanging histogram , as illustrated in Figure 6.12.

Figure 6.12 Hanging Histogram



You can use the HANGING option with only one fitted density curve. A hanging histogram aligns the tops of the histogram bars (displayed as lines) with the fitted curve. The lines are positioned at the midpoints of the histogram bins. A hanging histogram is a goodness-of-fit diagnostic in the sense that the closer the lines are to the horizontal axis, the better the fit. Hanging histograms are discussed by Tukey (1977), Wainer (1974), and Velleman and Hoaglin (1981).

IGAUSS< (*iGauss-options*) >

displays a fitted inverse Gaussian density curve on the histogram. The curve equation is

$$p(x) = \begin{cases} hv\left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp(-\frac{\lambda}{2\mu^2 x}(x-\mu)^2) & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function, and

μ = mean parameter ($\mu > 0$)
λ = shape parameter ($\lambda > 0$)
h = width of histogram interval
v = vertical scaling factor
and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

You can specify values for μ and λ with the MU= and LAMBDA= *iGauss-options*. By default, the sample mean is used for μ and a maximum likelihood estimate is computed for λ.

The IGAUSS option can appear only once in a HISTOGRAM statement. Table 6.19 lists secondary options you can specify with the IGAUSS option. See "Formulas for Fitted Curves" on page 336.

INDICES
requests capability indices based on the fitted distribution. Enclose the keyword INDICES in parentheses after the distribution keyword. See "Indices Using Fitted Curves" on page 353 for computational details and see Output 6.11.2.

K=NORMAL | QUADRATIC | TRIANGULAR
specifies the kernel function (normal, quadratic, or triangular) used to compute a kernel density estimate. Enclose the K= option in parentheses after the KERNEL option, as in the following statements:

```
proc capability;
   histogram length / kernel(k=quadratic);
run;
```

You can specify kernel functions for up to five estimates. You can also use the K= option together with the C= option, which specifies standardized bandwidths. If you specify more kernel functions than bandwidths, the last bandwidth in the list is repeated for the remaining estimates. Likewise, if you specify more bandwidths than kernel functions, the last kernel function is repeated for the remaining estimates. For example, the following statements compute three estimates with bandwidths of 0.5, 1.0, and 1.5:

```
proc capability;
   histogram length / kernel(c=0.5 1.0 1.5 k=normal quadratic);
run;
```

The first estimate uses a normal kernel, and the last two estimates use a quadratic kernel. By default, a normal kernel is used.

KERNEL<(*kernel-options*)>
superimposes up to five kernel density estimates on the histogram. You can specify the *kernel-options* described in the following table:

Option	Description
C=	specifies the smoothing parameter
COLOR=	specifies the color of the curve
FILL	specifies that the area under the curve is to be filled
K=	specifies the type of kernel function
L=	specifies the line style for the curve
LOWER=	specifies the lower bound for the curve
SYMBOL=	specifies the character used for the kernel density curve in line printer plots
UPPER=	specifies the upper bound for the curve
W=	specifies the width of the curve

You can request multiple kernel density estimates on the same histogram by specifying a list of values for either the C= or K= option. For more information, see the entries for these options. Also see Output 6.6.1 and "Kernel Density Estimates" on page 347. By default, kernel density estimates are computed using the AMISE method.

LAMBDA=*value*

specifies the shape parameter λ for fitted curves requested with the IGAUSS option. Enclose the LAMBDA= option in parentheses after the IGAUSS distribution keyword. If you do not specify a value for λ, the procedure calculates a maximum likelihood estimate.

LOGNORMAL< (*lognormal-options***) >**

displays a fitted lognormal density curve on the histogram. The curve equation is

$$p(x) = \begin{cases} \frac{hv}{\sigma\sqrt{2\pi}(x-\theta)} \exp\left(-\frac{(\log(x-\theta)-\zeta)^2}{2\sigma^2}\right) & \text{for } x > \theta \\ 0 & \text{for } x \leq \theta \end{cases}$$

where

θ = threshold parameter
ζ = scale parameter
σ = shape parameter ($\sigma > 0$)
h = width of histogram interval
v = vertical scaling factor
and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

Note that the lognormal distribution is also referred to as the S_L distribution in the Johnson system of distributions.

The parameter θ for the lognormal distribution must be less than the minimum data value. You can specify θ with the THETA= *lognormal-option*. The default value for θ is zero. If you specify THETA=EST, a maximum likelihood estimate is computed for θ. You can specify the parameters σ and ζ with the SIGMA= and ZETA= *lognormal-options*. By default, maximum likelihood estimates are computed for σ and ζ. For example, the following statements fit a lognormal distribution function with a default value of $\theta = 0$ and with maximum likelihood estimates for σ and ζ:

```
proc capability;
   histogram length / lognormal;
run;
```

The LOGNORMAL option can appear only once in a HISTOGRAM statement. Table 6.19 lists secondary options that you can specify with the LOGNORMAL option. See Example 6.9 and "Formulas for Fitted Curves" on page 336.

LOWER=*value-list*

specifies lower bounds for kernel density estimates requested with the KERNEL option. Enclose the LOWER= option in parentheses after the KERNEL option. You can specify up to five lower bounds for multiple kernel density estimates. If you specify more kernel estimates than lower bounds, the last lower bound is repeated for the remaining estimates.

MAXNBIN=*n*

specifies the maximum number of bins to be displayed in a comparative histogram. 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.

MIDPERCENTS

requests a table listing the midpoints and percent of observations in each histogram interval. For example, the following statements create the table in Figure 6.13:

```
proc capability;
   histogram Length / midpercents;
run;
```

Figure 6.13 Table of Midpoints and Observed Percentages

The CAPABILITY Procedure

Histogram Bins for Length	
Bin Midpoint	Observed Percent
10.02	12.000
10.08	32.000
10.14	28.000
10.20	18.000
10.26	6.000
10.32	4.000

If you specify the MIDPERCENTS option in parentheses after a density estimate option, a table listing the midpoints, observed percent of observations, and the estimated percent of the population in each interval (estimated from the fitted distribution) is printed.

The following statements create the table shown in Figure 6.14:

```
proc capability;
   histogram Length / gamma(theta=3 midpercents);
run;
```

Figure 6.14 Table of Observed and Expected Percentages

The CAPABILITY Procedure
Fitted Gamma Distribution for Length (Attachment Point Offset in mm)

Histogram Bin Percents for Gamma Distribution		
	Percent	
Bin Midpoint	Observed	Estimated
10.02	12.000	11.480
10.08	32.000	26.182
10.14	28.000	31.354
10.20	18.000	19.916
10.26	6.000	6.766
10.32	4.000	1.238

MIDPOINTS=*value-list* | **KEY** | **UNIFORM**

specifies how to determine the midpoints for the histogram intervals, where *values-list* determines the width of the histogram bars as the difference between consecutive midpoints. The procedure uses the same values for all variables. See Output 6.9.1.

The range of midpoints, extended at each end by half of the bar width, must cover the range of the data as well as any specification limits. For example, if you specify

```
midpoints=2 to 10 by 0.5
```

then all of the observations and specification limits should fall between 1.75 and 10.25. (Otherwise, a default list of midpoints is used.) You must use evenly spaced midpoints listed in increasing order.

KEY 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 that use the method of Terrell and Scott (1985). The procedure extends the midpoint list for the key cell in either direction as necessary until it spans the data in the remaining cells.

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

Neither KEY nor UNIFORM apply unless you use the CLASS statement. By default, if you use a CLASS statement, MIDPOINTS=KEY. However, if the key cell is empty then MIDPOINTS=UNIFORM. Otherwise, the procedure computes the midpoints by using the algorithm described in Terrell and Scott (1985). The default midpoints are primarily applicable to continuous data that are approximately normally distributed.

If you produce traditional graphics and use the MIDPOINTS= and HAXIS= options, you can use the ORDER= option in the AXIS statement you specified with the HAXIS= option. However, for the tick mark labels to coincide with the histogram interval midpoints, the range of the ORDER= list must encompass the range of the MIDPOINTS= list, as illustrated in the following statements:

```
proc capability;
   histogram length / midpoints=20 to 80 by 10
                      haxis=axis1;
   axis1 length=6 in order=10 20 30 40 50 60 70 80 90;
run;
```

MIDPTAXIS=*name*

is an alias for the HAXIS= option.

MU=*value-list*

specifies the parameter μ for fitted curves requested with the GUMBEL, IGAUSS, and NORMAL options. Enclose the MU= option in parentheses after the distribution keyword. For the normal and inverse Gaussian distributions, the default value of μ is the sample mean. If you do not specify a value for μ for the Gumbel distribution, the procedure calculates a maximum likelihood estimate.

NENDPOINTS=*n*

specifies the number of histogram interval endpoints and causes the endpoints, rather than interval midpoints, to be aligned with horizontal axis tick marks.

NMIDPOINTS=*n*

specifies the number of histogram intervals.

NOBARS
suppresses drawing of histogram bars. This option is useful when you want to display fitted curves only.

NOCURVELEGEND
NOCURVEL
suppresses the portion of the legend for fitted curves. If you use the INSET statement to display information about the fitted curve on the histogram, you can use the NOCURVELEGEND option to prevent the information about the fitted curve from being repeated in a legend at the bottom of the histogram. See Output 6.15.1.

NOLEGEND
suppresses legends for specification limits, fitted curves, and hidden observations. See Example 6.13. Specifying the NOLEGEND option is equivalent to specifying LEGEND=NONE.

NOPLOT
suppresses the creation of a plot. Use the NOPLOT option when you want only to print summary statistics for a fitted density or create either an OUTFIT= or an OUTHISTOGRAM= data set. See Example 6.11.

NOPRINT
suppresses printed output summarizing the fitted curve. Enclose the NOPRINT option in parentheses following the distribution option. See "Customizing a Histogram" on page 304 for an example.

NORMAL< (*normal-options*) >
displays a fitted normal density curve on the histogram. The curve equation is

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

where

μ = mean
σ = standard deviation ($\sigma > 0$)
h = width of histogram interval
v = vertical scaling factor
and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

Note that the normal distribution is also referred to as the S_N distribution in the Johnson system of distributions.

You can specify values for μ and σ with the MU= and SIGMA= *normal-options*, as shown in the following statements:

```
proc capability;
   histogram length / normal(mu=14 sigma=0.05);
run;
```

By default, the sample mean and sample standard deviation are used for μ and σ. The NORMAL option can appear only once in a HISTOGRAM statement. Table 6.19 lists secondary options that you can specify with the NORMAL option. See Figure 6.10 and "Formulas for Fitted Curves" on page 336.

NOSPECLEGEND

NOSPECL

suppresses the portion of the legend for specification limit reference lines. See Figure 6.11.

NOTABCONTENTS

suppresses the table of contents entries for tables produced by the HISTOGRAM statement. See the section "ODS Tables" on page 359 for descriptions of the tables produced by the HISTOGRAM statement.

OPTBOUNDRANGE=*value*

defines the sampling range for each parameter during maximum likelihood estimation for the Johnson S_U distribution. PROC UNIVARIATE computes initial estimates for each parameter by using the method of percentiles. The *value* determines the range of parameter values around the initial estimate that can be sampled for local optimization starting values. The default is 100.

OPTMAXITER=*value*

limits the number of iterations that are used by the optimizer in maximum likelihood estimation for the Johnson S_U distribution. The default is 500.

OPTMAXSTARTS=*N*

defines the maximum number of starting points to be used for local optimization in maximum likelihood estimation for the Johnson S_U distribution. That is, no more than *N* local optimizations are used in the multistart algorithm. The default value is 100.

OPTPRINT

prints the iteration history for the Johnson S_U distribution maximum likelihood estimation.

OPTSEED=*value*

specifies a positive integer seed for generating random number sequences in Johnson S_U distribution maximum likelihood estimation. You can use this option to replicate results from different runs.

OPTTOLERANCE=*value*

specifies the tolerance for declaring optimality in maximum likelihood estimation for the Johnson S_U distribution. The default value is 1E–8.

OUTFIT=*SAS-data-set*

creates a SAS data set that contains parameter estimates for fitted curves and related goodness-of-fit information. See "Output Data Sets" on page 355.

OUTHISTOGRAM=*SAS-data-set*

OUTHIST=*SAS-data-set*

creates a SAS data set that contains information about histogram intervals. Specifically, the data set contains the midpoints of the histogram intervals, the observed percent of observations in each interval, and the estimated percent of observations in each interval (estimated from each of the specified fitted curves). See "Output Data Sets" on page 355.

OUTKERNEL=*SAS-data-set*

creates a SAS data set containing information about kernel density estimates requested with the KERNEL option. See "OUTKERNEL= Output Data Set" on page 358 for details.

PARETO<(*Pareto-options***)>**

displays a fitted generalized Pareto density curve on the histogram. The curve equation is

$$p(x) = \begin{cases} \frac{hv}{\sigma}(1-\alpha(x-\theta)/\sigma)^{1/\alpha-1} & \text{if } \alpha \neq 0 \\ \frac{hv}{\sigma}\exp(-(x-\theta)/\sigma) & \text{if } \alpha = 0 \end{cases}$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)
α = shape parameter
h = width of histogram interval
v = vertical scaling factor
and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

The parameter θ must be less than the minimum data value. You can specify θ with the THETA= *Pareto-option*. The default value for θ is zero. If you specify THETA=EST, a maximum likelihood estimate is computed for θ. In addition, the generalized Pareto distribution has a shape parameter α and a scale parameter σ. You can specify these parameters with the ALPHA= and SIGMA= *Pareto-options*. By default, maximum likelihood estimates are computed for α and σ.

The PARETO option can appear only once in a HISTOGRAM statement. Table 6.19 lists secondary options you can specify with the PARETO option. See "Formulas for Fitted Curves" on page 336.

PCTAXIS=*name*|*value-list*

is an alias for the VAXIS= option.

PERCENTS=*value-list*

PERCENT=*value-list*

specifies a list of percents for which quantiles calculated from the data and quantiles estimated from the fitted curve are tabulated. The percents must be between 0 and 100. Enclose the PERCENTS= option in parentheses after the curve option. The default percents are 1, 5, 10, 25, 50, 75, 90, 95, and 99.

For example, the following statements create the table shown in Figure 6.15:

```
proc capability;
   histogram Length / lognormal(percents=1 3 5 95 97 99);
run;
```

Figure 6.15 Estimated and Observed Quantiles for the Lognormal Curve

The CAPABILITY Procedure
Fitted Lognormal Distribution for Length (Attachment Point Offset in mm)

Quantiles for Lognormal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	10.0180	9.95696
3.0	10.0180	9.98937
5.0	10.0310	10.00658
95.0	10.2780	10.24963
97.0	10.2930	10.26729
99.0	10.3220	10.30071

POWER< (*power-options*) >
displays a fitted power function density curve on the histogram. The curve equation is

$$p(x) = \begin{cases} hv\frac{\alpha}{\sigma}\left(\frac{x-\theta}{\sigma}\right)^{\alpha-1} & \text{for } \theta < x < \theta + \sigma \\ 0 & \text{for } x \le \theta \text{ or } x \ge \theta + \sigma \end{cases}$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)
α = shape parameter
h = width of histogram interval
v = vertical scaling factor
and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

The parameter θ must be less than or equal to the minimum data value. You can specify θ and σ with the THETA= and the SIGMA= *power-options*. The default values for θ and σ are 0 and 1, respectively. You can specify THETA=EST and SIGMA=EST to request maximum likelihood estimates for θ and σ.

In addition, the generalized Pareto distribution has a shape parameter α. You can specify α with the ALPHA= *power-option*. By default, a maximum likelihood estimate is computed for α.

The POWER option can appear only once in a HISTOGRAM statement. Table 6.19 lists secondary options you can specify with the POWER option. See "Formulas for Fitted Curves" on page 336.

RAYLEIGH< (*Rayleigh-options*) >
displays a fitted Rayleigh density curve on the histogram. The curve equation is

$$p(x) = \begin{cases} hv\frac{x-\theta}{\sigma^2}e^{-(x-\theta)^2/(2\sigma^2)} & \text{for } x \ge \theta \\ 0 & \text{for } x < \theta \end{cases}$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)
h = width of histogram interval
v = vertical scaling factor
and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

The parameter θ must be less than or equal to the minimum data value. You can specify θ with the THETA= *Rayleigh-option*. The default value for θ is zero. If you specify THETA=EST, a maximum likelihood estimate is computed for θ. You can specify σ with the SIGMA= *Rayleigh-option*. By default, a maximum likelihood estimate is computed for σ.

The RAYLEIGH option can appear only once in a HISTOGRAM statement. Table 6.19 lists secondary options you can specify with the RAYLEIGH option. See "Formulas for Fitted Curves" on page 336.

RTINCLUDE
includes the right endpoint of each histogram interval in that interval. By default, the left endpoint is included in the histogram interval.

SB<(*S_B-options*)>
displays a fitted Johnson S_B density curve on the histogram. The curve equation is

$$p(x) = \begin{cases} \frac{\delta h v}{\sigma\sqrt{2\pi}} \left[\left(\frac{x-\theta}{\sigma} \right) \left(1 - \frac{x-\theta}{\sigma} \right) \right]^{-1} \times & \\ \exp\left[-\frac{1}{2} \left(\gamma + \delta \log\left(\frac{x-\theta}{\theta+\sigma-x} \right) \right)^2 \right] & \text{for } \theta < x < \theta + \sigma \\ 0 & \text{for } x \leq \theta \text{ or } x \geq \theta + \sigma \end{cases}$$

where

θ = threshold parameter ($-\infty < \theta < \infty$)
σ = scale parameter ($\sigma > 0$)
δ = shape parameter ($\delta > 0$)
γ = shape parameter ($-\infty < \gamma < \infty$)
h = width of histogram interval
v = vertical scaling factor
and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

The S_B distribution is bounded below by the parameter θ and above by the value $\theta + \sigma$. The parameter θ must be less than the minimum data value. You can specify θ with the THETA= *S_B-option*, or you can request that θ be estimated with the THETA = EST *S_B-option*. The default value for θ is zero.

The sum $\theta + \sigma$ must be greater than the maximum data value. The default value for σ is one. You can specify σ with the SIGMA= S_B*-option*, or you can request that σ be estimated with the SIGMA = EST S_B*-option*. You can specify δ with the DELTA= S_B*-option*, and you can specify γ with the GAMMA= S_B*-option*. Note that the S_B*-options* are given in parentheses after the SB option.

By default, the method of percentiles is used to estimate the parameters of the S_B distribution. Alternatively, you can request the method of moments or the method of maximum likelihood with the FITMETHOD = MOMENTS or FITMETHOD = MLE options, respectively. Consider the following example:

```
proc capability;
   histogram length / sb;
   histogram length / sb( theta=est sigma=est );
   histogram length / sb( theta=0.5 sigma=8.4
                          delta=0.8 gamma=-0.6 );
run;
```

The first HISTOGRAM statement fits an S_B distribution with default values of $\theta = 0$ and $\sigma = 1$ and with percentile-based estimates for δ and γ. The second HISTOGRAM statement estimates all four parameters with the method of percentiles. The third HISTOGRAM statement displays an S_B curve with specified values for all four parameters.

The SB option can appear only once in a HISTOGRAM statement. Table 6.19 lists secondary options you can specify with the SB option.

SIGMA=*value-list*

specifies the parameter σ for fitted curves requested with the BETA, EXPONENTIAL, GAMMA, GUMBEL, LOGNORMAL, NORMAL, PARETO, POWER, RAYLEIGH, SB, SU, and WEIBULL options. Enclose the SIGMA= option in parentheses after the distribution keyword. The following table summarizes the use of the SIGMA= option.

Distribution Keyword	SIGMA= Specifies	Default Value	Alias
BETA	scale parameter σ	1	SCALE=
EXPONENTIAL	scale parameter σ	maximum likelihood estimate	SCALE=
GAMMA	scale parameter σ	maximum likelihood estimate	SCALE=
GUMBEL	scale parameter σ	maximum likelihood estimate	
LOGNORMAL	shape parameter σ	maximum likelihood estimate	SHAPE=
NORMAL	scale parameter σ	standard deviation	
PARETO	scale parameter σ	maximum likelihood estimate	
POWER	scale parameter σ	1	SCALE=
RAYLEIGH	scale parameter σ	maximum likelihood estimate	
SB	scale parameter σ	1	SCALE=
SU	scale parameter σ	percentile-based estimate	SCALE=
WEIBULL	scale parameter σ	maximum likelihood estimate	SCALE=

If you specify SIGMA=EST, an estimate is computed for σ. For syntax examples, see the entries for the distribution options.

SU<(*S_U -options*)>

displays a fitted Johnson S_U density curve on the histogram. The curve equation is

$$p(x) = \begin{cases} \frac{\delta h v}{\sigma\sqrt{2\pi}} \frac{1}{\sqrt{1+((x-\theta)/\sigma)^2}} \times \\ \exp\left[-\frac{1}{2}\left(\gamma + \delta \sinh^{-1}\left(\frac{x-\theta}{\sigma}\right)\right)^2\right] & \text{for } x > \theta \\ 0 & \text{for } x \le \theta \end{cases}$$

where

θ = location parameter ($-\infty < \theta < \infty$)
σ = scale parameter ($\sigma > 0$)
δ = shape parameter ($\delta > 0$)
γ = shape parameter ($-\infty < \gamma < \infty$)
h = width of histogram interval
v = vertical scaling factor
and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

You can specify the parameters with the THETA=, SIGMA=, DELTA=, and GAMMA= *S_U -options*, which are enclosed in parentheses after the SU option. If you do not specify these parameters, they are estimated.

By default, the method of percentiles is used to estimate the parameters of the S_U distribution. Alternatively, you can request the method of moments or the method of maximum likelihood with the FITMETHOD = MOMENTS or FITMETHOD = MLE options, respectively. Consider the following example:

```
proc capability;
   histogram length / su;
   histogram length / su( theta=0.5 sigma=8.4
                          delta=0.8 gamma=-0.6 );
run;
```

The first HISTOGRAM statement estimates all four parameters with the method of percentiles. The second HISTOGRAM statement displays an S_U curve with specified values for all four parameters.

The SU option can appear only once in a HISTOGRAM statement. Table 6.19 lists secondary options you can specify with the SU option.

THETA=*value-list*

THRESHOLD=*value-list*

specifies the lower threshold parameter θ for curves requested with the BETA, EXPONENTIAL, GAMMA, LOGNORMAL, PARETO, POWER, RAYLEIGH, SB, and WEIBULL options, and the location parameter θ for curves requested with the SU option. Enclose the THETA= option in parentheses after the curve option. See Example 6.8. The default *value* is zero. If you specify THETA=EST, an estimate is computed for θ.

UPPER=*value-list*

specifies upper bounds for kernel density estimates requested with the KERNEL option. Enclose the UPPER= option in parentheses after the KERNEL option. You can specify up to five upper bounds for multiple kernel density estimates. If you specify more kernel estimates than upper bounds, the last upper bound is repeated for the remaining estimates.

VSCALE=COUNT | PERCENT | 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. See Figure 6.11 for an illustration of VSCALE=COUNT. The default is PERCENT.

WEIBULL< (*Weibull-options***) >**

displays a fitted Weibull density curve on the histogram. The curve equation is

$$p(x) = \begin{cases} \frac{chv}{\sigma}(\frac{x-\theta}{\sigma})^{c-1}\exp(-(\frac{x-\theta}{\sigma})^c) & \text{for } x > \theta \\ 0 & \text{for } x \le \theta \end{cases}$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)
c = shape parameter ($c > 0$)
h = width of histogram interval
v = vertical scaling factor
and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

The parameter θ must be less than the minimum data value. You can specify θ with the THETA= *Weibull-option*. The default value for θ is zero. If you specify THETA=EST, a maximum likelihood estimate is computed for θ. You can specify σ and c with the SIGMA= and C= *Weibull-options*. By default, maximum likelihood estimates are computed for c and σ. For example, the following statements fit a Weibull distribution with $\theta = 15$ and with maximum likelihood estimates for σ and c:

```
proc capability;
   histogram length / weibull(theta=15);
run;
```

Note that the maximum likelihood estimate of c is calculated iteratively using the Newton-Raphson approximation. The CDELTA=, CINITIAL=, and MAXITER= *Weibull-options* control the approximation.

The WEIBULL option can appear only once in a HISTOGRAM statement. Table 6.19 lists secondary options that you can specify with the WEIBULL option. See Example 6.9 and "Formulas for Fitted Curves" on page 336.

ZETA=*value-list*
specifies a value for the scale parameter ζ for lognormal density curves requested with the LOGNORMAL option. Enclose the ZETA= option in parentheses after the LOGNORMAL option. By default, the procedure calculates a maximum likelihood estimate for ζ. You can specify the SCALE= option as an alias for the ZETA= option.

Options for Traditional Graphics

BARWIDTH=*value*
specifies the width of the histogram bars in screen percent units.

BMCFILL=*color*
specifies the fill color for a box-and-whisker plot in a bottom margin requested with the BMPLOT= option. By default, the box-and-whisker plot is not filled.

BMCFRAME=*color*
specifies the color for filling the frame of a bottom margin plot requested with the BMPLOT= option. By default, this area is not filled.

BMCOLOR=*color*
specifies the color of a carpet plot, or the outline color of a box-and-whisker plot, in a bottom margin plot requested with the BMPLOT= option.

BMMARGIN=*height*
specifies the height in screen percentage units of a bottom margin plot requested with the BMPLOT= option. By default, a bottom margin plot occupies 15 percent of the vertical display space.

CBARLINE=*color*
specifies the color of the outline of histogram bars. This option overrides the C= option in the SYMBOL1 statement.

CFILL=*color*
specifies a color used to fill the bars of the histogram (or the area under a fitted curve if you also specify the FILL option). See the entries for the FILL and PFILL= options for additional details. See Figure 6.11 and Output 6.8.1. Refer to *SAS/GRAPH: Help* for a list of colors. By default, bars are filled with an appropriate color from the ODS style.

CGRID=*color*
specifies the color for grid lines requested with the GRID option. By default, grid lines are the same color as the axes. If you use CGRID=, you do not need to specify the GRID option.

CLIPREF
draws reference lines requested with the HREF= and VREF= options behind the histogram bars. By default, reference lines are drawn in front of the histogram bars.

CLIPSPEC=CLIP | NOFILL
specifies that histogram bars are clipped at the upper and lower specification limit lines when there are no observations outside the specification limits. The bar intersecting the lower specification limit is clipped if there are no observations less than the lower limit; the bar intersecting the upper specification limit is clipped if there are no observations greater than the upper limit. If you specify CLIPSPEC=CLIP, the histogram bar is truncated at the specification limit. If you specify CLIPSPEC=NOFILL, the portion of a filled histogram bar outside the specification limit is left unfilled. Specifying CLIPSPEC=NOFILL when histogram bars are not filled has no effect.

CURVELEGEND=*name* | **NONE**
specifies the name of a LEGEND statement describing the legend for specification limits and fitted curves. Specifying CURVELEGEND=NONE suppresses the legend for fitted curves; this is equivalent to specifying the NOCURVELEGEND option.

FRONTREF
draws reference lines requested with the HREF= and VREF= options in front of the histogram bars. When the NOGSTYLE system option is specified, reference lines are drawn behind the histogram bars by default, and can be obscured by them.

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

INTERBAR=*value*
specifies the horizontal space in percent screen units between histogram bars. By default, the bars are contiguous.

LEGEND=*name* | **NONE**
specifies the name of a LEGEND statement describing the legend for specification limit reference lines and fitted curves. Specifying LEGEND=NONE suppresses all legend information and is equivalent to specifying the NOLEGEND option.

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.

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: Help* for a list of pattern values. By default, the bars and curve areas are not filled.

SPECLEGEND=*name* | **NONE**
specifies the name of a LEGEND statement describing the legend for specification limits and fitted curves. Specifying SPECLEGEND=NONE, which suppresses the portion of the legend for specification limit references lines, is equivalent to specifying the NOSPECLEGEND option.

VOFFSET=*value*
specifies the offset in percent screen units at the upper end of the vertical axis.

WBARLINE=*n*
specifies the width of bar outlines. By default, $n = 1$.

WGRID=*n*
specifies the width of the grid lines requested with the GRID option. By default, grid lines are the same width as the axes. If you use the WGRID= option, you do not need to specify the GRID option.

Options for Legacy Line Printer Charts

SYMBOL='*character*'

specifies the *character* used for the density curve or kernel density curve in line printer plots. Enclose the SYMBOL= option in parentheses after the distribution option or the KERNEL option. The default character is the first letter of the distribution keyword or '1' for the first kernel density estimate, '2' for the second kernel density estimate, and so on. If you use the SYMBOL= option with the KERNEL option, you can specify a list of up to five characters in parentheses for multiple kernel density estimates. If there are more estimates than characters, the last character specified is used for the remaining estimates.

Details: HISTOGRAM Statement

This section provides details on the following topics:

- formulas for fitted distributions
- formulas for kernel density estimates
- printed output
- OUTFIT=, OUTHISTOGRAM=, and OUTKERNEL= data sets
- graphical enhancements to histograms

Formulas for Fitted Curves

The following sections provide information about the families of parametric distributions that you can fit with the HISTOGRAM statement. Properties of these distributions are discussed by Johnson, Kotz, and Balakrishnan (1994) and Johnson, Kotz, and Balakrishnan (1995).

Beta Distribution

The fitted density function is

$$p(x) = \begin{cases} \frac{(x-\theta)^{\alpha-1}(\sigma+\theta-x)^{\beta-1}}{B(\alpha,\beta)\sigma^{(\alpha+\beta-1)}} hv & \text{for } \theta < x < \theta+\sigma \\ 0 & \text{for } x \le \theta \text{ or } x \ge \theta+\sigma \end{cases}$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and

θ = lower threshold parameter (lower endpoint parameter)
σ = scale parameter ($\sigma > 0$)
α = shape parameter ($\alpha > 0$)
β = shape parameter ($\beta > 0$)
h = width of histogram interval
v = vertical scaling factor, and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

NOTE: This notation is consistent with that of other distributions that you can fit with the HISTOGRAM statement. However, many texts, including Johnson, Kotz, and Balakrishnan (1995), write the beta density function as

$$p(x) = \begin{cases} \frac{(x-a)^{p-1}(b-x)^{q-1}}{B(p,q)(b-a)^{p+q-1}} & \text{for } a < x < b \\ 0 & \text{for } x \leq a \text{ or } x \geq b \end{cases}$$

The two notations are related as follows:

$\sigma = b - a$
$\theta = a$
$\alpha = p$
$\beta = q$

The range of the beta distribution is bounded below by a threshold parameter $\theta = a$ and above by $\theta + \sigma = b$. If you specify a fitted beta curve by using the BETA option, θ must be less than the minimum data value, and $\theta + \sigma$ must be greater than the maximum data value. You can specify θ and σ with the THETA= and SIGMA= *beta-options* in parentheses after the keyword BETA. By default, $\sigma = 1$ and $\theta = 0$. If you specify THETA=EST and SIGMA=EST, maximum likelihood estimates are computed for θ and σ.

In addition, you can specify α and β with the ALPHA= and BETA= *beta-options*, respectively. By default, the procedure calculates maximum likelihood estimates for α and β. For example, to fit a beta density curve to a set of data bounded below by 32 and above by 212 with maximum likelihood estimates for α and β, use the following statement:

```
histogram length / beta(theta=32 sigma=180);
```

The beta distributions are also referred to as Pearson Type I or II distributions. These include the *power-function* distribution ($\beta = 1$), the *arc-sine* distribution ($\alpha = \beta = \frac{1}{2}$), and the *generalized arc-sine* distributions ($\alpha + \beta = 1, \beta \neq \frac{1}{2}$).

You can use the DATA step function BETAINV to compute beta quantiles and the DATA step function PROBBETA to compute beta probabilities.

Exponential Distribution

The fitted density function is

$$p(x) = \begin{cases} \frac{hv}{\sigma} \exp(-(\frac{x-\theta}{\sigma})) & \text{for } x \geq \theta \\ 0 & \text{for } x < \theta \end{cases}$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)
h = width of histogram interval
v = vertical scaling factor, and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

The threshold parameter θ must be less than or equal to the minimum data value. You can specify θ with the THRESHOLD= *exponential-option*. By default, $\theta = 0$. If you specify THETA=EST, a maximum likelihood estimate is computed for θ. In addition, you can specify σ with the SCALE= *exponential-option*. By default, the procedure calculates a maximum likelihood estimate for σ. Note that some authors define the scale parameter as $\frac{1}{\sigma}$.

The exponential distribution is a special case of both the gamma distribution (with $\alpha = 1$) and the Weibull distribution (with $c = 1$). A related distribution is the *extreme value* distribution. If $Y = \exp(-X)$ has an exponential distribution, then X has an extreme value distribution.

Gamma Distribution

The fitted density function is

$$p(x) = \begin{cases} \frac{hv}{\Gamma(\alpha)\sigma}\left(\frac{x-\theta}{\sigma}\right)^{\alpha-1}\exp\left(-\left(\frac{x-\theta}{\sigma}\right)\right) & \text{for } x > \theta \\ 0 & \text{for } x \le \theta \end{cases}$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)
α = shape parameter ($\alpha > 0$)
h = width of histogram interval
v = vertical scaling factor, and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

The threshold parameter θ must be less than the minimum data value. You can specify θ with the THRESHOLD= *gamma-option*. By default, $\theta = 0$. If you specify THETA=EST, a maximum likelihood estimate is computed for θ. In addition, you can specify σ and α with the SCALE= and ALPHA= *gamma-options*. By default, the procedure calculates maximum likelihood estimates for σ and α.

The gamma distributions are also referred to as Pearson Type III distributions, and they include the chi-square, exponential, and Erlang distributions. The probability density function for the chi-square distribution is

$$p(x) = \begin{cases} \frac{1}{2\Gamma\left(\frac{\nu}{2}\right)}\left(\frac{x}{2}\right)^{\frac{\nu}{2}-1}\exp\left(-\frac{x}{2}\right) & \text{for } x > 0 \\ 0 & \text{for } x \le 0 \end{cases}$$

Notice that this is a gamma distribution with $\alpha = \frac{\nu}{2}$, $\sigma = 2$, and $\theta = 0$. The exponential distribution is a gamma distribution with $\alpha = 1$, and the Erlang distribution is a gamma distribution with α being a positive integer. A related distribution is the Rayleigh distribution. If $R = \frac{\max(X_1,\ldots,X_n)}{\min(X_1,\ldots,X_n)}$ where the X_i's are independent χ^2_ν variables, then $\log R$ is distributed with a χ_ν distribution having a probability density function of

$$p(x) = \begin{cases} \left[2^{\frac{\nu}{2}-1}\Gamma\left(\frac{\nu}{2}\right)\right]^{-1} x^{\nu-1}\exp\left(-\frac{x^2}{2}\right) & \text{for } x > 0 \\ 0 & \text{for } x \le 0 \end{cases}$$

If $\nu = 2$, the preceding distribution is referred to as the Rayleigh distribution.

You can use the DATA step function GAMINV to compute gamma quantiles and the DATA step function PROBGAM to compute gamma probabilities.

Gumbel Distribution

The fitted density function is

$$p(x) = \frac{hv}{\sigma} e^{-(x-\mu)/\sigma} \exp\left(-e^{-(x-\mu)/\sigma}\right)$$

where

μ = location parameter
σ = scale parameter ($\sigma > 0$)
h = width of histogram interval
v = vertical scaling factor, and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

You can specify μ and σ with the MU= and SIGMA= *Gumbel-options*, respectively. By default, the procedure calculates maximum likelihood estimates for these parameters.

NOTE: The Gumbel distribution is also referred to as Type 1 extreme value distribution.

NOTE: The random variable X has Gumbel (Type 1 extreme value) distribution if and only if e^X has Weibull distribution and $\exp((X-\mu)/\sigma)$ has standard exponential distribution.

Inverse Gaussian Distribution

The fitted density function is

$$p(x) = \begin{cases} hv\left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left(-\frac{\lambda}{2\mu^2 x}(x-\mu)^2\right) & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

where

μ = location parameter ($\mu > 0$)
λ = shape parameter ($\lambda > 0$)
h = width of histogram interval
v = vertical scaling factor, and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

The location parameter μ has to be greater then zero. You can specify μ with the MU= *iGauss-option*. In addition, you can specify shape parameter λ with the LAMBDA= *iGauss-option*. By default, the procedure uses the sample mean for μ and calculates a maximum likelihood estimate for λ.

NOTE: The special case where $\mu = 1$ and $\lambda = \phi$ corresponds to the Wald distribution.

Lognormal Distribution

The fitted density function is

$$p(x) = \begin{cases} \frac{hv}{\sigma\sqrt{2\pi}(x-\theta)} \exp\left(-\frac{(\log(x-\theta)-\zeta)^2}{2\sigma^2}\right) & \text{for } x > \theta \\ 0 & \text{for } x \leq \theta \end{cases}$$

where

θ = threshold parameter
ζ = scale parameter ($-\infty < \zeta < \infty$)
σ = shape parameter ($\sigma > 0$)
h = width of histogram interval
v = vertical scaling factor, and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

The threshold parameter θ must be less than the minimum data value. You can specify θ with the THRESHOLD= *lognormal-option*. By default, $\theta = 0$. If you specify THETA=EST, a maximum likelihood estimate is computed for θ. You can specify ζ and σ with the SCALE= and SHAPE= *lognormal-options*, respectively. By default, the procedure calculates maximum likelihood estimates for these parameters.

NOTE: The lognormal distribution is also referred to as the S_L distribution in the Johnson system of distributions.

NOTE: This book uses σ to denote the shape parameter of the lognormal distribution, whereas σ is used to denote the scale parameter of the beta, exponential, gamma, Gumbel, inverse Gaussian, normal, generalized Pareto, power function, Rayleigh, and Weibull distributions. The use of σ to denote the lognormal shape parameter is based on the fact that $\frac{1}{\sigma}(\log(X-\theta)-\zeta)$ has a standard normal distribution if X is lognormally distributed.

Normal Distribution

The fitted density function is

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

where
μ = mean
σ = standard deviation ($\sigma > 0$)
h = width of histogram interval
v = vertical scaling factor, and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

You can specify μ and σ with the MU= and SIGMA= *normal-options*, respectively. By default, the procedure estimates μ with the sample mean and σ with the sample standard deviation.

You can use the DATA step function PROBIT to compute normal quantiles and the DATA step function PROBNORM to compute probabilities.

NOTE: The normal distribution is also referred to as the S_N distribution in the Johnson system of distributions.

Generalized Pareto Distribution

The fitted density function is

$$p(x) = \begin{cases} \frac{hv}{\sigma}(1-\alpha(x-\theta)/\sigma)^{1/\alpha-1} & \text{if } \alpha \neq 0 \\ \frac{hv}{\sigma}\exp(-(x-\theta)/\sigma) & \text{if } \alpha = 0 \end{cases}$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)
α = shape parameter
h = width of histogram interval
v = vertical scaling factor, and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

The support of the distribution is $x > \theta$ for $\alpha \leq 0$ and $\theta < x < \sigma/\alpha$ for $\alpha > 0$.

NOTE: Special cases of the generalized Pareto distribution with $\alpha = 0$ and $\alpha = 1$ correspond respectively to the exponential distribution with mean σ and uniform distribution on the interval (θ, σ).

The threshold parameter θ must be less than the minimum data value. You can specify θ with the THETA= *Pareto-option*. By default, $\theta = 0$. You can also specify α and σ with the ALPHA= and SIGMA= *Pareto-options*,respectively. By default, the procedure calculates maximum likelihood estimates for these parameters.

NOTE: Maximum likelihood estimation of the parameters works well if $\alpha < \frac{1}{2}$, but not otherwise. In this case the estimators are asymptotically normal and asymptotically efficient. The asymptotic normal distribution of the maximum likelihood estimates has mean (α, σ) and variance-covariance matrix

$$\frac{1}{n}\begin{pmatrix} (1-\alpha)^2 & \sigma(1-\alpha) \\ \sigma(1-\alpha) & 2\sigma^2(1-\alpha) \end{pmatrix}.$$

NOTE: If no local minimum is found in the region

$$\{\alpha < 0, \sigma > 0\} \cup \{0 < \alpha \leq 1, \sigma/\alpha > \max(X_i)\},$$

there is no maximum likelihood estimator. More details on how to find maximum likelihood estimators and suggested algorithm can be found in Grimshaw(1993).

Power Function Distribution

The fitted density function is

$$p(x) = \begin{cases} hv\frac{\alpha}{\sigma}\left(\frac{x-\theta}{\sigma}\right)^{\alpha-1} & \text{for } \theta < x < \theta+\sigma \\ 0 & \text{for } x \le \theta \text{ or } x \ge \theta+\sigma \end{cases}$$

where

θ = lower threshold parameter (lower endpoint parameter)
σ = scale parameter ($\sigma > 0$)
α = shape parameter ($\alpha > 0$)
h = width of histogram interval
v = vertical scaling factor, and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

NOTE: This notation is consistent with that of other distributions that you can fit with the HISTOGRAM statement. However, many texts, including Johnson, Kotz, and Balakrishnan (1995), write the density function of power function distribution as

$$p(x) = \begin{cases} \frac{p}{b-a}\left(\frac{x-a}{b-a}\right)^{p-1} & \text{for } a < x < b \\ 0 & \text{for } x \le a \text{ or } x \ge b \end{cases}$$

The two parameterizations are related as follows:

$$\sigma = b - a$$

$$\theta = a$$

$$\alpha = p$$

NOTE: The family of power function distributions is a subclass of beta distribution with density function

$$p(x) = \begin{cases} hv\frac{(x-\theta)^{\alpha-1}(\sigma+\theta-x)^{\beta-1}}{B(\alpha,\beta)\sigma^{(\alpha+\beta-1)}} & \text{for } \theta < x < \theta+\sigma \\ 0 & \text{for } x \le \theta \text{ or } x \ge \theta+\sigma \end{cases}$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ with parameter $\beta = 1$. Therefore, all properties and estimation procedures of beta distribution apply.

The range of the power function distribution is bounded below by a threshold parameter $\theta = a$ and above by $\theta + \sigma = b$. If you specify a fitted power function curve by using the POWER option, θ must be less than the minimum data value and $\theta + \sigma$ must be greater than the maximum data value. You can specify θ and σ with the THETA= and SIGMA= *power-options* in parentheses after the keyword POWER. By default, $\sigma = 1$ and $\theta = 0$. If you specify THETA=EST and SIGMA=EST, maximum likelihood estimates are computed for θ and σ. However, three-parameter maximum likelihood estimation does not always converge.

In addition, you can specify α with the ALPHA= *power-option*. By default, the procedure calculates a maximum likelihood estimate for α. For example, to fit a power function density curve to a set of data bounded below by 32 and above by 212 with maximum likelihood estimate for α, use the following statement:

```
histogram Length / power(theta=32 sigma=180);
```

Rayleigh Distribution

The fitted density function is

$$p(x) = \begin{cases} h\,v\,\frac{x-\theta}{\sigma^2} e^{-(x-\theta)^2/(2\sigma^2)} & \text{for } x \geq \theta \\ 0 & \text{for } x < \theta \end{cases}$$

where

θ = lower threshold parameter (lower endpoint parameter)
σ = scale parameter ($\sigma > 0$)
h = width of histogram interval
v = vertical scaling factor, and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

NOTE: The Rayleigh distribution is a Weibull distribution with density function

$$p(x) = \begin{cases} h\,v\,\frac{k}{\lambda}\left(\frac{x-\theta}{\lambda}\right)^{k-1} \exp(-(\frac{x-\theta}{\lambda})^k) & \text{for } x \geq \theta \\ 0 & \text{for } x < \theta \end{cases}$$

and with shape parameter $k = 2$ and scale parameter $\lambda = \sqrt{2}\sigma$.

The threshold parameter θ must be less than the minimum data value. You can specify θ with the THETA= *Rayleigh-option*. By default, $\theta = 0$. In addition you can specify σ with the SIGMA= *Rayleigh-option*. By default, the procedure calculates maximum likelihood estimate for σ.

For example, to fit a Rayleigh density curve to a set of data bounded below by 32 with maximum likelihood estimate for σ, use the following statement:

```
histogram Length / rayleigh(theta=32);
```

Johnson S_B Distribution

The fitted density function is

$$p(x) = \begin{cases} \frac{\delta h v}{\sigma\sqrt{2\pi}}\left[\left(\frac{x-\theta}{\sigma}\right)\left(1-\frac{x-\theta}{\sigma}\right)\right]^{-1} \times \\ \exp\left[-\frac{1}{2}\left(\gamma + \delta\log(\frac{x-\theta}{\theta+\sigma-x})\right)^2\right] & \text{for } \theta < x < \theta+\sigma \\ 0 & \text{for } x \leq \theta \text{ or } x \geq \theta+\sigma \end{cases}$$

where

θ = threshold parameter ($-\infty < \theta < \infty$)

σ = scale parameter ($\sigma > 0$)
δ = shape parameter ($\delta > 0$)
γ = shape parameter ($-\infty < \gamma < \infty$)
h = width of histogram interval
v = vertical scaling factor, and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

The S_B distribution is bounded below by the parameter θ and above by the value $\theta + \sigma$. The parameter θ must be less than the minimum data value. You can specify θ with the THETA= S_B*-option*, or you can request that θ be estimated with the THETA = EST S_B*-option*. The default value for θ is zero. The sum $\theta + \sigma$ must be greater than the maximum data value. The default value for σ is one. You can specify σ with the SIGMA= S_B*-option*, or you can request that σ be estimated with the SIGMA = EST S_B*-option*.

By default, the method of percentiles given by Slifker and Shapiro (1980) is used to estimate the parameters. This method is based on four data percentiles, denoted by x_{-3z}, x_{-z}, x_z, and x_{3z}, which correspond to the four equally spaced percentiles of a standard normal distribution, denoted by $-3z$, $-z$, z, and $3z$, under the transformation

$$z = \gamma + \delta \log\left(\frac{x - \theta}{\theta + \sigma - x}\right)$$

The default value of z is 0.524. The results of the fit are dependent on the choice of z, and you can specify other values with the FITINTERVAL= option (specified in parentheses after the SB option). If you use the method of percentiles, you should select a value of z that corresponds to percentiles which are critical to your application.

The following values are computed from the data percentiles:

$$\begin{aligned} m &= x_{3z} - x_z \\ n &= x_{-z} - x_{-3z} \\ p &= x_z - x_{-z} \end{aligned}$$

It was demonstrated by Slifker and Shapiro (1980) that

$\frac{mn}{p^2} > 1$ for any S_U distribution
$\frac{mn}{p^2} < 1$ for any S_B distribution
$\frac{mn}{p^2} = 1$ for any S_L (lognormal) distribution

A tolerance interval around one is used to discriminate among the three families with this ratio criterion. You can specify the tolerance with the FITTOLERANCE= option (specified in parentheses after the SB option). The default tolerance is 0.01. Assuming that the criterion satisfies the inequality

$$\frac{mn}{p^2} < 1 - \text{tolerance}$$

the parameters of the S_B distribution are computed using the explicit formulas derived by Slifker and Shapiro (1980).

If you specify FITMETHOD = MOMENTS (in parentheses after the SB option) the method of moments is used to estimate the parameters. If you specify FITMETHOD = MLE (in parentheses after the SB option) the method of maximum likelihood is used to estimate the parameters. Note that maximum likelihood estimates may not always exist. Refer to Bowman and Shenton (1983) for discussion of methods for fitting Johnson distributions.

Johnson S_U Distribution

The fitted density function is

$$p(x) = \begin{cases} \frac{\delta h v}{\sigma\sqrt{2\pi}} \frac{1}{\sqrt{1+((x-\theta)/\sigma)^2}} \times \\ \exp\left[-\frac{1}{2}\left(\gamma + \delta \sinh^{-1}\left(\frac{x-\theta}{\sigma}\right)\right)^2\right] & \text{for } x > \theta \\ 0 & \text{for } x \leq \theta \end{cases}$$

where

θ = location parameter ($-\infty < \theta < \infty$)
σ = scale parameter ($\sigma > 0$)
δ = shape parameter ($\delta > 0$)
γ = shape parameter ($-\infty < \gamma < \infty$)
h = width of histogram interval
v = vertical scaling factor, and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

You can specify the parameters with the THETA=, SIGMA=, DELTA=, and GAMMA= *S_U-options*, which are enclosed in parentheses after the SU option. If you do not specify these parameters, they are estimated.

By default, the method of percentiles given by Slifker and Shapiro (1980) is used to estimate the parameters. This method is based on four data percentiles, denoted by x_{-3z}, x_{-z}, x_z, and x_{3z}, which correspond to the four equally spaced percentiles of a standard normal distribution, denoted by $-3z$, $-z$, z, and $3z$, under the transformation

$$z = \gamma + \delta \sinh^{-1}\left(\frac{x-\theta}{\sigma}\right)$$

The default value of z is 0.524. The results of the fit are dependent on the choice of z, and you can specify other values with the FITINTERVAL= option (specified in parentheses after the SU option). If you use the method of percentiles, you should select a value of z that corresponds to percentiles which are critical to your application.

The following values are computed from the data percentiles:

$$\begin{aligned} m &= x_{3z} - x_z \\ n &= x_{-z} - x_{-3z} \\ p &= x_z - x_{-z} \end{aligned}$$

It was demonstrated by Slifker and Shapiro (1980) that

$$\begin{array}{ll} \frac{mn}{p^2} > 1 & \text{for any } S_U \text{ distribution} \\ \frac{mn}{p^2} < 1 & \text{for any } S_B \text{ distribution} \\ \frac{mn}{p^2} = 1 & \text{for any } S_L \text{ (lognormal) distribution} \end{array}$$

A tolerance interval around one is used to discriminate among the three families with this ratio criterion. You can specify the tolerance with the FITTOLERANCE= option (specified in parentheses after the SU option). The default tolerance is 0.01. Assuming that the criterion satisfies the inequality

$$\frac{mn}{p^2} > 1 + \text{tolerance}$$

the parameters of the S_U distribution are computed using the explicit formulas derived by Slifker and Shapiro (1980).

If you specify FITMETHOD = MOMENTS (in parentheses after the SU option) the method of moments is used to estimate the parameters. If you specify FITMETHOD = MLE (in parentheses after the SU option) the method of maximum likelihood is used to estimate the parameters. Note that maximum likelihood estimates may not always exist. Refer to Bowman and Shenton (1983) for discussion of methods for fitting Johnson distributions.

Weibull Distribution

The fitted density function is

$$p(x) = \begin{cases} \frac{chv}{\sigma}(\frac{x-\theta}{\sigma})^{c-1}\exp(-(\frac{x-\theta}{\sigma})^c) & \text{for } x > \theta \\ 0 & \text{for } x \le \theta \end{cases}$$

where
θ = threshold parameter
σ = scale parameter ($\sigma > 0$)
c = shape parameter ($c > 0$)
h = width of histogram interval
v = vertical scaling factor, and

$$v = \begin{cases} n & \text{the sample size, for VSCALE=COUNT} \\ 100 & \text{for VSCALE=PERCENT} \\ 1 & \text{for VSCALE=PROPORTION} \end{cases}$$

The threshold parameter θ must be less than the minimum data value. You can specify θ with the THRESHOLD= *Weibull-option*. By default, $\theta = 0$. If you specify THETA=EST, a maximum likelihood estimate is computed for θ. You can specify σ and c with the SCALE= and SHAPE= *Weibull-options*, respectively. By default, the procedure calculates maximum likelihood estimates for σ and c.

The exponential distribution is a special case of the Weibull distribution where c = 1.

Kernel Density Estimates

You can use the KERNEL option to superimpose kernel density estimates on histograms. Smoothing the data distribution with a kernel density estimate can be more effective than using a histogram to examine features that might be obscured by the choice of histogram bins or sampling variation. A kernel density estimate can also be more effective than a parametric curve fit when the process distribution is multimodal. See Example 6.12.

The general form of the kernel density estimator is

$$\hat{f}_\lambda(x) = \frac{1}{n\lambda}\sum_{i=1}^{n} K_0\left(\frac{x - x_i}{\lambda}\right)$$

where $K_0(\cdot)$ is a kernel function, λ is the bandwidth, n is the sample size, and x_i is the ith observation.

The KERNEL option provides three kernel functions (K_0): normal, quadratic, and triangular. You can specify the function with the K= *kernel-option* in parentheses after the KERNEL option. Values for the K= option are NORMAL, QUADRATIC, and TRIANGULAR (with aliases of N, Q, and T, respectively). By default, a normal kernel is used. The formulas for the kernel functions are

Normal	$K_0(t) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}t^2)$	for $-\infty < t < \infty$
Quadratic	$K_0(t) = \frac{3}{4}(1 - t^2)$	for $\lvert t\rvert \le 1$
Triangular	$K_0(t) = 1 - \lvert t\rvert$	for $\lvert t\rvert \le 1$

The value of λ, referred to as the bandwidth parameter, determines the degree of smoothness in the estimated density function. You specify λ indirectly by specifying a standardized bandwidth c with the C= *kernel-option*. If Q is the interquartile range, and n is the sample size, then c is related to λ by the formula

$$\lambda = cQn^{-\frac{1}{5}}$$

For a specific kernel function, the discrepancy between the density estimator $\hat{f}_\lambda(x)$ and the true density $f(x)$ is measured by the mean integrated square error (MISE):

$$\text{MISE}(\lambda) = \int_x \{E(\hat{f}_\lambda(x)) - f(x)\}^2 dx + \int_x var(\hat{f}_\lambda(x))dx$$

The MISE is the sum of the integrated squared bias and the variance. An approximate mean integrated square error (AMISE) is

$$\text{AMISE}(\lambda) = \frac{1}{4}\lambda^4\left(\int_t t^2K(t)dt\right)^2\int_x \left(f''(x)\right)^2 dx + \frac{1}{n\lambda}\int_t K(t)^2 dt$$

A bandwidth that minimizes AMISE can be derived by treating $f(x)$ as the normal density having parameters μ and σ estimated by the sample mean and standard deviation. If you do not specify a bandwidth parameter or if you specify C=MISE, the bandwidth that minimizes AMISE is used. The value of AMISE can be used to compare different density estimates. For each estimate, the bandwidth parameter c, the kernel function type, and the value of AMISE are reported in the SAS log.

The general kernel density estimates assume that the domain of the density to estimate can take on all values on a real line. However, sometimes the domain of a density is an interval bounded on one or both sides. For

example, if a variable Y is a measurement of only positive values, then the kernel density curve should be bounded so that it is zero for negative Y values.

The CAPABILITY procedure uses a reflection technique to create the bounded kernel density curve, as described in Silverman (1986, pp 30-31). It adds the reflections of kernel density that are outside the boundary to the bounded kernel estimates. The general form of the bounded kernel density estimator is computed by replacing $K_0\left(\frac{x-x_i}{\lambda}\right)$ in the original equation with

$$\left\{K_0\left(\frac{x-x_i}{\lambda}\right)+K_0\left(\frac{(x-x_l)+(x_i-x_l)}{\lambda}\right)+K_0\left(\frac{(x_u-x)+(x_u-x_i)}{\lambda}\right)\right\}$$

where x_l is the lower bound and x_u is the upper bound.

Without a lower bound, $x_l=\infty$ and $K_0\left(\frac{(x-x_l)+(x_i-x_l)}{\lambda}\right)$ equals zero. Similarly, without an upper bound, $x_u=\infty$ and $K_0\left(\frac{(x_u-x)+(x_u-x_i)}{\lambda}\right)$ equals zero.

When C=MISE is used with a bounded kernel density, the CAPABILITY procedure uses a bandwidth that minimizes the AMISE for its corresponding unbounded kernel.

Printed Output

If you request a fitted parametric distribution, printed output summarizing the fit is produced in addition to the graphical display. Figure 6.16 shows the printed output for a fitted lognormal distribution requested by the following statements:

```
proc capability data=Hang;
   spec target=14 lsl=13.95 usl=14.05;
   hist / lognormal(indices midpercents);
run;
```

Figure 6.16 Sample Summary of Fitted Distribution

The CAPABILITY Procedure
Fitted Lognormal Distribution for Width (Width in cm)

Parameters for Lognormal Distribution

Parameter	Symbol	Estimate
Threshold	Theta	0
Scale	Zeta	2.638966
Shape	Sigma	0.001497
Mean		13.99873
Std Dev		0.020952

Goodness-of-Fit Tests for Lognormal Distribution

Test		Statistic	DF		p Value
Kolmogorov-Smirnov	D	0.09148348		Pr > D	>0.150
Cramer-von Mises	W-Sq	0.05040427		Pr > W-Sq	>0.500
Anderson-Darling	A-Sq	0.33476355		Pr > A-Sq	>0.500
Chi-Square	Chi-Sq	2.87938822	3	Pr > Chi-Sq	0.411

Figure 6.16 *continued*

Percent Outside Specifications for Lognormal Distribution			
Lower Limit		**Upper Limit**	
LSL	13.950000	**USL**	14.050000
Obs Pct < LSL	2.000000	**Obs Pct > USL**	0
Est Pct < LSL	0.992170	**Est Pct > USL**	0.728125

Capability Indices Based on Lognormal Distribution	
Cp	0.795463
CPL	0.776822
CPU	0.814021
Cpk	0.776822
Cpm	0.792237

Histogram Bin Percents for Lognormal Distribution		
	Percent	
Bin Midpoint	**Observed**	**Estimated**
13.95	4.000	2.963
13.97	18.000	15.354
13.99	26.000	33.872
14.01	38.000	32.055
14.03	10.000	13.050
14.05	4.000	2.281

Quantiles for Lognormal Distribution		
	Quantile	
Percent	**Observed**	**Estimated**
1.0	13.9440	13.9501
5.0	13.9656	13.9643
10.0	13.9710	13.9719
25.0	13.9860	13.9846
50.0	14.0018	13.9987
75.0	14.0129	14.0129
90.0	14.0218	14.0256
95.0	14.0241	14.0332
99.0	14.0470	14.0475

The summary is organized into the following parts:

- Parameters
- Chi-Square Goodness-of-Fit Test
- EDF Goodness-of-Fit Tests

- Specifications
- Indices Using the Fitted Curve
- Histogram Intervals
- Quantiles

These parts are described in the sections that follow.

Parameters

This section lists the parameters for the fitted curve as well as the estimated mean and estimated standard deviation. See "Formulas for Fitted Curves" on page 336.

Chi-Square Goodness-of-Fit Test

The chi-square goodness-of-fit statistic for a fitted parametric distribution is computed as follows:

$$\chi^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

where

O_i = observed value in ith histogram interval
E_i = expected value in ith histogram interval
m = number of histogram intervals
p = number of estimated parameters

The degrees of freedom for the chi-square test is equal to $m - p - 1$. You can save the observed and expected interval values in the OUTFIT= data set discussed in "Output Data Sets" on page 355.

Note that empty intervals are not combined, and the range of intervals used to compute χ^2 begins with the first interval containing observations and ends with the final interval containing observations.

EDF Goodness-of-Fit Tests

When you fit a parametric distribution, the HISTOGRAM statement provides a series of goodness-of-fit tests based on the empirical distribution function (EDF). The EDF tests offer advantages over the chi-square goodness-of-fit test, including improved power and invariance with respect to the histogram midpoints. For a thorough discussion, refer to D'Agostino and Stephens (1986).

The empirical distribution function is defined for a set of n independent observations $X_1, \ldots, X_n$ with a common distribution function $F(x)$. Denote the observations ordered from smallest to largest as $X_{(1)}, \ldots, X_{(n)}$. The empirical distribution function, $F_n(x)$, is defined as

$$\begin{aligned} F_n(x) &= 0, \quad x < X_{(1)} \\ F_n(x) &= \tfrac{i}{n}, \quad X_{(i)} \le x < X_{(i+1)} \quad i = 1, \ldots, n-1 \\ F_n(x) &= 1, \quad X_{(n)} \le x \end{aligned}$$

Note that $F_n(x)$ is a step function that takes a step of height $\frac{1}{n}$ at each observation. This function estimates the distribution function $F(x)$. At any value x, $F_n(x)$ is the proportion of observations less than or equal to x,

while $F(x)$ is the probability of an observation less than or equal to x. EDF statistics measure the discrepancy between $F_n(x)$ and $F(x)$.

The computational formulas for the EDF statistics make use of the probability integral transformation $U = F(X)$. If $F(X)$ is the distribution function of X, the random variable U is uniformly distributed between 0 and 1.

Given n observations $X_{(1)}, \ldots, X_{(n)}$, the values $U_{(i)} = F(X_{(i)})$ are computed by applying the transformation, as shown in the following sections.

The HISTOGRAM statement provides three EDF tests:

- Kolmogorov-Smirnov
- Anderson-Darling
- Cramér-von Mises

These tests are based on various measures of the discrepancy between the empirical distribution function $F_n(x)$ and the proposed parametric cumulative distribution function $F(x)$.

The following sections provide formal definitions of the EDF statistics.

Kolmogorov-Smirnov Statistic The Kolmogorov-Smirnov statistic (D) is defined as

$$D = \sup_x |F_n(x) - F(x)|$$

The Kolmogorov-Smirnov statistic belongs to the supremum class of EDF statistics. This class of statistics is based on the largest vertical difference between $F(x)$ and $F_n(x)$.

The Kolmogorov-Smirnov statistic is computed as the maximum of D^+ and D^-, where D^+ is the largest vertical distance between the EDF and the distribution function when the EDF is greater than the distribution function, and D^- is the largest vertical distance when the EDF is less than the distribution function.

$$\begin{aligned} D^+ &= \max_i \left(\frac{i}{n} - U_{(i)} \right) \\ D^- &= \max_i \left(U_{(i)} - \frac{i-1}{n} \right) \\ D &= \max \left(D^+, D^- \right) \end{aligned}$$

Anderson-Darling Statistic The Anderson-Darling statistic and the Cramér-von Mises statistic belong to the quadratic class of EDF statistics. This class of statistics is based on the squared difference $(F_n(x) - F(x))^2$. Quadratic statistics have the following general form:

$$Q = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 \, \psi(x) dF(x)$$

The function $\psi(x)$ weights the squared difference $(F_n(x) - F(x))^2$.

The Anderson-Darling statistic (A^2) is defined as

$$A^2 = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 \left[F(x) \left(1 - F(x) \right) \right]^{-1} dF(x)$$

Here the weight function is $\psi(x) = [F(x)\,(1 - F(x))]^{-1}$.

The Anderson-Darling statistic is computed as

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\left[(2i-1)\log U_{(i)} + (2n+1-2i)\log\left(\{1-U_{(i)}\right)\right]$$

Cramér-von Mises Statistic The Cramér-von Mises statistic (W^2) is defined as

$$W^2 = n\int_{-\infty}^{+\infty}(F_n(x) - F(x))^2\,dF(x)$$

Here the weight function is $\psi(x) = 1$.

The Cramér-von Mises statistic is computed as

$$W^2 = \sum_{i=1}^{n}\left(U_{(i)} - \frac{2i-1}{2n}\right)^2 + \frac{1}{12n}$$

Probability Values for EDF Tests Once the EDF test statistics are computed, the associated probability values (p-values) must be calculated.

For the Gumbel, inverse Gaussian, generalized Pareto, and Rayleigh distributions, the procedure computes associated probability values (p-values) by resampling from the estimated distribution. It generates k random samples of size n, where k is specified by the EDFNSAMPLES= option and n is the number of observations in the original data. EDF test statistics are computed for each sample, and the p-value is the proportion of samples whose EDF statistic is greater than or equal to the statistic computed for the original data. You can use the EDFSEED= option to specify a seed value for generating the sample values.

For the beta, exponential, gamma, lognormal, normal, power function, and Weibull distributions, the CAPABILITY procedure uses internal tables of probability levels similar to those given by D'Agostino and Stephens (1986). If the value is between two probability levels, then linear interpolation is used to estimate the probability value. The probability value depends upon the parameters that are known and the parameters that are estimated for the distribution you are fitting. Table 6.23 summarizes different combinations of estimated parameters for which EDF tests are available.

Table 6.23 Availability of EDF Tests

Distribution	Parameters			Tests Available
	Threshold	Scale	Shape	
Beta	θ known	σ known	α, β known	all
	θ known	σ known	$\alpha, \beta < 5$ unknown	all
Exponential	θ known,	σ known		all
	θ known	σ unknown		all
	θ unknown	σ known		all
	θ unknown	σ unknown		all

Table 6.23 (*continued*)

Distribution	Parameters			Tests Available
	Threshold	Scale	Shape	
Gamma	θ known	σ known	α known	all
	θ known	σ unknown	α known	all
	θ known	σ known	α unknown	all
	θ known	σ unknown	$\alpha > 1$ unknown	all
	θ unknown	σ known	$\alpha > 1$ known	all
	θ unknown	σ unknown	$\alpha > 1$ known	all
	θ unknown	σ known	$\alpha > 1$ unknown	all
	θ unknown	σ unknown	$\alpha > 1$ unknown	all
Lognormal	θ known	ζ known	σ known	all
	θ known	ζ known	σ unknown	A^2 and W^2
	θ known	ζ unknown	σ known	A^2 and W^2
	θ known	ζ unknown	σ unknown	all
	θ unknown	ζ known	$\sigma < 3$ known	all
	θ unknown	ζ known	$\sigma < 3$ unknown	all
	θ unknown	ζ unknown	$\sigma < 3$ known	all
	θ unknown	ζ unknown	$\sigma < 3$ unknown	all
Normal	θ known	σ known		all
	θ known	σ unknown		A^2 and W^2
	θ unknown	σ known		A^2 and W^2
	θ unknown	σ unknown		all
Weibull	θ known	σ known	c known	all
	θ known	σ unknown	c known	A^2 and W^2
	θ known	σ known	c unknown	A^2 and W^2
	θ known	σ unknown	c unknown	A^2 and W^2
	θ unknown	σ known	$c > 2$ known	all
	θ unknown	σ unknown	$c > 2$ known	all
	θ unknown	σ known	$c > 2$ unknown	all
	θ unknown	σ unknown	$c > 2$ unknown	all

Specifications

This section is included in the summary only if you provide specification limits, and it tabulates the limits as well as the observed percentages and estimated percentages outside the limits.

The estimated percentages are computed only if fitted distributions are requested and are based on the probability that an observed value exceeds the specification limits, assuming the fitted distribution. The observed percentages are the percents of observations outside the specification limits.

Indices Using Fitted Curves

This section is included in the summary only if you specify the INDICES option in parentheses after a distribution option, as in the statements that produce Figure 6.16. Standard process capability indices, such as C_p and C_{pk}, are not appropriate if the data are not normally distributed. The INDICES option computes

generalizations of the standard indices by using the fact that for the normal distribution, 3σ is both the distance from the lower 0.135 percentile to the median (or mean) and the distance from the median (or mean) to the upper 99.865 percentile. These percentiles are estimated from the fitted distribution, and the appropriate percentile-to-median distances are substituted for 3σ in the standard formulas.

Writing T for the target, LSL and USL for the lower and upper specification limits, and P_α for the 100αth percentile, the generalized capability indices are as follows:

$$CPL = \frac{P_{0.5} - LSL}{P_{0.5} - P_{0.00135}}$$

$$CPU = \frac{USL - P_{0.5}}{P_{0.99865} - P_{0.5}}$$

$$C_p = \frac{USL - LSL}{P_{0.99865} - P_{0.00135}}$$

$$C_{pk} = \min\left(\frac{P_{0.5} - \mathrm{LSL}}{P_{0.5} - P_{0.00135}}, \frac{\mathrm{USL} - P_{0.5}}{P_{0.99865} - P_{0.5}}\right)$$

$$K = 2 \times \frac{\left|\frac{1}{2}(\mathrm{USL} + \mathrm{LSL}) - P_{0.5}\right|}{\mathrm{USL} - \mathrm{LSL}}$$

$$C_{pm} = \frac{\min\left(\frac{T-LSL}{P_{0.5}-P_{0.00135}}, \frac{USL-T}{P_{0.99865}-P_{0.5}}\right)}{\sqrt{1 + \left(\frac{\mu - T}{\sigma}\right)^2}}$$

If the data are normally distributed, these formulas reduce to the formulas for the standard capability indices, which are given in the section "Standard Capability Indices" on page 235.

The following guidelines apply to the use of generalized capability indices requested with the INDICES option:

- When you choose the family of parametric distributions for the fitted curve, consider whether an appropriate family can be derived from assumptions about the process.
- Whenever possible, examine the data distribution with a histogram, probability plot, or quantile-quantile plot.
- Apply goodness-of-fit tests to assess how well the parametric distribution models the data.
- Consider whether a generalized index has a meaningful practical interpretation in your application.

At the time of this writing, there is ongoing research concerning the application of generalized capability indices, and it is important to note that other approaches can be used with nonnormal data:

- Transform the data to normality, then compute and report standard capability indices on the transformed scale.
- Report the proportion of nonconforming output estimated from the fitted distribution.
- If it is not possible to adequately model the data distribution with a parametric density, smooth the data distribution with a kernel density estimate and simply report the proportion of nonconforming output.

Refer to Rodriguez and Bynum (1992) for additional discussion.

Histogram Intervals

This section is included in the summary only if you specify the MIDPERCENTS option in parentheses after the distribution option, as in the statements that produce Figure 6.16. This table lists the interval midpoints along with the observed and estimated percentages of the observations that lie in the interval. The estimated percentages are based on the fitted distribution.

In addition, you can specify the MIDPERCENTS option to request a table of interval midpoints with the observed percent of observations that lie in the interval. See the entry for the MIDPERCENTS option on page 323.

Quantiles

This table lists observed and estimated quantiles. You can use the PERCENTS= option to specify the list of quantiles to appear in this list. The list in Figure 6.16 is the default list. See the entry for the PERCENTS= option on page 328.

Output Data Sets

You can create two output data sets with the HISTOGRAM statement: the OUTFIT= data set and the OUTHISTOGRAM= data set. These data sets are described in the following sections.

OUTFIT= Data Set

The OUTFIT= data set contains the parameters of fitted density curves, information about chi-square and EDF goodness-of-fit tests, specification limit information, and capability indices based on the fitted distribution. Because you can specify multiple HISTOGRAM statements with the CAPABILITY procedure, you can create several OUTFIT= data sets. For each variable plotted with the HISTOGRAM statement, the OUTFIT= data set contains one observation for each fitted distribution requested in the HISTOGRAM statement. If you use a BY statement, the OUTFIT= data set contains several observations for each BY group (one observation for each variable and fitted density combination). ID variables are not saved in the OUTFIT= data set.

The OUTFIT= data set contains the variables listed in Table 6.24. By default, an OUTFIT= data set contains _MIDPT1_ and _MIDPTN_ variables, whose values identify histogram intervals by their midpoints. When the ENDPOINTS= or NENDPOINTS option is specified, intervals are identified by endpoint values instead. If the RTINCLUDE option is specified, the variables _MAXPT1_ and _MAXPTN_ contain upper endpoint values. Otherwise, the variables _MINPT1_ and _MINPTN_ contain lower endpoint values.

Table 6.24 Variables in the OUTFIT= Data Set

Variable	Description
ADASQ	Anderson-Darling EDF goodness-of-fit statistic
ADP	p-value for Anderson-Darling EDF goodness-of-fit test
CHISQ	chi-square goodness-of-fit statistic
_CP_	generalized capability index C_p based on the fitted curve
CPK	generalized capability index C_{pk} based on the fitted curve
CPL	generalized capability index CPL based on the fitted curve
CPM	generalized capability index C_{pm} based on the fitted curve
CPU	generalized capability index CPU based on the fitted curve
CURVE	name of fitted distribution (abbreviated to 8 characters)
CVMWSQ	Cramer-von Mises EDF goodness-of-fit statistic
CVMP	p-value for Cramer-von Mises EDF goodness-of-fit test
_DF_	degrees of freedom for chi-square goodness-of-fit test
ESTGTR	estimated percent of population greater than upper specification limit
ESTLSS	estimated percent of population less than lower specification limit
ESTSTD	estimated standard deviation
EXPECT	estimated mean
_K_	generalized capability index K based on the fitted curve
KSD	Kolmogorov-Smirnov EDF goodness-of-fit statistic
KSP	p-value for Kolmogorov-Smirnov EDF goodness-of-fit test
LOCATN	location parameter for fitted distribution. For the Gumbel, inverse Gaussian, and normal distributions, this is either the value of μ specified with the MU= option or the value estimated by the procedure. For all other distributions, this is either the value specified or estimated according to the THETA= option, or zero.
LSL	lower specification limit
MAXPT1	upper endpoint of first interval used to calculate the value of the chi-square statistic.
MAXPTN	upper endpoint of last interval used to calculate the value of the chi-square statistic.
MIDPT1	midpoint of first interval used to calculate the value of the chi-square statistic. This is the leftmost interval that contains at least one value of the variable.
MIDPTN	midpoint of last interval used to calculate the value of the chi-square statistic. This is the rightmost interval that contains at least one value of the variable.
MINPT1	lower endpoint of first interval used to calculate the value of the chi-square statistic.
MINPTN	lower endpoint of last interval used to calculate the value of the chi-square statistic.
OBSGTR	observed percent of data greater than upper specification limit
OBSLSS	observed percent of data less than the lower specification limit
PCHISQ	p-value for chi-square goodness-of-fit test

Table 6.24 *(continued)*

Variable	Description
SCALE	value of scale parameter for fitted distribution. For the lognormal distribution, this is the value of ζ specified or estimated according to the ZETA= option. For all other distributions, this is the value specified or estimated according to the SIGMA= option.
SHAPE1	value of shape parameter for fitted distribution. For the beta, gamma, generalized Pareto, and power function distributions, this is the value of α, either specified with the ALPHA= option or estimated by the procedure. For the lognormal distribution, this is the value of σ, either specified with the SIGMA= option or estimated by the procedure. For the Weibull distribution, this is the value of c, either specified with the C= option or estimated by the procedure. For the Johnson S_B and S_U distributions, this is the value of δ, either specified with the DELTA= option or estimated by the procedure. For distributions without a shape parameter (Gumbel, normal, exponential, and Rayleigh distributions), _SHAPE1_ is set to missing.
SHAPE2	value of shape parameter for fitted distribution. For the beta distribution, this is the value of β, either specified with the BETA= option or estimated by the procedure. For the Johnson S_B and S_U distributions, this is the value of γ, either specified with the GAMMA= option or estimated by the procedure. For all other distributions, _SHAPE2_ is set to missing.
TARGET	target value
USL	upper specification limit
VAR	variable name
WIDTH	width of histogram interval

OUTHISTOGRAM= Data Set

The OUTHISTOGRAM= data set contains information about histogram intervals. Because you can specify multiple HISTOGRAM statements with the CAPABILITY procedure, you can create multiple OUTHISTOGRAM= data sets.

The data set contains a group of observations for each variable plotted with the HISTOGRAM statement. The group contains an observation for each interval of the histogram, beginning with the leftmost interval that contains a value of the variable and ending with the rightmost interval that contains a value of the variable. These intervals will not necessarily coincide with the intervals displayed in the histogram because the histogram may be padded with empty intervals at either end. If you superimpose one or more fitted curves on the histogram, the OUTHISTOGRAM= data set contains multiple groups of observations for each variable (one group for each curve). If you use a BY statement, the OUTHISTOGRAM= data set contains groups of observations for each BY group. ID variables are not saved in the OUTHISTOGRAM= data set.

The OUTHISTOGRAM= data set contains the variables listed in Table 6.25. By default, an OUTHISTOGRAM= data set contains the _MIDPT_ variable, whose values identify histogram intervals by their midpoints. When the ENDPOINTS= or NENDPOINTS option is specified, intervals are identified by

endpoint values instead. If the RTINCLUDE option is specified, the _MAXPT_ variable contains an interval's upper endpoint value. Otherwise, the _MINPT_ variable contains the interval's lower endpoint value.

Table 6.25 Variables in the OUTHISTOGRAM= Data Set

Variable	Description
COUNT	number of variable values in histogram interval
CURVE	name of fitted distribution (if requested in HISTOGRAM statement)
EXPPCT	estimated percent of population in histogram interval determined from optional fitted distribution
MAXPT	upper endpoint of histogram interval
MIDPT	midpoint of histogram interval
MINPT	lower endpoint of histogram interval
OBSPCT	percent of variable values in histogram interval
VAR	variable name

OUTKERNEL= Output Data Set

An OUTKERNEL= data set contains information about kernel density estimates requested with the KERNEL option. Because you can specify multiple HISTOGRAM statements with the CAPABILITY procedure, you can create multiple OUTKERNEL= data sets.

An OUTKERNEL= data set contains a group of observations for each kernel density estimate requested with the HISTOGRAM statement. These observations span a range of analysis variable values recorded in the _VALUE_ variable. The procedure determines the increment between values, and therefore the number of observations in the group. The variable _DENSITY_ contains the kernel density calculated for the corresponding analysis variable value.

When a density curve is overlaid on a histogram, the curve is scaled so that the area under the curve equals the total area of the histogram bars. The scaled density values are saved in the variable _COUNT_, _PERCENT_, or _PROPORTION_, depending on the histogram's vertical axis scale, determined by the VSCALE= option. Only one of these variables appears in a given OUTKERNEL= data set.

Table 6.26 lists the variables in an OUTKERNEL= data set.

Table 6.26 Variables in the OUTKERNEL= Data Set

Variable	Description
_C_	standardized bandwidth parameter
COUNT	kernel density scaled for VSCALE=COUNT
DENSITY	kernel density
PERCENT	kernel density scaled for VSCALE=PERCENT (default)
PROPORTION	kernel density scaled for VSCALE=PROPORTION
TYPE	kernel function
VALUE	variable value at which kernel function is calculated
VAR	variable name

ODS Tables

The following table summarizes the ODS tables related to fitted distributions that you can request with the HISTOGRAM statement.

Table 6.27 ODS Tables Produced with the HISTOGRAM Statement

Table Name	Description	Option
Bins	histogram bins	MIDPERCENTS suboption with any distribution option, such as NORMAL(MIDPERCENTS)
FitIndices	capability indices computed from fitted distribution	INDICES suboption with any distribution option, such as LOGNORMAL(INDICES)
FitQuantiles	quantiles of fitted distribution	any distribution option such as NORMAL
GoodnessOfFit	goodness-of-fit tests for fitted distribution	any distribution option such as NORMAL
ParameterEstimates	parameter estimates for fitted distribution	any distribution option such as NORMAL
Specifications	percents outside specification limits based on empirical and fitted distributions	any distribution option such as NORMAL

ODS Graphics

Before you create ODS Graphics output, ODS Graphics must be enabled (for example, by using the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" (Chapter 21, *SAS/STAT User's Guide*).

The appearance of a graph produced with ODS Graphics is determined by the style associated with the ODS destination where the graph is produced. HISTOGRAM options used to control the appearance of traditional graphics are ignored for ODS Graphics output.

When ODS Graphics is in effect, the HISTOGRAM statement assigns a name to the graph it creates. You can use this name to reference the graph when using ODS. The name is listed in Table 6.28.

Table 6.28 ODS Graphics Produced by the HISTOGRAM Statement

ODS Graph Name	Plot Description
Histogram	histogram

See Chapter 4, "SAS/QC Graphics," for more information about ODS Graphics and other methods for producing charts.

SYMBOL and PATTERN Statement Options

In earlier releases of SAS/QC software, graphical features (such as colors and line types) of specification lines, histogram bars, and fitted curves were controlled with options in SYMBOL and PATTERN statements when producing traditional graphics. These options are still supported, although they have been superseded by options in the HISTOGRAM and SPEC statements. The following tables summarize the two sets of options. **NOTE:** These statements have no effect on ODS Graphics output.

Table 6.29 Graphical Enhancement of Histogram Outlines and Specification Lines

Feature	Statement and Options	Alternative Statement and Options
Outline of Histogram Bars	HISTOGRAM Statement	SYMBOL1 Statement
color	CBARLINE=*color*	C=*color*
width		W=*value*
Target Reference Line	SPEC Statement	SYMBOL1 Statement
position	TARGET=*value*	
color	CTARGET=*color*	C=*color*
line type	LTARGET=*linetype*	L=*linetype*
width	WTARGET=*value*	W=*value*
Lower Specification Line	SPEC Statement	SYMBOL2 Statement
position	LSL=*value*	
color	CLSL=*color*	C=*color*
line type	LLSL=*linetype*	L=*linetype*
width	WLSL=*value*	W=*value*
Upper Specification Line	SPEC Statement	SYMBOL3 Statement
position	USL=*value*	
color	CUSL=*color*	C=*color*
line type	LUSL=*linetype*	L=*linetype*
width	WUSL=*value*	W=*value*

Table 6.30 Graphical Enhancement of Areas Under Histograms and Curves

Area Under Histogram or Curve	Statement and Options	Alternative Statement and Options
Histogram or Curve	HISTOGRAM Statement	PATTERN1 Statement
pattern	PFILL=*pattern*	V=*pattern*
color	CFILL=*color*	C=*color*
Left of Lower Specification Limit	SPEC Statement	PATTERN2 Statement
pattern	PLEFT=*pattern*	V=*pattern*
color	CLEFT=*color*	C=*color*
Right of Upper Specification Limit	SPEC Statement	PATTERN3 Statement
pattern	PRIGHT=*pattern*	V=*pattern*
color	CRIGHT=*color*	C=*color*

Table 6.31 Graphical Enhancement of Fitted Curves

Feature	Statement and Options	Alternative Statement and Options
Normal Curve	Normal-options	SYMBOL4 Statement
color	COLOR=*color*	C=*color*
line type	L=*linetype*	L=*linetype*
width	W=*value*	W=*value*
Lognormal Curve	Lognormal-options	SYMBOL5 Statement
color	COLOR=*color*	C=*color*
line type	L=*linetype*	L=*linetype*
width	W=*value*	W=*value*
Exponential Curve	Exponential-options	SYMBOL6 Statement
color	COLOR=*color*	C=*color*
line type	L=*linetype*	L=*linetype*
width	W=*value*	W=*value*
Weibull Curve	Weibull-options	SYMBOL7 Statement
color	COLOR=*color*	C=*color*
line type	L=*linetype*	L=*linetype*
width	W=*value*	W=*value*
Gamma Curve	Gamma-options	SYMBOL8 Statement
color	COLOR=*color*	C=*color*
line type	L=*linetype*	L=*linetype*
width	W=*value*	W=*value*
Beta Curve	Beta-options	SYMBOL9 Statement
color	COLOR=*color*	C=*color*
line type	L=*linetype*	L=*linetype*
width	W=*value*	W=*value*
Johnson S_B Curve	S_B-options	SYMBOL10 Statement
color	COLOR=*color*	C=*color*
line type	L=*linetype*	L=*linetype*
width	W=*value*	W=*value*
Johnson S_U Curve	S_U-options	SYMBOL11 Statement
color	COLOR=*color*	C=*color*
line type	L=*linetype*	L=*linetype*
width	W=*value*	W=*value*
Rayleigh Curve	Rayleigh-options	SYMBOL12 Statement
color	COLOR=*color*	C=*color*
line type	L=*linetype*	L=*linetype*
width	W=*value*	W=*value*
Generalized Pareto Curve	Pareto-options	SYMBOL13 Statement
color	COLOR=*color*	C=*color*
line type	L=*linetype*	L=*linetype*
width	W=*value*	W=*value*

Table 6.31 (*continued*)

Feature	Statement and Options	Alternative Statement and Options
Gumbel Curve	Gumbel-options	SYMBOL14 Statement
color	COLOR=*color*	C=*color*
line type	L=*linetype*	L=*linetype*
width	W=*value*	W=*value*
Power Function Curve	Power-options	SYMBOL15 Statement
color	COLOR=*color*	C=*color*
line type	L=*linetype*	L=*linetype*
width	W=*value*	W=*value*
Inverse Gaussian Curve	IGauss-options	SYMBOL16 Statement
color	COLOR=*color*	C=*color*
line type	L=*linetype*	L=*linetype*
width	W=*value*	W=*value*

Examples: HISTOGRAM Statement

This section provides advanced examples of the HISTOGRAM statement.

Example 6.8: Fitting a Beta Curve

NOTE: See *Fitting a Beta Curve on a Histogram* in the SAS/QC Sample Library.

You can use a beta distribution to model the distribution of a quantity that is known to vary between lower and upper bounds. In this example, a manufacturing company uses a robotic arm to attach hinges on metal sheets. The attachment point should be offset 10.1 mm from the left edge of the sheet. The actual offset varies between 10.0 and 10.5 mm due to variation in the arm. Offsets for 50 attachment points are saved in the following data set:

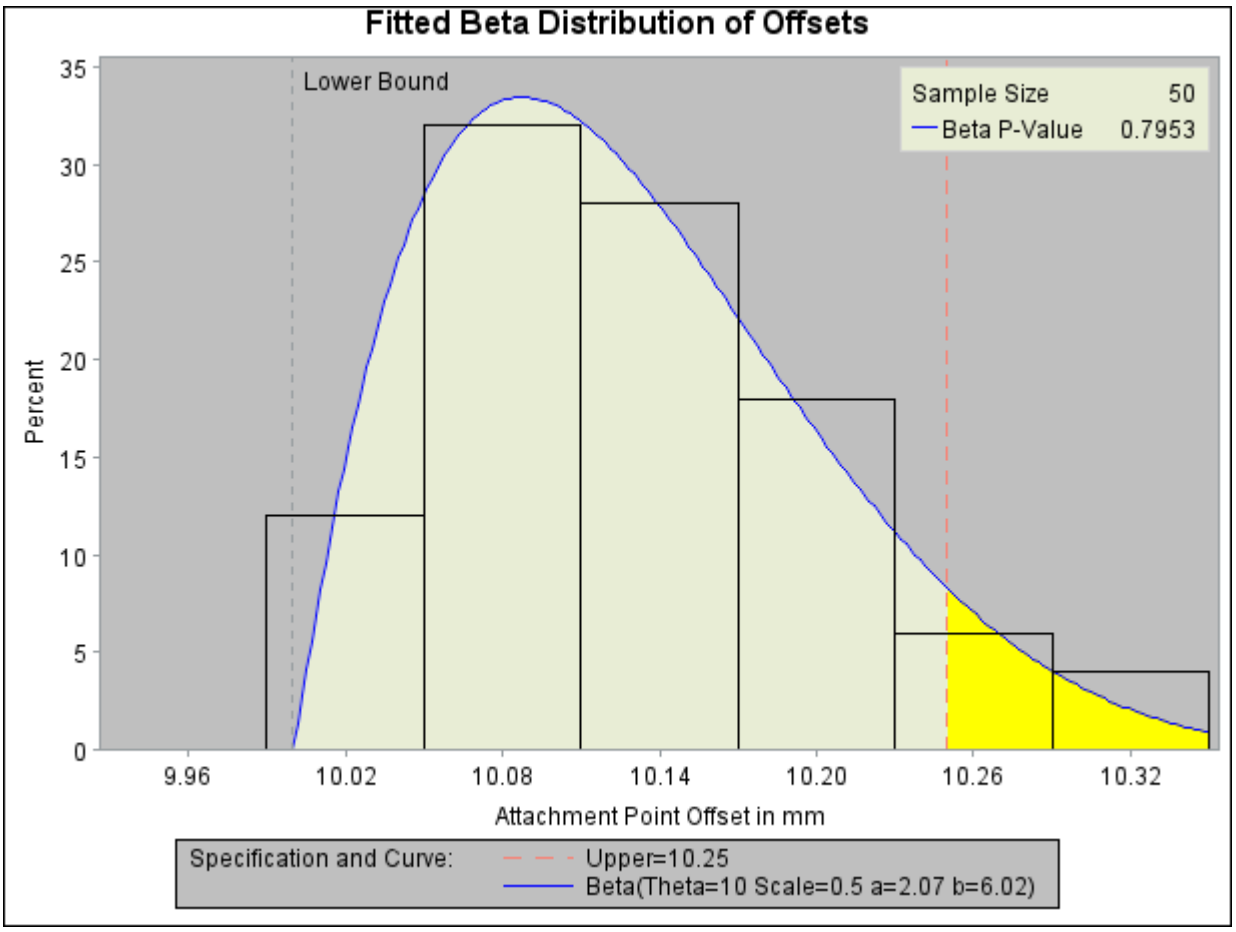
```
data Measures;
   input Length @@;
   label Length = 'Attachment Point Offset in mm';
   datalines;
10.147 10.070 10.032 10.042 10.102
10.034 10.143 10.278 10.114 10.127
10.122 10.018 10.271 10.293 10.136
10.240 10.205 10.186 10.186 10.080
10.158 10.114 10.018 10.201 10.065
10.061 10.133 10.153 10.201 10.109
10.122 10.139 10.090 10.136 10.066
10.074 10.175 10.052 10.059 10.077
10.211 10.122 10.031 10.322 10.187
10.094 10.067 10.094 10.051 10.174
;
```

The following statements create a histogram with a fitted beta density curve:

```
ods graphics off;
legend2 frame cframe=ligr cborder=black position=center;
title1 'Fitted Beta Distribution of Offsets';
proc capability data=Measures;
   specs usl=10.25 lusl=20 cusl=salmon cright=yellow pright=solid;
   histogram Length /
      beta(theta=10 scale=0.5 color=blue fill)
      cfill     = ywh
      cframe    = ligr
      href      = 10
      hreflabel = 'Lower Bound'
      lhref     = 2
      legend    = legend2
      vaxis     = axis1;
   axis1 label=(a=90 r=0);
   inset n = 'Sample Size'
         beta(pchisq = 'P-Value') / pos=ne  cfill=ywh;
run;
```

The histogram is shown in Output 6.8.1. The THETA= *beta-option* specifies the lower threshold. The SCALE= *beta-option* specifies the range between the lower threshold and the upper threshold (in this case, 0.5 mm). Note that in general, the default THETA= and SCALE= values are zero and one, respectively.

Output 6.8.1 Superimposing a Histogram with a Fitted Beta Curve



The FILL *beta-option* specifies that the area under the curve is to be filled with the CFILL= color. (If FILL were omitted, the CFILL= color would be used to fill the histogram bars instead.) The CRIGHT= option in the SPEC statement specifies the color under the curve to the right of the upper specification limit. If the CRIGHT= option were not specified, the entire area under the curve would be filled with the CFILL= color. When a lower specification limit is available, you can use the CLEFT= option in the SPEC statement to specify the color under the curve to the left of this limit.

The HREF= option draws a reference line at the lower bound, and the HREFLABEL= option adds the label *Lower Bound*. The option LHREF=2 specifies a dashed line type. The INSET statement adds an inset with the sample size and the *p*-value for a chi-square goodness-of-fit test.

In addition to displaying the beta curve, the BETA option summarizes the curve fit, as shown in Output 6.8.2. The output tabulates the parameters for the curve, the chi-square goodness-of-fit test whose *p*-value is shown in Output 6.8.1, the observed and estimated percents above the upper specification limit, and the observed and estimated quantiles. For instance, based on the beta model, the percent of offsets greater than the upper specification limit is 6.6%. For computational details, see the section "Formulas for Fitted Curves" on page 336.

Output 6.8.2 Summary of Fitted Beta Distribution

Fitted Beta Distribution of Offsets

The CAPABILITY Procedure
Fitted Beta Distribution for Length (Attachment Point Offset in mm)

Parameters for Beta Distribution

Parameter	Symbol	Estimate
Threshold	Theta	10
Scale	Sigma	0.5
Shape	Alpha	2.06832
Shape	Beta	6.022479
Mean		10.12782
Std Dev		0.072339

Goodness-of-Fit Tests for Beta Distribution

Test	Statistic		DF	p Value	
Chi-Square	**Chi-Sq**	1.02463588	3	**Pr > Chi-Sq**	0.795

Percent Outside Specifications for Beta Distribution

	Upper Limit
USL	10.250000
Obs Pct > USL	8.000000
Est Pct > USL	6.618103

Quantiles for Beta Distribution

	Quantile	
Percent	**Observed**	**Estimated**
1.0	10.0180	10.0124
5.0	10.0310	10.0285
10.0	10.0380	10.0416
25.0	10.0670	10.0718
50.0	10.1220	10.1174
75.0	10.1750	10.1735
90.0	10.2255	10.2292
95.0	10.2780	10.2630
99.0	10.3220	10.3237

Example 6.9: Fitting Lognormal, Weibull, and Gamma Curves

NOTE: See *Superimposing Fitted Curves on a Histogram* in the SAS/QC Sample Library.

To find an appropriate model for a process distribution, you should consider curves from several distribution families. As shown in this example, you can use the HISTOGRAM statement to fit more than one type of distribution and display the density curves on the same histogram.

The gap between two plates is measured (in cm) for each of 50 welded assemblies selected at random from the output of a welding process assumed to be in statistical control. The lower and upper specification limits for the gap are 0.3 cm and 0.8 cm, respectively. The measurements are saved in a data set named Plates.

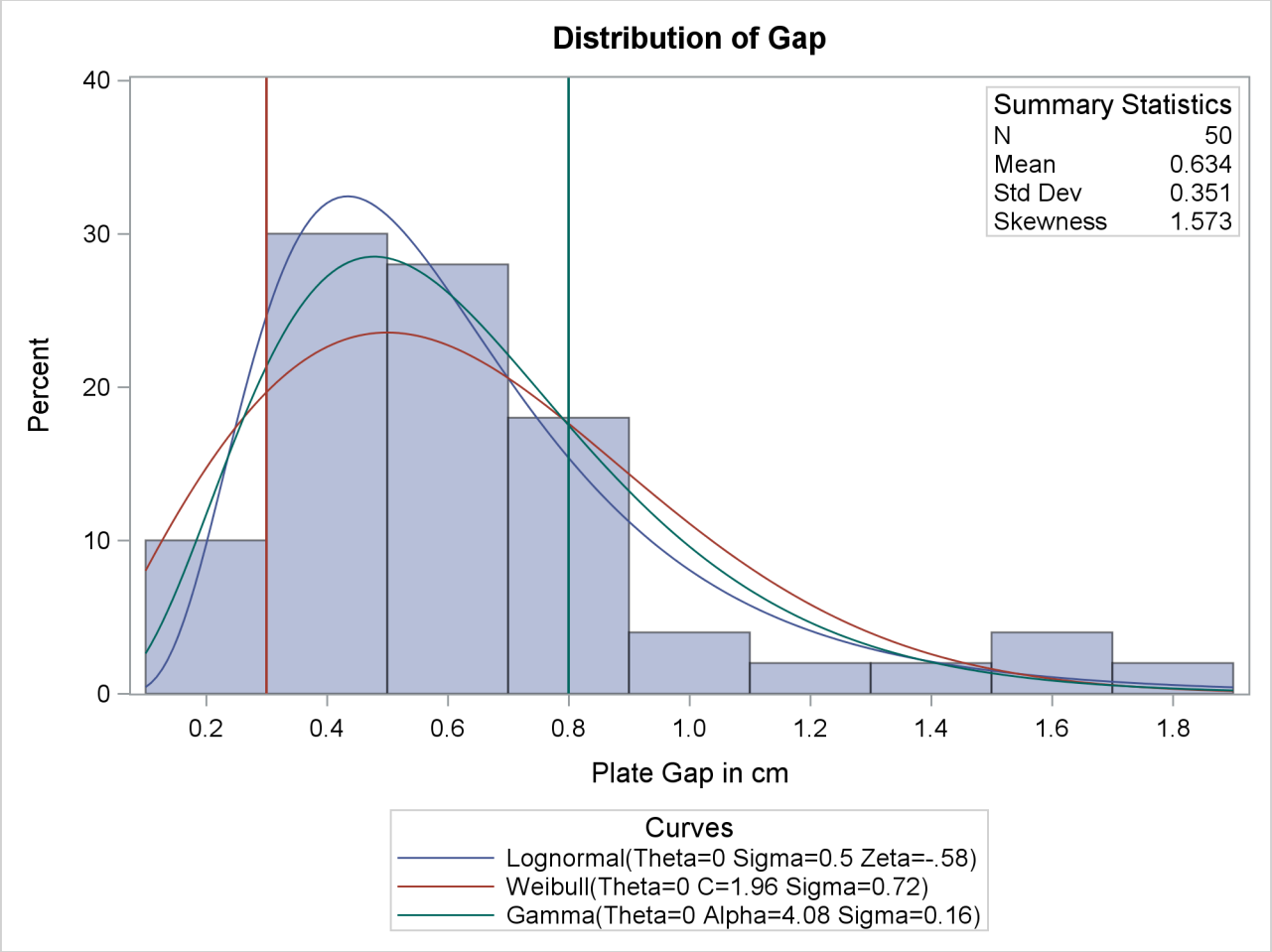
```
data Plates;
   label Gap='Plate Gap in cm';
   input Gap @@;
   datalines;
0.746  0.357  0.376  0.327  0.485 1.741  0.241  0.777  0.768  0.409
0.252  0.512  0.534  1.656  0.742 0.378  0.714  1.121  0.597  0.231
0.541  0.805  0.682  0.418  0.506 0.501  0.247  0.922  0.880  0.344
0.519  1.302  0.275  0.601  0.388 0.450  0.845  0.319  0.486  0.529
1.547  0.690  0.676  0.314  0.736 0.643  0.483  0.352  0.636  1.080
;
```

The following statements fit three distributions (lognormal, Weibull, and gamma) and display their density curves on a single histogram:

```
ods graphics on;
proc capability data=Plates;
   var Gap;
   specs  lsl = 0.3 usl = 0.8;
   histogram /
       midpoints=0.2 to 1.8 by 0.2
       lognormal
       weibull
       gamma
       nospeclegend;
   inset n mean(5.3) std='Std Dev'(5.3) skewness(5.3)
         / pos = ne  header = 'Summary Statistics';
run;
```

The LOGNORMAL, WEIBULL, and GAMMA options superimpose fitted curves on the histogram in Output 6.9.1. Note that a threshold parameter $\theta = 0$ is assumed for each curve. In applications where the threshold is not zero, you can specify θ with the THETA= option.

Output 6.9.1 Superimposing a Histogram with Fitted Curves



The LOGNORMAL, WEIBULL, and GAMMA options also produce the summaries for the fitted distributions shown in Output 6.9.2, Output 6.9.3, and Output 6.9.4.

Output 6.9.2 Summary of Fitted Lognormal Distribution

The CAPABILITY Procedure
Fitted Lognormal Distribution for Gap (Plate Gap in cm)

Parameters for Lognormal Distribution		
Parameter	**Symbol**	**Estimate**
Threshold	Theta	0
Scale	Zeta	-0.58375
Shape	Sigma	0.499546
Mean		0.631932
Std Dev		0.336436

Output 6.9.2 *continued*

Goodness-of-Fit Tests for Lognormal Distribution					
Test		**Statistic**	**DF**	**p Value**	
Kolmogorov-Smirnov	**D**	0.06441431		**Pr > D**	>0.150
Cramer-von Mises	**W-Sq**	0.02823022		**Pr > W-Sq**	>0.500
Anderson-Darling	**A-Sq**	0.24308402		**Pr > A-Sq**	>0.500
Chi-Square	**Chi-Sq**	7.51762213	6	**Pr > Chi-Sq**	0.276

Percent Outside Specifications for Lognormal Distribution			
Lower Limit		**Upper Limit**	
LSL	0.300000	**USL**	0.800000
Obs Pct < LSL	10.000000	**Obs Pct > USL**	20.000000
Est Pct < LSL	10.719540	**Est Pct > USL**	23.519008

Quantiles for Lognormal Distribution		
	Quantile	
Percent	**Observed**	**Estimated**
1.0	0.23100	0.17449
5.0	0.24700	0.24526
10.0	0.29450	0.29407
25.0	0.37800	0.39825
50.0	0.53150	0.55780
75.0	0.74600	0.78129
90.0	1.10050	1.05807
95.0	1.54700	1.26862
99.0	1.74100	1.78313

Output 6.9.2 provides four goodness-of-fit tests for the lognormal distribution: the chi-square test and three tests based on the EDF (Anderson-Darling, Cramer-von Mises, and Kolmogorov-Smirnov). See "Chi-Square Goodness-of-Fit Test" on page 350 and "EDF Goodness-of-Fit Tests" on page 350 for more information. The EDF tests are superior to the chi-square test because they are not dependent on the set of midpoints used for the histogram.

At the $\alpha = 0.10$ significance level, all four tests support the conclusion that the two-parameter lognormal distribution with scale parameter $\hat{\zeta} = -0.58$, and shape parameter $\hat{\sigma} = 0.50$ provides a good model for the distribution of plate gaps.

Output 6.9.3 Summary of Fitted Weibull Distribution

The CAPABILITY Procedure
Fitted Weibull Distribution for Gap (Plate Gap in cm)

Parameters for Weibull Distribution

Parameter	Symbol	Estimate
Threshold	Theta	0
Scale	Sigma	0.719208
Shape	C	1.961159
Mean		0.637641
Std Dev		0.339248

Goodness-of-Fit Tests for Weibull Distribution

Test		Statistic	DF		p Value
Cramer-von Mises	**W-Sq**	0.1593728		**Pr > W-Sq**	0.016
Anderson-Darling	**A-Sq**	1.1569354		**Pr > A-Sq**	<0.010
Chi-Square	**Chi-Sq**	15.0252997	6	**Pr > Chi-Sq**	0.020

Percent Outside Specifications for Weibull Distribution

Lower Limit		Upper Limit	
LSL	0.300000	**USL**	0.800000
Obs Pct < LSL	10.000000	**Obs Pct > USL**	20.000000
Est Pct < LSL	16.473319	**Est Pct > USL**	29.165543

Quantiles for Weibull Distribution

	Quantile	
Percent	Observed	Estimated
1.0	0.23100	0.06889
5.0	0.24700	0.15817
10.0	0.29450	0.22831
25.0	0.37800	0.38102
50.0	0.53150	0.59661
75.0	0.74600	0.84955
90.0	1.10050	1.10040
95.0	1.54700	1.25842
99.0	1.74100	1.56691

Output 6.9.3 provides two EDF goodness-of-fit tests for the Weibull distribution: the Anderson-Darling and the Cramer-von Mises tests. (See Table 6.23 for a complete list of the EDF tests available in the HISTOGRAM statement.) The probability values for the chi-square and EDF tests are all less than 0.10, indicating that the data do not support a Weibull model.

Output 6.9.4 Summary of Fitted Gamma Distribution

The CAPABILITY Procedure
Fitted Gamma Distribution for Gap (Plate Gap in cm)

Parameters for Gamma Distribution		
Parameter	**Symbol**	**Estimate**
Threshold	Theta	0
Scale	Sigma	0.155198
Shape	Alpha	4.082646
Mean		0.63362
Std Dev		0.313587

Goodness-of-Fit Tests for Gamma Distribution					
Test		**Statistic**	**DF**	**p Value**	
Kolmogorov-Smirnov	**D**	0.0969533		**Pr > D**	>0.250
Cramer-von Mises	**W-Sq**	0.0739847		**Pr > W-Sq**	>0.250
Anderson-Darling	**A-Sq**	0.5810661		**Pr > A-Sq**	0.137
Chi-Square	**Chi-Sq**	12.3075959	6	**Pr > Chi-Sq**	0.055

Percent Outside Specifications for Gamma Distribution			
Lower Limit		**Upper Limit**	
LSL	0.300000	**USL**	0.800000
Obs Pct < LSL	10.000000	**Obs Pct > USL**	20.000000
Est Pct < LSL	12.111039	**Est Pct > USL**	25.696522

Quantiles for Gamma Distribution		
	Quantile	
Percent	**Observed**	**Estimated**
1.0	0.23100	0.13326
5.0	0.24700	0.21951
10.0	0.29450	0.27938
25.0	0.37800	0.40404
50.0	0.53150	0.58271
75.0	0.74600	0.80804
90.0	1.10050	1.05392
95.0	1.54700	1.22160
99.0	1.74100	1.57939

Output 6.9.4 provides four goodness-of-fit tests for the gamma distribution. The probability value for the chi-square test is less than 0.10, indicating that the data do not support a gamma model.

Based on this analysis, the fitted lognormal distribution is the best model for the distribution of plate gaps. You can use this distribution to calculate useful quantities. For instance, you can compute the probability that the gap of a randomly sampled plate exceeds the upper specification limit, as follows:

$$\begin{aligned} \Pr[\text{gap} > \text{USL}] &= \Pr\left[Z > \tfrac{1}{\sigma}(\log(\text{USL} - \theta) - \zeta)\right] \\ &= 1 - \Phi\left[\tfrac{1}{\sigma}(\log(\text{USL} - \theta) - \zeta)\right] \end{aligned}$$

where Z has a standard normal distribution, and $\Phi(\cdot)$ is the standard normal cumulative distribution function. Note that $\Phi(\cdot)$ can be computed with the DATA step function PROBNORM. In this example, USL = 0.8 and $\Pr[\text{gap} > 0.8] = 0.2352$. This value is expressed as a percent (*Est Pct* > *USL*) in Output 6.9.2.

Example 6.10: Comparing Goodness-of-Fit Tests

NOTE: See *Comparing Goodness-of-Fit Tests* in the SAS/QC Sample Library.

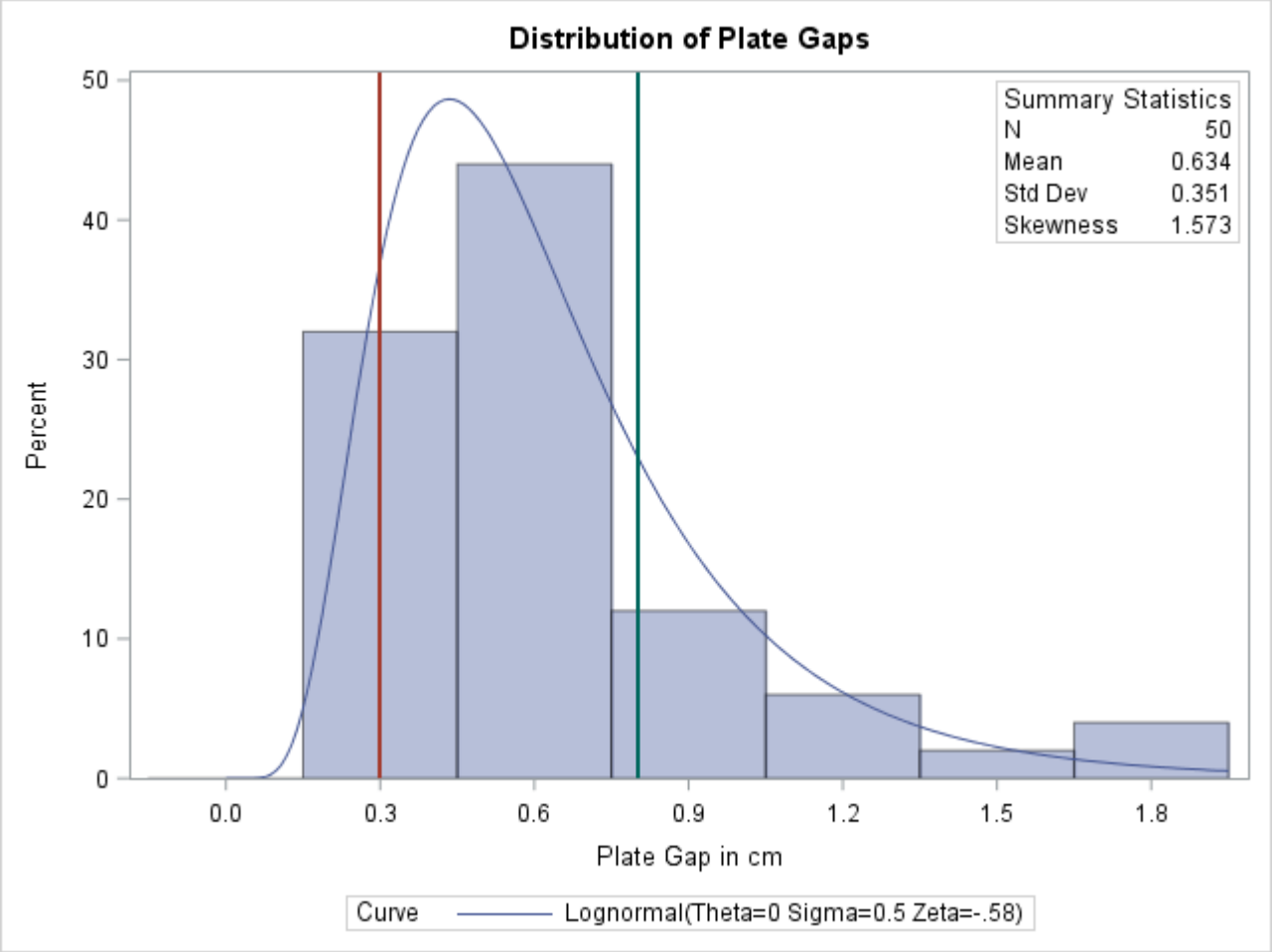
A weakness of the chi-square goodness-of-fit test is its dependence on the choice of histogram midpoints. An advantage of the EDF tests is that they give the same results regardless of the midpoints, as illustrated in this example.

In Example 6.9, the option MIDPOINTS=0.2 TO 1.8 BY 0.2 was used to specify the histogram midpoints for Gap. The following statements refit the lognormal distribution by using default midpoints (0.3 to 1.8 by 0.3).

```
data Plates;
   label Gap='Plate Gap in cm';
   input Gap @@;
   datalines;
0.746  0.357  0.376  0.327  0.485 1.741  0.241  0.777  0.768  0.409
0.252  0.512  0.534  1.656  0.742 0.378  0.714  1.121  0.597  0.231
0.541  0.805  0.682  0.418  0.506 0.501  0.247  0.922  0.880  0.344
0.519  1.302  0.275  0.601  0.388 0.450  0.845  0.319  0.486  0.529
1.547  0.690  0.676  0.314  0.736 0.643  0.483  0.352  0.636  1.080
;


title1 'Distribution of Plate Gaps';
proc capability data=Plates noprint;
   var Gap;
   specs  lsl = 0.3 usl = 0.8;
   histogram / lognormal
               nospeclegend
               odstitle  = title;
   inset n mean(5.3) std='Std Dev'(5.3) skewness(5.3) /
      pos    = ne
      header = 'Summary Statistics';
run;
```

The histogram is shown in Output 6.10.1.

Output 6.10.1 Lognormal Curve Fit with Default Midpoints



A summary of the lognormal fit is shown in Output 6.10.2. The *p*-value for the chi-square goodness-of-fit test is 0.082. Because this value is less than 0.10 (a typical cutoff level), the conclusion is that the lognormal distribution is not an appropriate model for the data. This is the *opposite* conclusion drawn from the chi-square test in Example 6.9, which is based on a different set of midpoints and has a *p*-value of 0.2756 (see Output 6.9.2). Moreover, the results of the EDF goodness-of-fit tests are the same because these tests do not depend on the midpoints. When available, the EDF tests provide more powerful alternatives to the chi-square test. For a thorough discussion of EDF tests, refer to D'Agostino and Stephens (1986).

Output 6.10.2 Printed Output for the Lognormal Curve

Distribution of Plate Gaps

The CAPABILITY Procedure
Fitted Lognormal Distribution for Gap (Plate Gap in cm)

Parameters for Lognormal Distribution		
Parameter	**Symbol**	**Estimate**
Threshold	Theta	0
Scale	Zeta	-0.58375
Shape	Sigma	0.499546
Mean		0.631932
Std Dev		0.336436

Goodness-of-Fit Tests for Lognormal Distribution					
Test		**Statistic**	**DF**	**p Value**	
Kolmogorov-Smirnov	**D**	0.06441431		**Pr > D**	>0.150
Cramer-von Mises	**W-Sq**	0.02823022		**Pr > W-Sq**	>0.500
Anderson-Darling	**A-Sq**	0.24308402		**Pr > A-Sq**	>0.500
Chi-Square	**Chi-Sq**	6.69789360	3	**Pr > Chi-Sq**	0.082

Example 6.11: Computing Capability Indices for Nonnormal Distributions

NOTE: See *Nonnormal Distribution Capability Indices* in the SAS/QC Sample Library.

Standard capability indices such as C_{pk} are generally considered meaningful only if the process output has a normal (or reasonably normal) distribution. In practice, however, many processes have nonnormal distributions. This example, which is a continuation of Example 6.9 and Example 6.10, shows how you can use the HISTOGRAM statement to compute generalized capability indices based on fitted nonnormal distributions.

The following statements produce printed output that is partially listed in Output 6.11.1 and Output 6.11.2:

```
data Plates;
   label Gap='Plate Gap in cm';
   input Gap @@;
   datalines;
0.746  0.357  0.376  0.327  0.485 1.741  0.241  0.777  0.768  0.409
0.252  0.512  0.534  1.656  0.742 0.378  0.714  1.121  0.597  0.231
0.541  0.805  0.682  0.418  0.506 0.501  0.247  0.922  0.880  0.344
0.519  1.302  0.275  0.601  0.388 0.450  0.845  0.319  0.486  0.529
1.547  0.690  0.676  0.314  0.736 0.643  0.483  0.352  0.636  1.080
;


proc capability data=Plates checkindices(alpha=0.05);
   specs lsl=0.3 usl= 0.8;
   histogram Gap / lognormal(indices) noplot;
run;
```

The PROC CAPABILITY statement computes the standard capability indices that are shown in Output 6.11.1.

Output 6.11.1 Standard Capability Indices for Variable Gap

Distribution of Plate Gaps

The CAPABILITY Procedure
Variable: Gap (Plate Gap in cm)

Process Capability Indices			
Index	**Value**	**95% Confidence Limits**	
Cp	0.237112	0.190279	0.283853
CPL	0.316422	0.203760	0.426833
CPU	0.157803	0.059572	0.254586
Cpk	0.157803	0.060270	0.255336

Warning: Normality is rejected for alpha = 0.05 using the Shapiro-Wilk test

The CHECKINDICES option in the PROC statement requests a goodness-of-fit test for normality in conjunction with the indices and displays the warning that normality is rejected at the significance level $\alpha = 0.05$.

Example 6.9 concluded that the fitted lognormal distribution summarized in Output 6.9.2 is a good model, so one might consider computing generalized capability indices based on this distribution. These indices are requested with the INDICES option and are shown in Output 6.11.2. Formulas and recommendations for these indices are given in "Indices Using Fitted Curves" on page 353.

Output 6.11.2 Fitted Lognormal Distribution Information

Capability Indices Based on Lognormal Distribution	
Cp	0.210804
CPL	0.595156
CPU	0.124927
Cpk	0.124927

Example 6.12: Computing Kernel Density Estimates

NOTE: See *Superimposing Kernel Density Estimates* in the SAS/QC Sample Library.

This example illustrates the use of kernel density estimates to visualize a nonnormal data distribution.

The effective channel length (in microns) is measured for 1225 field effect transistors. The channel lengths are saved as values of the variable Length in a SAS data set named Channel:

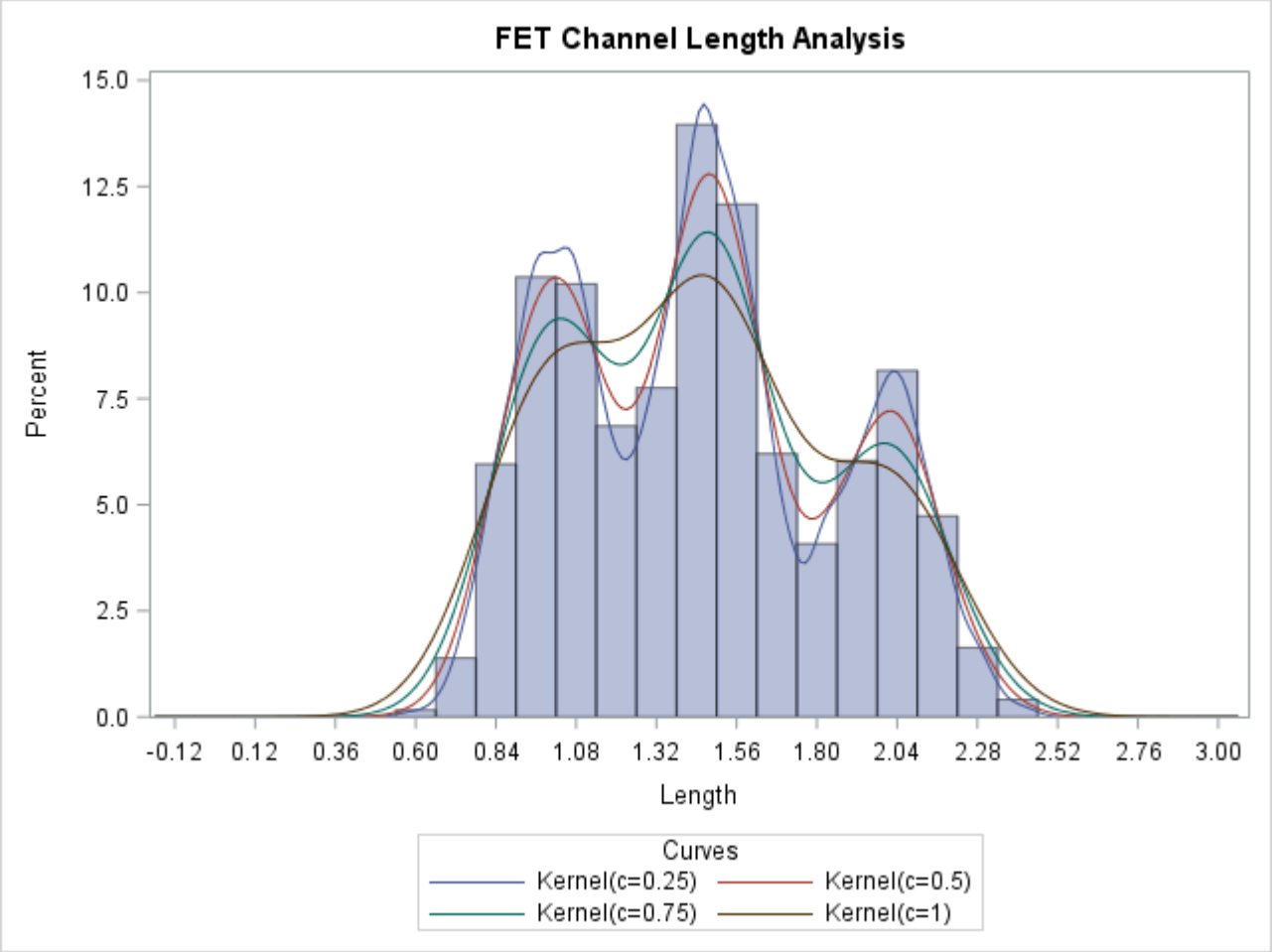
```
data Channel;
   length Lot $ 16;
   input Length @@;
   select;
      when (_n_ <= 425) Lot='Lot 1';
      when (_n_ >= 926) Lot='Lot 3';
      otherwise Lot='Lot 2';
   end;
   datalines;
0.91 1.01 0.95 1.13 1.12 0.86 0.96 1.17 1.36 1.10
0.98 1.27 1.13 0.92 1.15 1.26 1.14 0.88 1.03 1.00
0.98 0.94 1.09 0.92 1.10 0.95 1.05 1.05 1.11 1.15
1.11 0.98 0.78 1.09 0.94 1.05 0.89 1.16 0.88 1.19
1.01 1.08 1.19 0.94 0.92 1.27 0.90 0.88 1.38 1.02

   ... more lines ...

2.13 2.05 1.90 2.07 2.15 1.96 2.15 1.89 2.15 2.04
1.95 1.93 2.22 1.74 1.91
;
```

When you use kernel density estimates to explore a data distribution, you should try several choices for the bandwidth parameter c because this determines the smoothness and closeness of the fit. You can specify a list of C= values with the KERNEL option to request multiple density estimates, as shown in the following statements:

```
title "FET Channel Length Analysis";
proc capability data=Channel noprint;
   histogram Length / kernel(c = 0.25 0.50 0.75 1.00)
                      odstitle = title;
run;
```

The display, shown in Output 6.12.1, demonstrates the effect of c. In general, larger values of c yield smoother density estimates, and smaller values yield estimates that more closely fit the data distribution.

Output 6.12.1 Multiple Kernel Density Estimates



Output 6.12.1 reveals strong trimodality in the data, which are explored further in "Creating a One-Way Comparative Histogram" on page 276.

Example 6.13: Fitting a Three-Parameter Lognormal Curve

NOTE: See *Three-Parameter Lognormal Distribution* in the SAS/QC Sample Library.

If you request a lognormal fit with the LOGNORMAL option, a *two-parameter* lognormal distribution is assumed. This means that the shape parameter σ and the scale parameter ζ are unknown (unless specified) and that the threshold θ is known (it is either specified with the THETA= option or assumed to be zero).

If it is necessary to estimate θ in addition to ζ and σ, the distribution is referred to as a *three-parameter* lognormal distribution. The equation for this distribution is the same as the equation given in section "Lognormal Distribution" on page 340 but the method of maximum likelihood must be modified. This example shows how you can request a three-parameter lognormal distribution.

A manufacturing process (assumed to be in statistical control) produces a plastic laminate whose strength must exceed a minimum of 25 psi. Samples are tested, and a lognormal distribution is observed for the strengths. It is important to estimate θ to determine whether the process is capable of meeting the strength requirement. The strengths for 49 samples are saved in the following data set:

```
data Plastic;
   label Strength='Strength in psi';
   input Strength @@;
   datalines;
30.26 31.23 71.96 47.39 33.93 76.15 42.21
81.37 78.48 72.65 61.63 34.90 24.83 68.93
43.27 41.76 57.24 23.80 34.03 33.38 21.87
31.29 32.48 51.54 44.06 42.66 47.98 33.73
25.80 29.95 60.89 55.33 39.44 34.50 73.51
43.41 54.67 99.43 50.76 48.81 31.86 33.88
35.57 60.41 54.92 35.66 59.30 41.96 45.32
;
```

The following statements use the LOGNORMAL option in the HISTOGRAM statement to display the fitted three-parameter lognormal curve shown in Output 6.13.1:

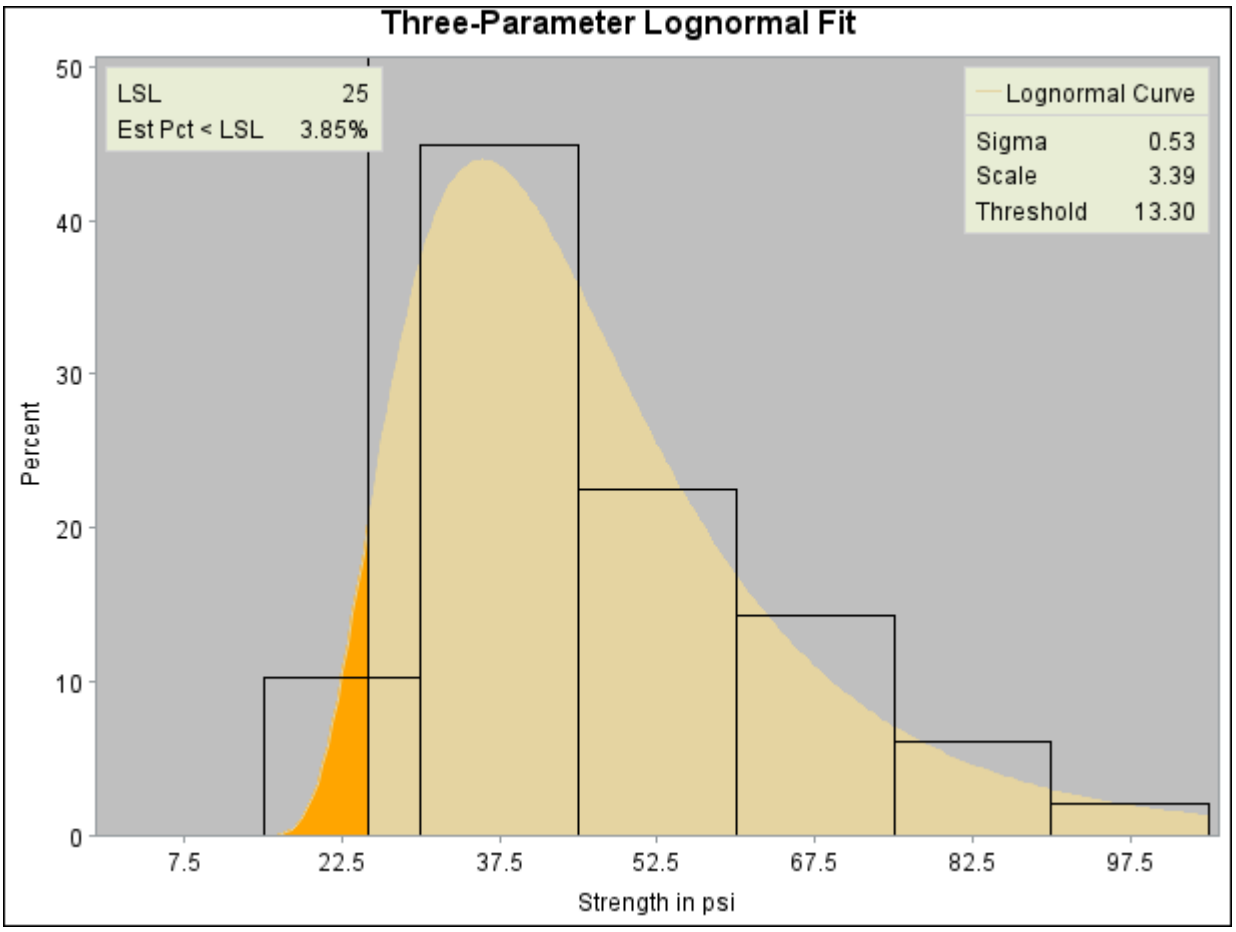
```
ods graphics off;
title 'Three-Parameter Lognormal Fit';
proc capability data=Plastic noprint;
   spec lsl=25 cleft=orange clsl=black;
   histogram Strength / lognormal(fill color = paoy
                                  theta = est)
                        cfill  = paoy
                        cframe = ligr
                        nolegend;
   inset lsl='LSL' lslpct / cfill = ywh pos=nw;
   inset lognormal        / format=6.2 pos=ne cfill = ywh;
run;
```

Specifying THETA=EST requests a *local* maximum likelihood estimate (LMLE) for θ, as described by Cohen (1951). This estimate is then used to compute maximum likelihood estimates for σ and ζ. The sample program CAPL3A illustrates a similar computational method implemented as a SAS/IML program.

NOTE: See *Three-Parameter Weibull Distribution* in the SAS/QC Sample Library.

Note that you can specify THETA=EST as a *Weibull-option* to fit a three-parameter Weibull distribution.

Output 6.13.1 Three-Parameter Lognormal Fit



Example 6.14: Annotating a Folded Normal Curve

NOTE: See *Cpk for Folded Normal Distribution* in the SAS/QC Sample Library.

This example shows how to display a fitted curve that is not supported by the HISTOGRAM statement.

The offset of an attachment point is measured (in mm) for a number of manufactured assemblies, and the measurements are saved in a data set named Assembly.

```
data Assembly;
   label Offset = 'Offset (in mm)';
   input Offset @@;
   datalines;
11.11 13.07 11.42  3.92 11.08  5.40 11.22 14.69  6.27  9.76
 9.18  5.07  3.51 16.65 14.10  9.69 16.61  5.67  2.89  8.13
 9.97  3.28 13.03 13.78  3.13  9.53  4.58  7.94 13.51 11.43
11.98  3.90  7.67  4.32 12.69  6.17 11.48  2.82 20.42  1.01
 3.18  6.02  6.63  1.72  2.42 11.32 16.49  1.22  9.13  3.34
 1.29  1.70  0.65  2.62  2.04 11.08 18.85 11.94  8.34  2.07
 0.31  8.91 13.62 14.94  4.83 16.84  7.09  3.37  0.49 15.19
```

```
 5.16  4.14  1.92 12.70  1.97  2.10  9.38  3.18  4.18  7.22
15.84 10.85  2.35  1.93  9.19  1.39 11.40 12.20 16.07  9.23
 0.05  2.15  1.95  4.39  0.48 10.16  4.81  8.28  5.68 22.81
 0.23  0.38 12.71  0.06 10.11 18.38  5.53  9.36  9.32  3.63
12.93 10.39  2.05 15.49  8.12  9.52  7.77 10.70  6.37  1.91
 8.60 22.22  1.74  5.84 12.90 13.06  5.08  2.09  6.41  1.40
15.60  2.36  3.97  6.17  0.62  8.56  9.36 10.19  7.16  2.37
12.91  0.95  0.89  3.82  7.86  5.33 12.92  2.64  7.92 14.06
;
```

The assembly process is in statistical control, and it is decided to fit a *folded normal distribution* to the offset measurements. A variable X has a folded normal distribution if $X = |Y|$, where Y is distributed as $N(\mu, \sigma)$. The fitted density is

$$h(x) = \frac{1}{\sqrt{2\pi}\sigma}\left[\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) + \exp\left(-\frac{(x+\mu)^2}{2\sigma^2}\right)\right] \quad , \; x \geq 0$$

You can use SAS/IML software to compute preliminary estimates of μ and σ based on a method of moments given by Elandt (1961). These estimates are computed by solving equation (19) of Elandt (1961), which is given by

$$f(\theta) = \frac{\left(\frac{2}{\sqrt{2\pi}}e^{-\theta^2/2} - \theta\left[1 - 2\Phi(\theta)\right]\right)^2}{1+\theta^2} = A$$

where $\Phi(\cdot)$ is the standard normal distribution function, and

$$A = \frac{\bar{x}^2}{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

Then the estimates of σ and μ are given by

$$\begin{aligned} \hat{\sigma}_0 &= \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2}{1+\hat{\theta}^2}} \\ \hat{\mu}_0 &= \hat{\theta} \cdot \hat{\sigma}_0 \end{aligned}$$

Begin by using the MEANS procedure to compute the first and second moments and using the DATA step to compute the constant A.

```
proc means data = Assembly noprint;
   var Offset;
   output out=Stat mean=m1 var=var n=n min = min;
run;

* Compute constant A from equation (19) of Elandt (1961) ;
data Stat;
   keep m2 a min;
   set Stat;
   a  = (m1*m1);
   m2 = ((n-1)/n)*var + a;
   a  = a/m2;
run;
```

Next, use the SAS/IML subroutine NLPDD to solve equation (19) by minimizing $(f(\theta) - A)^2$, and compute $\hat{\mu}_0$ and $\hat{\sigma}_0$.

```
proc iml;
   use Stat;
   read all var {m2}  into m2;
   read all var {a}   into a;
   read all var {min} into min;

   * f(t) is the function in equation (19) of Elandt (1961) ;
   start f(t) global(a);
      y = .39894*exp(-0.5*t*t);
      y = (2*y-(t*(1-2*probnorm(t))))**2/(1+t*t);
      y = (y-a)**2;
      return(y);
   finish;

   * Minimize (f(t)-A)**2 and estimate mu and sigma ;
   if ( min < 0 ) then do;
      print "Warning: Observations are not all nonnegative.";
      print "     The folded normal is inappropriate.";
      stop;
   end;
   if ( a < 0.637 ) then do;
      print "Warning: the folded normal may be inappropriate";
   end;
   opt = { 0 0 };
   con = { 1e-6 };
   x0  = { 2.0 };
   tc  = { . . . . . 1e-12 . . . . . .};
   call nlpdd(rc,etheta0,"f",x0,opt,con,tc);
   esig0 = sqrt(m2/(1+etheta0*etheta0));
   emu0  = etheta0*esig0;

   create Prelim var {emu0 esig0 etheta0};
   append;
   close Prelim;
      * Define the log likelihood of the folded normal ;
      start g(p) global(x);
         y = 0.0;
         do i = 1 to nrow(x);
            z = exp( (-0.5/p[2])*(x[i]-p[1])*(x[i]-p[1]) );
            z = z + exp( (-0.5/p[2])*(x[i]+p[1])*(x[i]+p[1]) );
            y = y + log(z);
         end;
         y = y - nrow(x)*log( sqrt( p[2] ) );
         return(y);
      finish;
      * Maximize the log likelihood with subroutine NLPDD ;
      use Assembly;
      read all var {Offset} into x;
      esig0sq = esig0*esig0;
      x0      = emu0||esig0sq;
      opt     = { 1 0 };
      con     = { . 0.0, .  .  };
      call nlpdd(rc,xr,"g",x0,opt,con);
```

```
      emu     = xr[1];
      esig    = sqrt(xr[2]);
      etheta  = emu/esig;
      create Parmest var{emu esig etheta};
      append;
      close Parmest;
   quit;

title 'The Data Set Prelim';
proc print data=Prelim noobs;
run;
```

The preliminary estimates are saved in the data set Prelim, as shown in Output 6.14.1.

Output 6.14.1 Preliminary Estimates of μ, σ, and θ

The Data Set Prelim

EMU0	ESIG0	ETHETA0
6.51735	6.54953	0.99509

Now, using $\hat{\mu}_0$ and $\hat{\sigma}_0$ as initial estimates, call the NLPDD subroutine to maximize the log likelihood, $l(\mu, \sigma)$, of the folded normal distribution, where, up to a constant,

$$l(\mu,\sigma) = -n\log\sigma + \sum_{i=1}^{n}\log\left[\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) + \exp\left(-\frac{(x_i+\mu)^2}{2\sigma^2}\right)\right]$$

```
* Define the log likelihood of the folded normal ;
start g(p) global(x);
   y = 0.0;
   do i = 1 to nrow(x);
      z = exp( (-0.5/p[2])*(x[i]-p[1])*(x[i]-p[1]) );
      z = z + exp( (-0.5/p[2])*(x[i]+p[1])*(x[i]+p[1]) );
      y = y + log(z);
   end;
   y = y - nrow(x)*log( sqrt( p[2] ) );
   return(y);
finish;

* Maximize the log likelihood with subroutine NLPDD ;
use assembly;
read all var {offset} into x;
esig0sq = esig0*esig0;
x0      = emu0||esig0sq;
opt     = { 1 0 };
con     = { . 0.0, .  .  };
call nlpdd(rc,xr,"g",x0,opt,con);
emu     = xr[1];
esig    = sqrt(xr[2]);
etheta  = emu/esig;

create parmest var{emu esig etheta};
append;
```

```
   close parmest;
quit;

title 'The Data Set PARMEST';
proc print data=Parmest noobs;
   var emu esig etheta;
run;
```

The data set Parmest saves the maximum likelihood estimates $\hat{\mu}$ and $\hat{\sigma}$ (as well as $\hat{\mu}/\hat{\sigma}$), as shown in Output 6.14.2.

Output 6.14.2 Final Estimates of μ, σ, and θ

The Data Set PARMEST

EMU	ESIG	ETHETA
6.66761	6.39650	1.04239

To annotate the curve on a histogram, begin by computing the width and endpoints of the histogram intervals. The following statements save these values in an OUTFIT= data set called OUT. Note that a plot is not produced at this point.

```
ods graphics off;
proc capability data = Assembly noprint;
   histogram Offset / outfit = Out normal(noprint) noplot;
run;

title 'OUTFIT= Data Set Out';
proc print data=Out noobs round;
   var _var_ _curve_ _locatn_ _scale_ _chisq_ _df_ _pchisq_
       _midpt1_ _width_ _midptn_ _expect_ _eststd_ _adasq_
       _adp_ _cvmwsq_ _cvmp_ _ksd_ _ksp_;
run;
```

Output 6.14.3 provides a partial listing of the data set Out. The width and endpoints of the histogram bars are saved as values of the variables _WIDTH_, _MIDPT1_, and _MIDPTN_. See "Output Data Sets" on page 355.

Output 6.14.3 The OUTFIT= Data Set Out

OUTFIT= Data Set Out

VAR	_CURVE_	_LOCATN_	_SCALE_	_CHISQ_	_DF_	_PCHISQ_	_MIDPT1_	_WIDTH_	_MIDPTN_
Offset	NORMAL	7.62	5.24	31.17	5	0	1.5	3	22.5

EXPECT	_ESTSTD_	_ADASQ_	_ADP_	_CVMWSQ_	_CVMP_	_KSD_	_KSP_
7.62	5.24	1.9	0.01	0.28	0.01	0.09	0.01

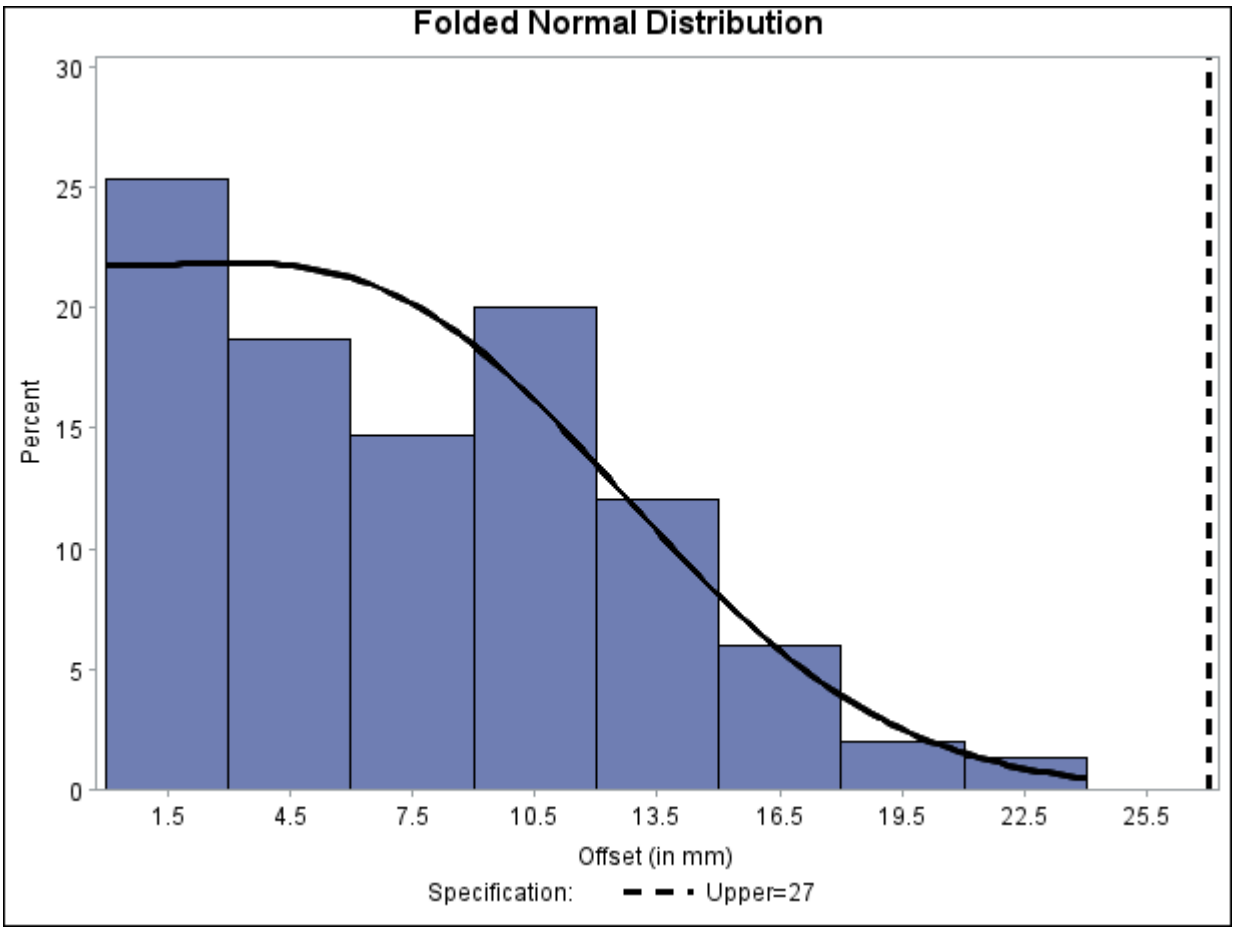
The following statements create an annotate data set named Anno, which contains the coordinates of the fitted curve:

```
data Anno;
   merge Parmest Out;
   length function color $ 8;

   function = 'point';
   color    = 'black';
   size     =  2;
   xsys     = '2';
   ysys     = '2';
   when     = 'a';
   constant = 39.894*_width_;;
   left     =  _midpt1_ - .5*_width_;
   right    =  _midptn_ + .5*_width_;
   inc      = (right-left)/100;
   do x = left to right by inc;
      z1 = (x-emu)/esig;
      z2 = (x+emu)/esig;
      y  = (constant/esig)*(exp(-0.5*z1*z1)+exp(-0.5*z2*z2));
      output;
      function = 'draw';
   end;
run;
```

The following statements read the ANNOTATE= data set and display the histogram and fitted curve, as shown in Output 6.14.4:

```
ods graphics off;
title "Folded Normal Distribution";
proc capability data=Assembly noprint;
   spec usl=27 cusl=black lusl=2 wusl=2;
   histogram Offset / annotate = Anno;
run;
```

Output 6.14.4 Histogram with Annotated Folded Normal Curve



INSET Statement: CAPABILITY Procedure

Overview: INSET Statement

Graphical displays such as histograms and probability plots are commonly used for process capability analysis. You can use the INSET statement to enhance these plots by adding a box or table (referred to as an *inset*) of summary statistics directly to the graph. An inset typically displays statistics calculated by the CAPABILITY procedure but can also display values provided in a SAS data set. A typical application of the INSET statement is to augment a histogram with the sample size, mean, standard deviation, and process capability index C_{pk}.

Note that the INSET statement by itself does not produce a display and must be used with the CDFPLOT, COMPHISTOGRAM, HISTOGRAM, PPPLOT, PROBPLOT, or QQPLOT statement.

You can use options in the INSET statement to

- specify the position of the inset

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

The INSET statement is not applicable when you produce line printer plots by specifying the LINEPRINTER option in the PROC CAPABILITY statement.

Getting Started: INSET Statement

This section introduces the INSET statement with examples that illustrate commonly used options. Complete syntax for the INSET statement is presented in the section "Syntax: INSET Statement" on page 389, and advanced examples are given in the section "Examples: INSET Statement" on page 409.

Displaying Summary Statistics on a Histogram

NOTE: See *Histograms with INSET Statement Features* in the SAS/QC Sample Library.

In a plant producing copper wire, an important quality characteristic is the torsion strength, measured as the twisting force in pounds per inch necessary to break the wire. The following statements create the SAS data set Wire, which contains the torsion strengths (Strength) for 50 different wire samples:

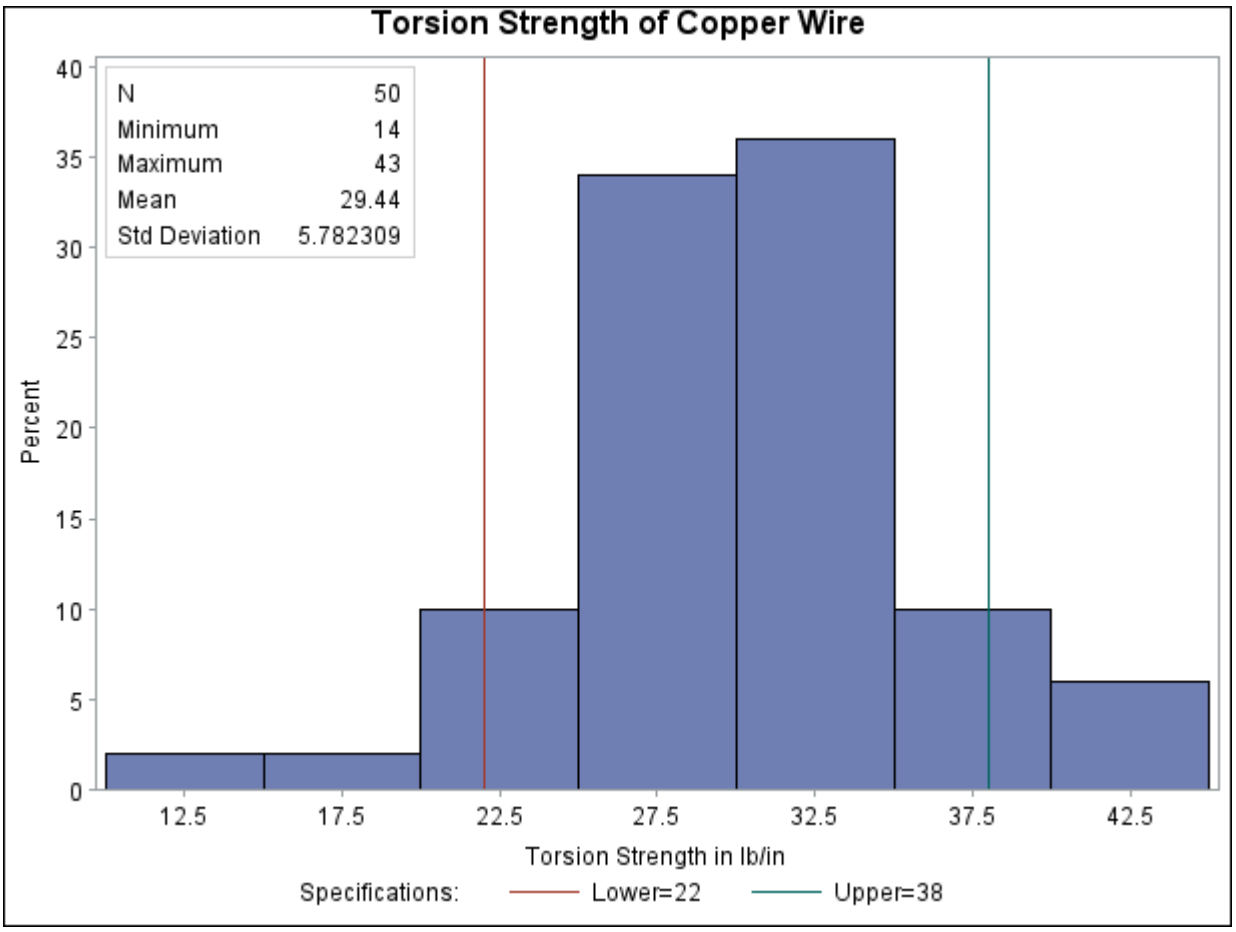
```
data Wire;
   label Strength='Torsion Strength in lb/in';
   input Strength @@;
   datalines;
25 25 36 31 26 36 29 37 37 20
34 27 21 35 30 41 33 21 26 26
19 25 14 32 30 29 31 26 22 24
34 33 28 26 43 30 40 32 32 31
25 26 27 34 33 27 33 29 30 31
;
```

A histogram is used to examine the data distribution. For a more complete report, the sample size, minimum value, maximum value, mean, and standard deviation are displayed on the histogram. The following statements illustrate how to inset these statistics:

```
ods graphics off;
title 'Torsion Strength of Copper Wire';
proc capability data=Wire noprint;
   spec lsl=22 usl=38;
   histogram Strength;
   inset n min max mean std;
run;
```

The resulting histogram is displayed in Figure 6.17. The INSET statement immediately follows the plot statement that creates the graphical display (in this case, the HISTOGRAM statement). Specify the keywords for inset statistics (such as N, MIN, MAX, MEAN, and STD) immediately after the word INSET. The inset statistics appear in the order in which you specify the keywords.

Figure 6.17 A Histogram with an Inset



A complete list of keywords that you can use with the INSET statement is provided in "Summary of INSET Keywords" on page 391. Note that the set of keywords available for a particular display depends on both the plot statement that precedes the INSET statement and the options that you specify in the plot statement.

The following examples illustrate options commonly used for enhancing the appearance of an inset.

Formatting Values and Customizing Labels

NOTE: See *Histograms with INSET Statement Features* in the SAS/QC Sample Library.

By default, each inset statistic is identified with an appropriate label, and each numeric value is printed using an appropriate format. However, you may want to provide your own labels and formats. For example, in Figure 6.17 the default format for the standard deviation prints an excessive number of decimal places. The following statements correct this problem, as well as customizing some of the labels displayed in the inset:

```
ods graphics on;
proc capability data=Wire noprint;
   spec lsl=22 usl=38;
   histogram Strength;
   inset n='Sample Size' min max mean std='Std Dev' (5.2);
run;
```

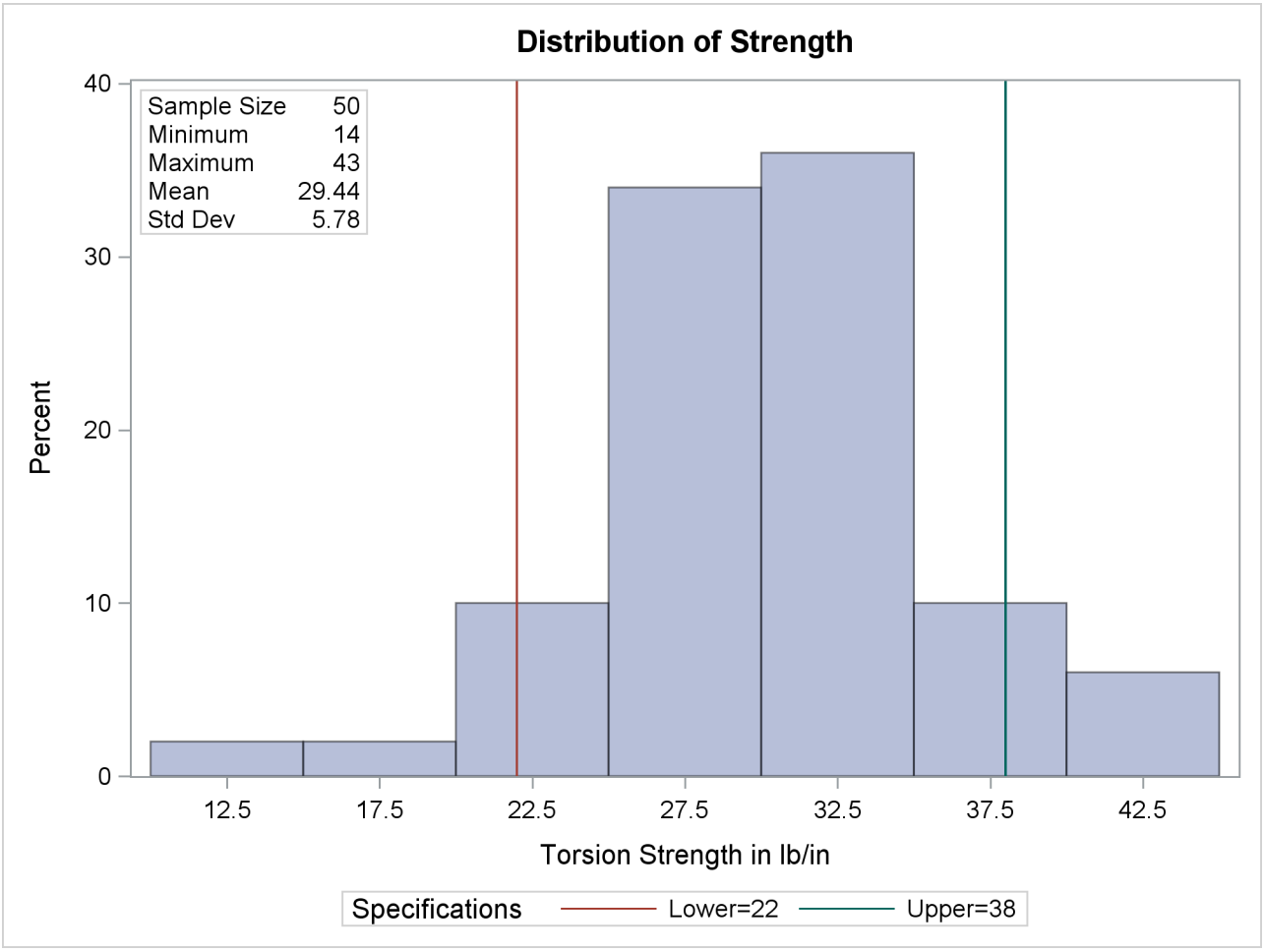
The ODS GRAPHICS ON statement specified before the PROC CAPABILITY statement enables ODS Graphics, so the histogram is created using ODS Graphics instead of traditional graphics.

The resulting histogram is displayed in Figure 6.18. You can provide your own label by specifying the keyword for that statistic followed by an equal sign (=) and the label in quotes. The label can have up to 24 characters.

The format 5.2 specified in parentheses after the keyword STD displays the standard deviation with a field width of five and two decimal places. In general, you can specify any numeric SAS format in parentheses after an inset keyword. You can also specify a format to be used for all the statistics in the INSET statement with the FORMAT= option (see the next section, "Adding a Header and Positioning the Inset" on page 388). For more information about SAS formats, refer to *SAS Formats and Informats: Reference*.

Note that if you specify both a label and a format for a statistic, the label must appear before the format, as with the keyword STD in the previous statements.

Figure 6.18 Formatting Values and Customizing Labels in an Inset



Adding a Header and Positioning the Inset

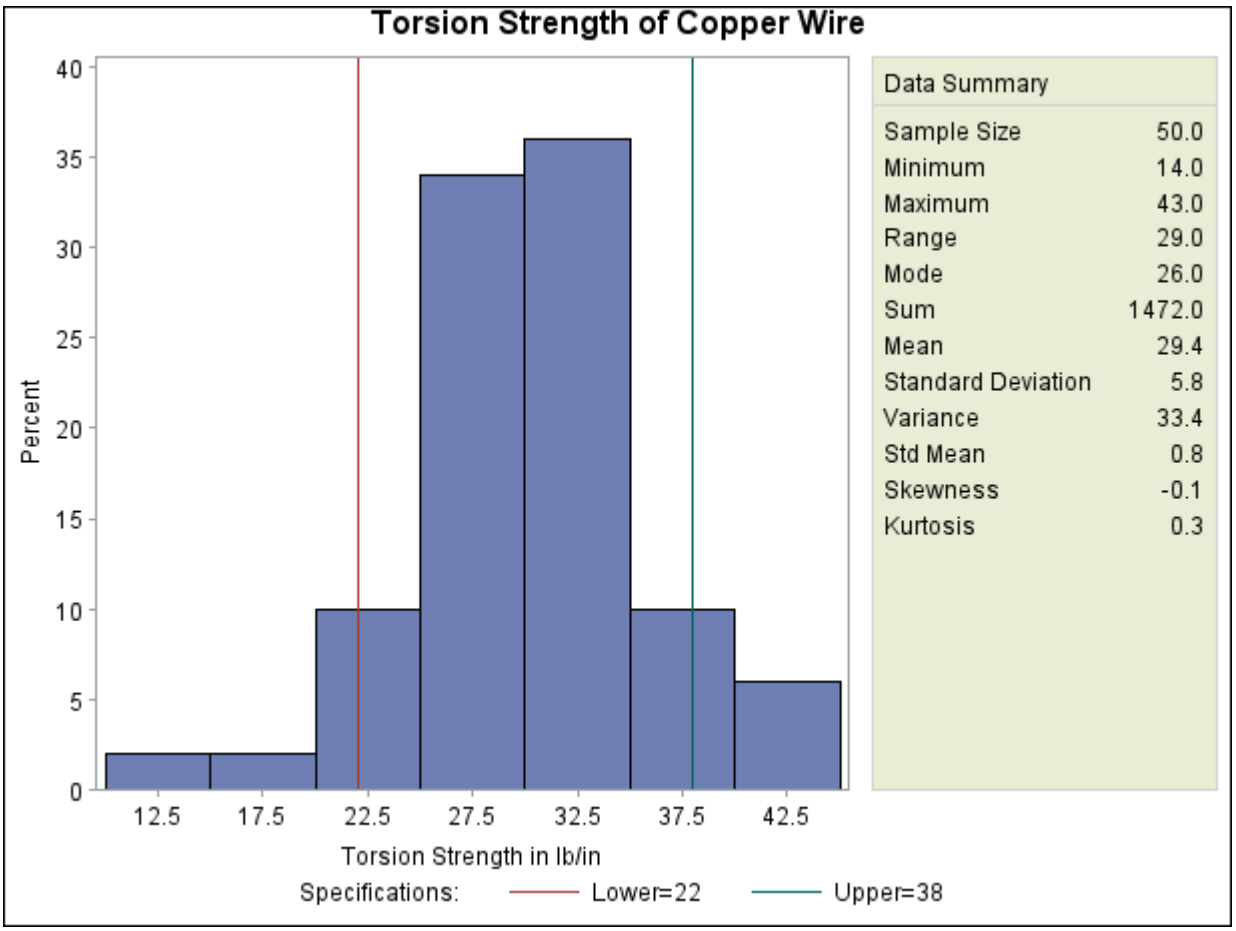
NOTE: See *Histograms with INSET Statement Features* in the SAS/QC Sample Library.

In the previous examples, the inset is displayed in the upper left corner of the plot, the default position for insets added to histograms. You can control the inset position with the POSITION= option. In addition, you can display a header at the top of the inset with the HEADER= option. The following statements create the chart shown in Figure 6.19:

```
ods graphics off;
title 'Torsion Strength of Copper Wire';
proc capability data=Wire noprint;
   spec lsl=22 usl=38;
   histogram Strength;
   inset n='Sample Size' min max range mode sum mean
      std='Standard Deviation' var stdmean skewness
      kurtosis / format = 6.1
                 pos    = rm
                 header = 'Data Summary' cfill = ywh;
run;
```

The header (in this case, *Data Summary*) can be up to 40 characters. Note that a long list of inset statistics is requested. Consequently, POSITION=RM is specified to position the inset in the right margin. For more information about positioning, see "Details: INSET Statement" on page 404. Also note that the FORMAT= option is used to format all inset statistics. The options, such as HEADER=, POSITION=, and FORMAT=, are specified after the slash (/) in the INSET statement. For more details on INSET statement options, see "Dictionary of Options" on page 401.

Figure 6.19 Adding a Header and Repositioning the Inset



Syntax: INSET Statement

The syntax for the INSET statement is as follows:

INSET *keyword-list* < / *options* > **;**

You can use any number of INSET statements in the CAPABILITY procedure. Each INSET statement produces an inset and must follow one of the plot statements: CDFPLOT, COMPHISTOGRAM, HISTOGRAM, PPPLOT, PROBPLOT, or QQPLOT. The inset appears in all displays produced by the plot statement that immediately precedes it. The statistics are displayed in the order in which they are specified. For example, the following statements produce a cumulative distribution plot with two insets and a histogram with one inset:

```
proc capability data=Wire;
   cdfplot Strength;
      inset mean std min max n;
      inset p1 p5 p10;
   histogram Strength;
      inset var skewness kurtosis;
run;
```

The statistics displayed in an inset are computed for a specific process variable from observations for the current BY group. For example, in the following statements, there are two process variables (Strength and Diameter) and a BY variable (Batch). If there are three different batches (levels of Batch), then a total of six histograms are produced. The statistics in each inset are computed for a particular variable and batch. The labels in the inset are the same for each histogram.

```
proc capability data=Wire2;
   by Batch;
   histogram Strength Diameter / normal;
   inset mean std min max normal(mu sigma);
run;
```

The components of the INSET statement are described as follows.

keyword-list

can include any of the keywords listed in "Summary of INSET Keywords" on page 391. Some keywords allow *secondary keywords* to be specified in parentheses immediately after the *primary keyword*. Also, some inset statistics are available only if you request plot statements and options for which those statistics are calculated. For example, consider the following statements:

```
proc capability data=Wire;
   histogram Strength / normal;
   inset mean std normal(ad adpval);
run;
```

The keywords MEAN and STD display the sample mean and standard deviation of Strength. The primary keyword NORMAL with the secondary keywords AD and ADPVAL display the Anderson-Darling goodness-of-fit test statistic and *p*-value in the inset as well. The statistics specified with the NORMAL keyword are available only because a normal distribution has been fit to the data by using the NORMAL option in the HISTOGRAM statement. See the section "Summary of INSET Keywords" for a list of available keywords.

Typically, you specify keywords, to display statistics computed by the CAPABILITY procedure. However, you can also specify the keyword DATA= followed by the name of a SAS data set to display customized statistics. This data set must contain two variables:

- a character variable named _LABEL_ whose values provide labels for inset entries.
- a variable named _VALUE_, which can be either character or numeric, and whose values provide values for inset entries.

The label and value from each observation in the DATA= data set occupy one line in the inset. The position of the DATA= keyword in the keyword list determines the position of its lines in the inset.

By default, inset statistics are identified with appropriate labels, and numeric values are printed using appropriate formats. However, you can provide customized labels and formats. You provide the customized label by specifying the keyword for that statistic followed by an equal sign (=) and the label in quotes. Labels can have up to 24 characters. You provide the numeric format in parentheses after the keyword. Note that if you specify both a label and a format for a statistic, the label must appear before the format. For an example, see "Formatting Values and Customizing Labels" on page 386.

options

appear after the slash (/) and control the appearance of the inset. For example, the following INSET statement uses two appearance options (POSITION= and CTEXT=):

```
inset mean std min max / position=ne ctext=yellow;
```

The POSITION= option determines the location of the inset, and the CTEXT= option specifies the color of the text of the inset.

See "Summary of Options" on page 401 for a list of all available options, and "Dictionary of Options" on page 401 for detailed descriptions. Note the difference between keywords and options; keywords specify the information to be displayed in an inset, whereas options control the appearance of the inset.

Summary of INSET Keywords

Summary Statistics and Process Capability Indices

Table 6.32 Summary Statistics

Keyword	Description
CSS	Corrected sum of squares
CV	Coefficient of variation
GEOMEAN	Geometric mean
KURTOSIS | KURT	Kurtosis
MAX	Largest value
MEAN	Sample mean
MIN	Smallest value
MODE	Most frequent value
N	Sample size
NEXCL	Number of observations excluded by MAXNBIN= or MAXSIGMAS= option
NMISS	Number of missing values
NOBS	Number of observations
RANGE	Range
SKEWNESS | SKEW	Skewness
STD | STDDEV	Standard deviation
STDMEAN | STDERR	Standard error of the mean
SUM	Sum of the observations
SUMWGT	Sum of the weights
USS	Uncorrected sum of squares
VAR	Variance

Table 6.33 Percentile Statistics

Keyword	Description
P1	1st percentile
P5	5th percentile
P10	10th percentile
Q1 | P25	Lower quartile (25th percentile)
MEDIAN | Q2 | P50	Median (50th percentile)
Q3 | P75	Upper quartile (75th percentile)
P90	90th percentile
P95	95th percentile
P99	99th percentile
QRANGE	Interquartile range (Q3 - Q1)

Table 6.34 lists keywords for distribution-free confidence limits for percentiles requested with the CIPCTLDF option.

Table 6.34 Keywords for Distribution-Free Confidence Limits for Percentiles

Keyword	Description
P1_LCL_DF	1st percentile lower confidence limit
P1_UCL_DF	1st percentile upper confidence limit
P5_LCL_DF	5th percentile lower confidence limit
P5_UCL_DF	5th percentile upper confidence limit
P10_LCL_DF	10th percentile lower confidence limit
P10_UCL_DF	10th percentile upper confidence limit
Q1_LCL_DF | P25_LCL_DF	Lower quartile (25th percentile) lower confidence limit
Q1_UCL_DF | P25_UCL_DF	Lower quartile (25th percentile) upper confidence limit
MEDIAN_LCL_DF | Q2_LCL_DF | P50_LCL_DF	Median (50th percentile) lower confidence limit
MEDIAN_UCL_DF | Q2_UCL_DF | P50_UCL_DF	Median (50th percentile) upper confidence limit
Q3_LCL_DF | P75_LCL_DF	Upper quartile (75th percentile) lower confidence limit
Q3_UCL_DF | P75_UCL_DF	Upper quartile (75th percentile) upper confidence limit
P90_LCL_DF	90th percentile lower confidence limit
P90_UCL_DF	90th percentile upper confidence limit
P95_LCL_DF	95th percentile lower confidence limit
P95_UCL_DF	95th percentile upper confidence limit
P99_LCL_DF	99th percentile lower confidence limit
P99_UCL_DF	99th percentile upper confidence limit

Table 6.35 lists keywords for percentile confidence limits computed assuming normality requested with the CIPCTLNORMAL option.

Table 6.35 Keywords Percentile Confidence Limits Assuming Normality

Keyword	Description
P1_LCL	1st percentile lower confidence limit
P1_UCL	1st percentile upper confidence limit
P5_LCL	5th percentile lower confidence limit
P5_UCL	5th percentile upper confidence limit
P10_LCL	10th percentile lower confidence limit
P10_UCL	10th percentile upper confidence limit
Q1_LCL | P25_LCL	Lower quartile (25th percentile) lower confidence limit
Q1_UCL | P25_UCL	Lower quartile (25th percentile) upper confidence limit
MEDIAN_LCL | Q2_LCL | P50_LCL	Median (50th percentile) lower confidence limit
MEDIAN_UCL | Q2_UCL | P50_UCL	Median (50th percentile) upper confidence limit
Q3_LCL | P75_LCL	Upper quartile (75th percentile) lower confidence limit
Q3_UCL | P75_UCL	Upper quartile (75th percentile) upper confidence limit
P90_LCL	90th percentile lower confidence limit
P90_UCL	90th percentile upper confidence limit
P95_LCL	95th percentile lower confidence limit
P95_UCL	95th percentile upper confidence limit
P99_LCL	99th percentile lower confidence limit
P99_UCL	99th percentile upper confidence limit

Table 6.36 Robust Statistics

Keyword	Description
GINI	Gini's mean difference
MAD	Median absolute difference about the median
QN	Q_n, alternative to MAD
SN	S_n, alternative to MAD
STD_GINI	Gini's standard deviation
STD_MAD	MAD standard deviation
STD_QN	Q_n standard deviation
STD_QRANGE	Interquartile range standard deviation
STD_SN	S_n standard deviation

Table 6.37 Hypothesis Testing

Keyword	Description
MSIGN	Sign statistic
NORMALTEST	Test statistic for normality
PNORMAL	Probability value for the test of normality
SIGNRANK	Signed rank statistic
PROBM	Probability of greater absolute value for the sign statistic
PROBN	Probability value for the test of normality
PROBS	Probability value for the signed rank test
PROBT	Probability value for the Student's *t* test
T	Statistics for Student's *t* test

Table 6.38 Input Data Set

Keyword	Description
DATA=	(label, value) pairs from input data set

Table 6.39 Capability Indices and Confidence Limits

Keyword	Description
CP	Capability index C_p
CPLCL	Lower confidence limit for C_p
CPUCL	Upper confidence limit for C_p
CPK	Capability index C_{pk}
CPKLCL	Lower confidence limit for C_{pk}
CPKUCL	Upper confidence limit for C_{pk}
CPL	Capability index CPL
CPM	Capability index C_{pm}
CPMLCL	Lower confidence limit for C_{pm}
CPMUCL	Upper confidence interval for C_{pm}
CPU	Capability index CPU
K	Capability index K

Table 6.40 Specification Limits and Related Information

Keyword	Description
LSL	Lower specification limit
USL	Upper specification limit
TARGET	Target value
PCTGTR	Percent of nonmissing observations that exceed the upper specification limit
PCTLSS	Percent of nonmissing observations that are less than the lower specification limit
PCTBET	Percent of nonmissing observations between the upper and lower specification limits (inclusive)

Statistics Available with Parametric Density Estimates

You can request parametric density estimates with all plot statements in the CAPABILITY procedure (CDFPLOT, COMPHISTOGRAM, HISTOGRAM, PPPLOT, PROBPLOT, and QQPLOT). You can display parameters and statistics associated with these estimates in an inset by specifying a distribution keyword followed by secondary keywords in parentheses. For example, the following statements create a histogram for Strength with a fitted exponential density curve:

```
proc capability data=Wire;
   histogram Strength / exp;
   inset exp(sigma theta);
run;
```

The secondary keywords SIGMA and THETA for the EXP distribution keyword request an inset displaying the values of the exponential scale parameter σ and threshold parameter θ. You must request the distribution option in the plot statement to display the corresponding distribution statistics in an inset. Specifying a distribution keyword with no secondary keywords produces an inset displaying the full set of parameters for that distribution. See Output 6.15.1 for an example of an inset with statistics from a fitted normal curve.

The following table describes the available distribution keywords. Note that some keywords are not available with all plot statements.

Table 6.41 Density Estimation Primary Keywords

Keyword	Distribution	Plot Statement Availability
BETA	beta	all but COMPHISTOGRAM
EXPONENTIAL	exponential	all but COMPHISTOGRAM
GAMMA	gamma	all but COMPHISTOGRAM
GUMBEL	Gumbel	all but COMPHISTOGRAM
IGAUSS	inverse Gaussian	CDFPLOT, HISTOGRAM, PPPLOT
LOGNORMAL	lognormal	all but COMPHISTOGRAM
NORMAL	normal	all
PARETO	generalized Pareto	all but COMPHISTOGRAM
POWER	power function	all but COMPHISTOGRAM
RAYLEIGH	Rayleigh	all but COMPHISTOGRAM
SB	Johnson S_B	HISTOGRAM
SU	Johnson S_U	HISTOGRAM
WEIBULL	Weibull	all but COMPHISTOGRAM
WEIBULL2	2-parameter Weibull	PROBPLOT, QQPLOT

Table 6.42 lists the secondary keywords available with each distribution keyword listed in Table 6.41. In many cases, aliases can be used (for example, ALPHA in place of SHAPE1).

Table 6.42 Density Estimation Secondary Keywords

Secondary Keyword	Alias	Description
Secondary Keywords Available with the BETA Keyword		
ALPHA	SHAPE1	First shape parameter α
BETA	SHAPE2	Second shape parameter β
SIGMA	SCALE	Scale parameter σ
THETA	THRESHOLD	Lower threshold parameter θ
MEAN		Mean of the fitted distribution
STD		Standard deviation of the fitted distribution
Secondary Keywords Available with the EXPONENTIAL Keyword		
SIGMA	SCALE	Scale parameter σ
THETA	THRESHOLD	Threshold parameter θ
MEAN		Mean of the fitted distribution
STD		Standard deviation of the fitted distribution
Secondary Keywords Available with the GAMMA Keyword		
ALPHA	SHAPE	Shape parameter α
SIGMA	SCALE	Scale parameter σ
THETA	THRESHOLD	Threshold parameter θ
MEAN		Mean of the fitted distribution
STD		Standard deviation of the fitted distribution
Secondary Keywords Available with the GUMBEL Keyword		
MU		Location parameter μ
SIGMA	SCALE	Scale parameter σ

Table 6.42 (*continued*)

Secondary Keyword	Alias	Description
MEAN		Mean of the fitted distribution
STD		Standard deviation of the fitted distribution
Secondary Keywords Available with the IGAUSS Keyword		
MU		Mean parameter μ
LAMBDA		Shape parameter λ
MEAN		Mean of the fitted distribution
STD		Standard deviation of the fitted distribution
Secondary Keywords Available with the LOGNORMAL Keyword		
SIGMA	SHAPE	Shape parameter σ
THETA	THRESHOLD	Threshold parameter θ
ZETA	SCALE	Scale parameter ζ
MEAN		Mean of the fitted distribution
STD		Standard deviation of the fitted distribution
Secondary Keywords Available with the NORMAL Keyword		
MU	MEAN	Mean parameter μ
SIGMA	STD	Scale parameter σ
Secondary Keywords Available with the PARETO Keyword		
ALPHA		Shape parameter α
SIGMA	SCALE	Scale parameter σ
THETA	THRESHOLD	Threshold parameter θ
MEAN		Mean of the fitted distribution
STD		Standard deviation of the fitted distribution
Secondary Keywords Available with the POWER Keyword		
ALPHA		Shape parameter α
SIGMA	SCALE	Scale parameter σ
THETA	THRESHOLD	Threshold parameter θ
MEAN		Mean of the fitted distribution
STD		Standard deviation of the fitted distribution
Secondary Keywords Available with the RAYLEIGH Keyword		
SIGMA	SCALE	Scale parameter σ
THETA	THRESHOLD	Threshold parameter θ
MEAN		Mean of the fitted distribution
STD		Standard deviation of the fitted distribution
Secondary Keywords Available with the SB Keyword		
DELTA	SHAPE1	Shape parameter δ
GAMMA	SHAPE2	Shape parameter γ
SIGMA	SCALE	Scale parameter σ
THETA	THRESHOLD	Threshold parameter θ
MEAN		Mean of the fitted distribution
STD		Standard deviation of the fitted distribution
Secondary Keywords Available with the SU Keyword		
DELTA	SHAPE1	Shape parameter δ
GAMMA	SHAPE2	Shape parameter γ

Table 6.42 (*continued*)

Secondary Keyword	Alias	Description
SIGMA	SCALE	Scale parameter σ
THETA		Location parameter θ
MEAN		Mean of the fitted distribution
STD		Standard deviation of the fitted distribution
Secondary Keywords Available with the WEIBULL Keyword		
C	SHAPE	Shape parameter c
SIGMA	SCALE	Scale parameter σ
THETA	THRESHOLD	Threshold parameter θ
MEAN		Mean of the fitted distribution
STD		Standard deviation of the fitted distribution
Secondary Keywords Available with the WEIBULL2 Keyword		
C	SHAPE	Shape parameter c
SIGMA	SCALE	Scale parameter σ
THETA	THRESHOLD	Known lower threshold θ_0
MEAN		Mean of the fitted distribution
STD		Standard deviation of the fitted distribution

The secondary keywords listed in Table 6.43 can be used with any distribution keyword but *only* with the HISTOGRAM and COMPHISTOGRAM plot statements.

Table 6.43 Statistics Computed from Any Parametric Density Estimate

Secondary Keyword	Description
CP	Capability index C_p
CPK	Capability index C_{pk}
CPL	Capability index CPL
CPM	Capability index C_{pm}
CPU	Capability index CPU
ESTPCTLSS	Estimated percentage less than the lower specification limit
ESTPCTGTR	Estimated percentage greater than the upper specification limit
K	Capability index K

The secondary keywords listed in Table 6.44 can be used with any distribution keyword but *only* with the HISTOGRAM plot statement (see Example 6.15).

Table 6.44 Goodness-of-Fit Statistics for Fitted Curves

Secondary Keyword	Description
CHISQ	Chi-square statistic
DF	Degrees of freedom for the chi-square test
PCHISQ	Probability value for the chi-square test
AD	Anderson-Darling EDF test statistic
ADPVAL	Anderson-Darling EDF test *p*-value
CVM	Cramér-von Mises EDF test statistic
CVMPVAL	Cramér-von Mises EDF test *p*-value
KSD	Kolmogorov-Smirnov EDF test statistic
KSDPVAL	Kolmogorov-Smirnov EDF test *p*-value

Table 6.45 lists primary keywords available only with the HISTOGRAM and COMPHISTOGRAM plot statements. These keywords display fill areas on a histogram. If you fit a parametric density on a histogram and request that the area under the curve be filled, these keywords display the percentage of the distribution area that lies below the lower specification limit, between the specification limits, or above the upper specification limit. If you do not fill the area beneath a parametric density estimate, these keywords display the observed proportion of observations (that is, the area in the bars of the histogram).

You should use these options with the FILL, CFILL=, and PFILL= options in the HISTOGRAM and COMPHISTOGRAM statements and with the CLEFT=, CRIGHT=, PLEFT=, and PRIGHT= options in the SPEC statements. See Output 6.16.1 for an example.

Table 6.45 Curve Area Keywords

Keyword	Alias	Description
BETWEENPCT	BETPCT	Area between the specification limits
LSLPCT		Area below the lower specification limit
USLPCT		Area above the upper specification limit

Statistics Available with Nonparametric Kernel Density Estimates

You can request nonparametric kernel density estimates with the HISTOGRAM and COMPHISTOGRAM plot statements. You can display statistics associated with these estimates by specifying a kernel density keyword followed by secondary keywords in parentheses. For example, the following statements create a histogram for Strength with a fitted kernel density estimate:

```
proc capability data=Wire;
   histogram Strength / kernel;
   inset kernel(c amise);
run;
```

The secondary keywords C and AMISE for the KERNEL keyword display the values of the standardized bandwidth c and the approximate mean integrated square error.

Note that you can specify more than one kernel density estimate on a single histogram. If you specify multiple kernel density estimates, you can request inset statistics for all of the estimates with the KERNEL keyword, or you can display inset statistics for up to five individual curves with KERNEL*n* keywords, as in the following example:

```
proc capability data=Wire;
   histogram Strength / kernel(c = 1 2 3);
   inset kernel2(c) kernel3(c);
run;
```

Three kernel density estimates are displayed on the histogram, but the inset displays the value of c only for the second and third estimates.

Table 6.46 lists the kernel density keywords. Table 6.47 lists the available secondary keywords.

Table 6.46 Kernel Density Estimate Primary Keywords

Keyword	Description
KERNEL	displays statistics for all kernel estimates
KERNEL*n*	displays statistics for only the *n*th kernel density estimate $n = 1, 2, 3, 4,$ or 5

Table 6.47 Secondary Keywords Available with the KERNEL Keyword

Secondary Keyword	Description
TYPE	kernel type: normal, quadratic, or triangular
BANDWIDTH	bandwidth λ for the density estimate
BWIDTH	alias for BANDWIDTH
C	standardized bandwidth c for the density estimate: $c = \frac{\lambda}{Q} n^{\frac{1}{5}}$ where n = sample size, λ = bandwidth, and Q = interquartile range
AMISE	approximate mean integrated square error (MISE) for the kernel density

Summary of Options

The following table lists the INSET statement options. See the section "Dictionary of Options" for complete descriptions of the options.

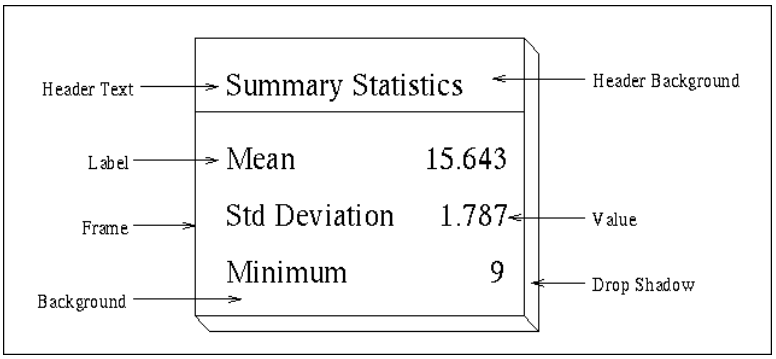
Table 6.48 INSET Options

Option	Description
CFILL=	specifies color of inset background
CFILLH=	specifies color of header background
CFRAME=	specifies color of frame
CHEADER=	specifies color of header text
CSHADOW=	specifies color of drop shadow
CTEXT=	specifies color of inset text
DATA	specifies data units for POSITION=(x, y) coordinates
FONT=	specifies font of text
FORMAT=	specifies format of values in inset
GUTTER=	specifies gutter width for inset in top or bottom margin
HEADER=	specifies header text
HEIGHT=	specifies height of inset text
NCOLS=	specifies number of columns for inset in top or bottom margin
NOFRAME	suppresses frame around inset
POSITION=	specifies position of inset
REFPOINT=	specifies reference point of inset positioned with POSITION=(x, y) coordinates

Dictionary of Options

The following sections provide detailed descriptions of options for the INSET statement. Terms used in this section are illustrated in Figure 6.20.

Figure 6.20 The Inset



General Options

You can specify the following general options:

DATA
specifies that data coordinates are to be used in positioning the inset with the POSITION= option. The DATA option is available only when you specify POSITION= (x, y), and it must be placed immediately after the coordinates (x, y). For details, see the entry for the POSITION= option or "Positioning the Inset Using Coordinates" on page 406. See Figure 6.23 for an example.

FORMAT=*format*
specifies a format for all the values displayed in an inset. If you specify a format for a particular statistic, then this format overrides the format you specified with the FORMAT= option. See Figure 6.19 or Output 6.15.1 for an example.

GUTTER=*value*
specifies the gutter width in percent screen units for an inset located in the top or bottom margin of ODS Graphics output. The gutter is the space between columns of (label, value) pairs in an inset. The default value is four. This option is ignored if ODS Graphics is disabled.

HEADER= '*string*'
specifies the header text. The *string* cannot exceed 40 characters. If you do not specify the HEADER= option, no header line appears in the inset. If all the keywords listed in the INSET statement are secondary keywords corresponding to a fitted curve on a histogram, a default header is displayed that indicates the distribution and identifies the curve. See Figure 6.19 for an example of a specified header and Output 6.15.1 for an example of the default header for a fitted normal curve.

NCOLS=*n*
specifies the number of columns of (label, value) pairs displayed in an inset located in the top or bottom margin of ODS Graphics output. The default value is three. This option is ignored if ODS Graphics is disabled.

NOFRAME
suppresses the frame drawn around the text.

POSITION=*position*

POS=*position*
determines the position of the inset. The *position* can be a compass point keyword, a margin keyword, or a pair of coordinates (x, y). You can specify coordinates in axis percent units or axis data units. For more information, see "Details: INSET Statement" on page 404. By default, POSITION=NW, which positions the inset in the upper left (northwest) corner of the display.

NOTE: In this release of the CAPABILITY procedure, you cannot specify coordinates with the POSITION= option when producing ODS Graphics output.

Options for Traditional Graphics

You can specify the following options if you are producing traditional graphics:

CFILL=*color* | BLANK
specifies the color of the background (including the header background if you do not specify the CFILLH= option). See Output 6.15.1 for an example.

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

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

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

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

CSHADOW=*color*

CS=*color*
specifies the color of the drop shadow. See Output 6.16.1 for an example. By default, if you do not specify the CSHADOW= option, a drop shadow is not displayed.

CTEXT=*color*

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

FONT=*font*
specifies the font of the text. By default, the font is SIMPLEX if the inset is located in the interior of the plot, and the font is the same as the other text displayed on the plot if the inset is located in the exterior of the plot.

HEIGHT=*value*
specifies the height of the text.

REFPOINT=BR | BL | TR | TL

RP=BR | BL | TR | TL
specifies the reference point for an inset that is positioned by a pair of coordinates with the POSITION= option. Use the REFPOINT= option with POSITION= coordinates. The REFPOINT= option specifies which corner of the inset frame you want positioned at coordinates (x, y). The keywords BL, BR, TL, and TR represent bottom left, bottom right, top left, and top right, respectively. See Figure 6.24 for an example. The default is REFPOINT=BL.

If you specify the position of the inset as a compass point or margin keyword, the REFPOINT= option is ignored. For more information, see "Positioning the Inset Using Coordinates" on page 406.

Details: INSET Statement

This section provides details on three different methods of positioning the inset with the POSITION= option. With the POSITION= option, you can specify

- compass points
- keywords for margin positions
- coordinates in data units or percent axis units

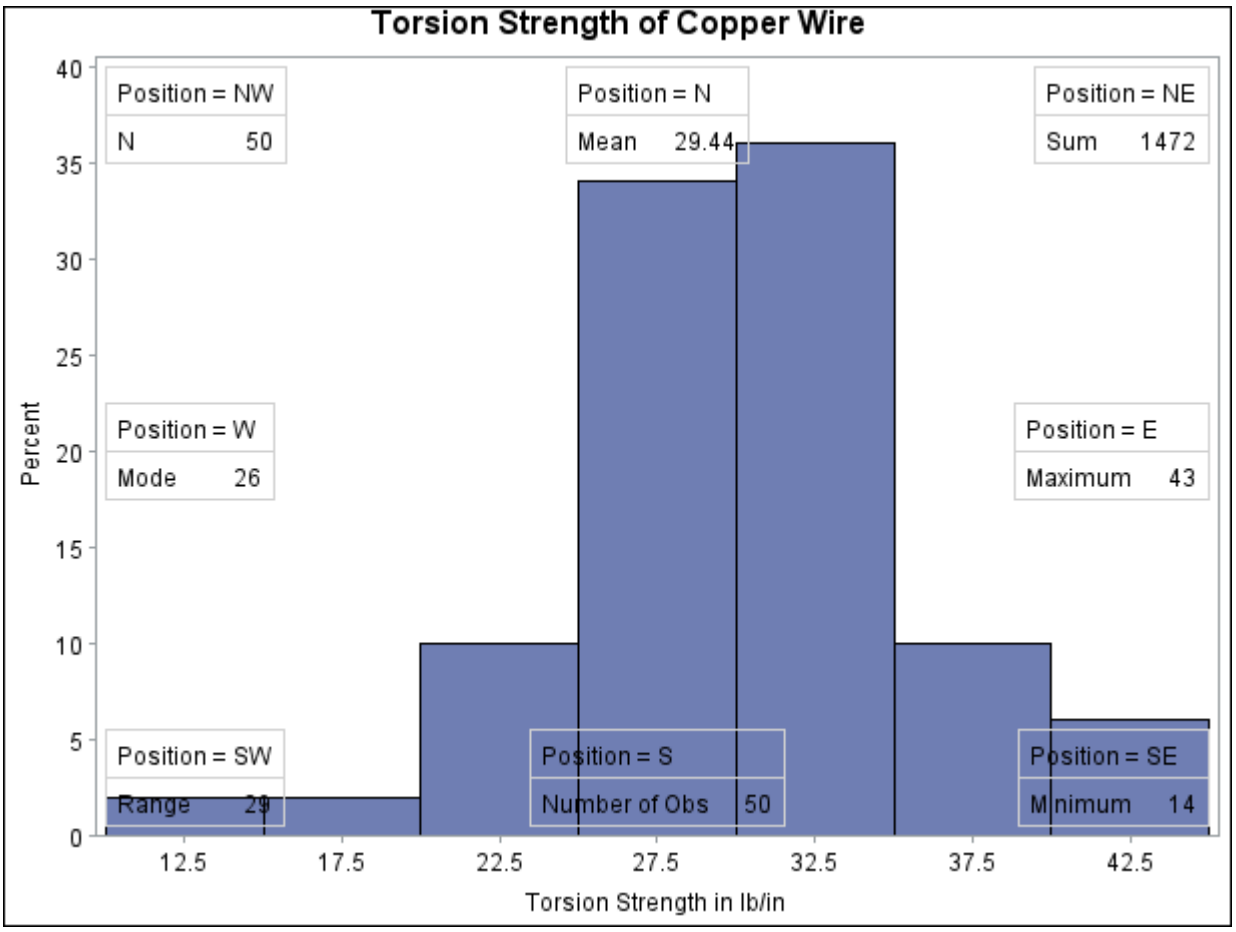
Positioning the Inset Using Compass Points

NOTE: See *Positioning the Inset* in the SAS/QC Sample Library.

You can specify the eight compass points N, NE, E, SE, S, SW, W, and NW as keywords for the POSITION= option. The following statements create the display in Figure 6.21, which demonstrates all eight compass positions. The default is NW.

```
ods graphics off;
title 'Torsion Strength of Copper Wire';
proc capability data=Wire;
   histogram Strength / odstitle = title;
   inset n     / header='Position = NW' pos=nw;
   inset mean  / header='Position = N ' pos=n ;
   inset sum   / header='Position = NE' pos=ne;
   inset max   / header='Position = E ' pos=e ;
   inset min   / header='Position = SE' pos=se;
   inset nobs  / header='Position = S ' pos=s ;
   inset range / header='Position = SW' pos=sw;
   inset mode  / header='Position = W ' pos=w ;
run;
```

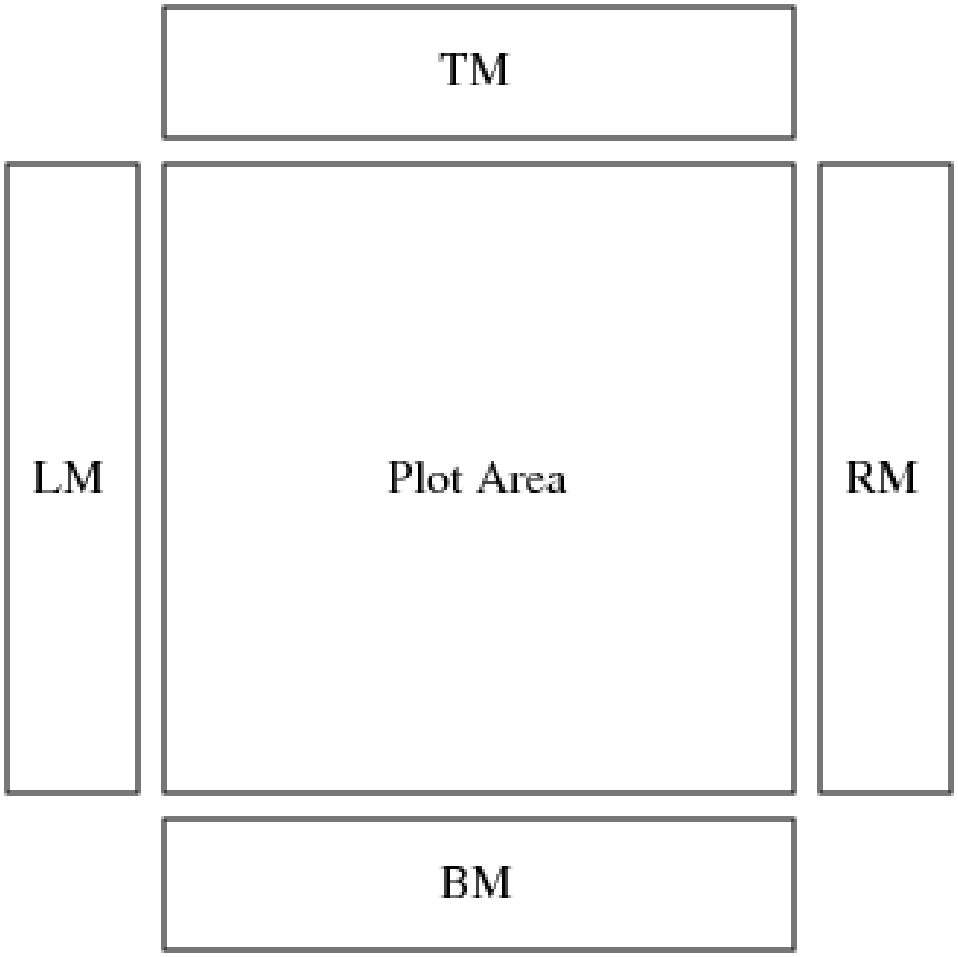
Figure 6.21 Insets Positioned Using Compass Points



Positioning the Inset in the Margins

You can also position the inset in one of the four margins surrounding the plot area using the margin keywords LM, RM, TM, or BM, as illustrated in Figure 6.22.

Figure 6.22 Positioning Insets in the Margins



For an example of an inset placed in the right margin, see Figure 6.19. Margin positions are recommended if a large number of statistics are listed in the INSET statement. If you attempt to display a lengthy inset in the interior of the plot, it is likely that the inset will collide with the data display.

Positioning the Inset Using Coordinates

If you are producing traditional graphics, you can also specify the position of the inset with coordinates: POSITION= (x, y). The coordinates can be given in axis percent units (the default) or in axis data units.

NOTE: In this release of the CAPABILITY procedure, you cannot position insets by using coordinates when producing ODS Graphics output.

Data Unit Coordinates

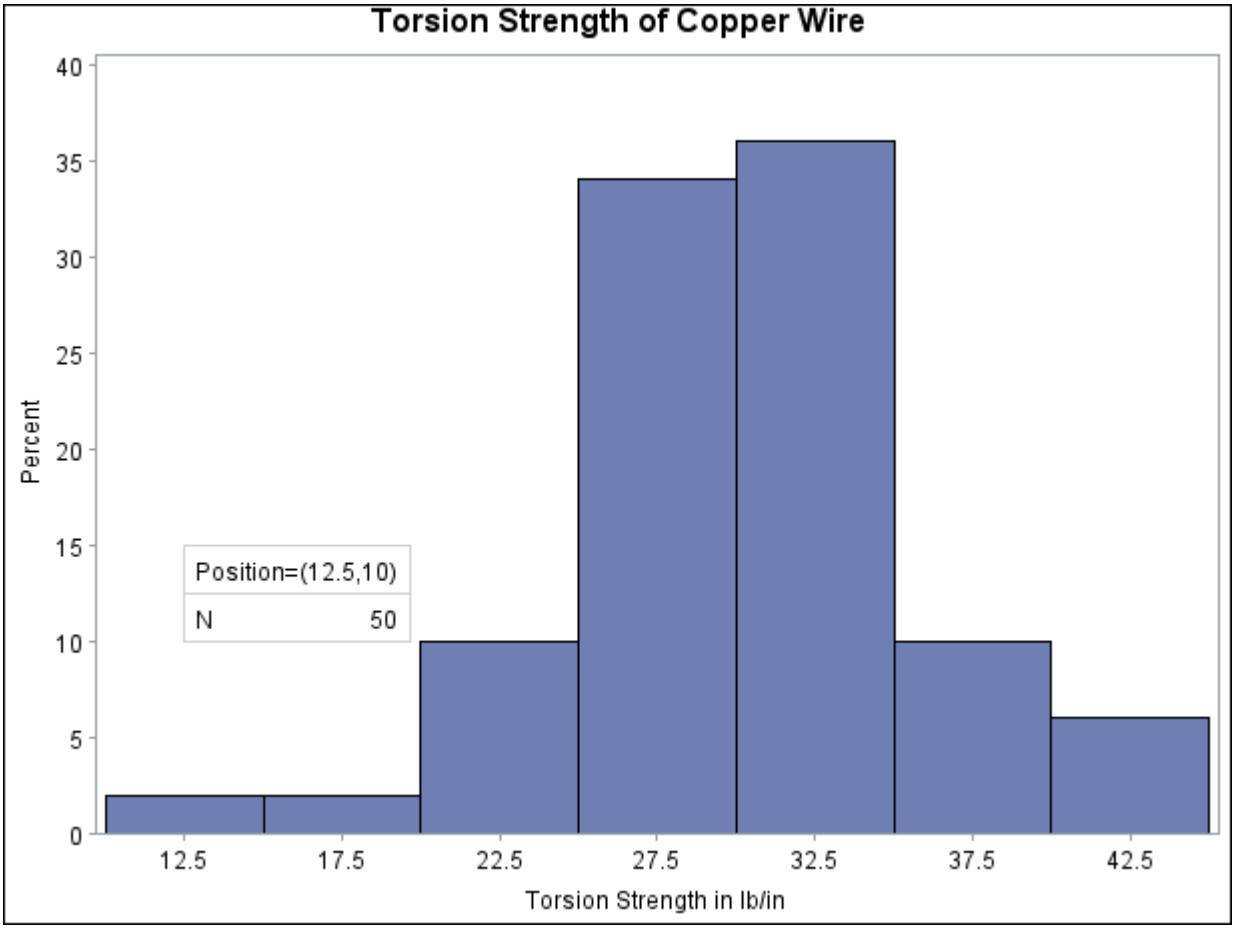
NOTE: See *Positioning the Inset* in the SAS/QC Sample Library.

If you specify the DATA option immediately following the coordinates, the inset is positioned using axis data units. For example, the following statements place the bottom left corner of the inset at 12.5 on the horizontal axis and 10 on the vertical axis:

```
title 'Torsion Strength of Copper Wire';
proc capability data=Wire;
   histogram Strength;
   inset n / header   = 'Position=(12.5,10)'
             position = (12.5,10) data;
run;
```

The histogram is displayed in Figure 6.23. By default, the specified coordinates determine the position of the bottom left corner of the inset. You can change this reference point with the REFPOINT= option, as in the next example.

Figure 6.23 Inset Positioned Using Data Unit Coordinates



Axis Percent Unit Coordinates

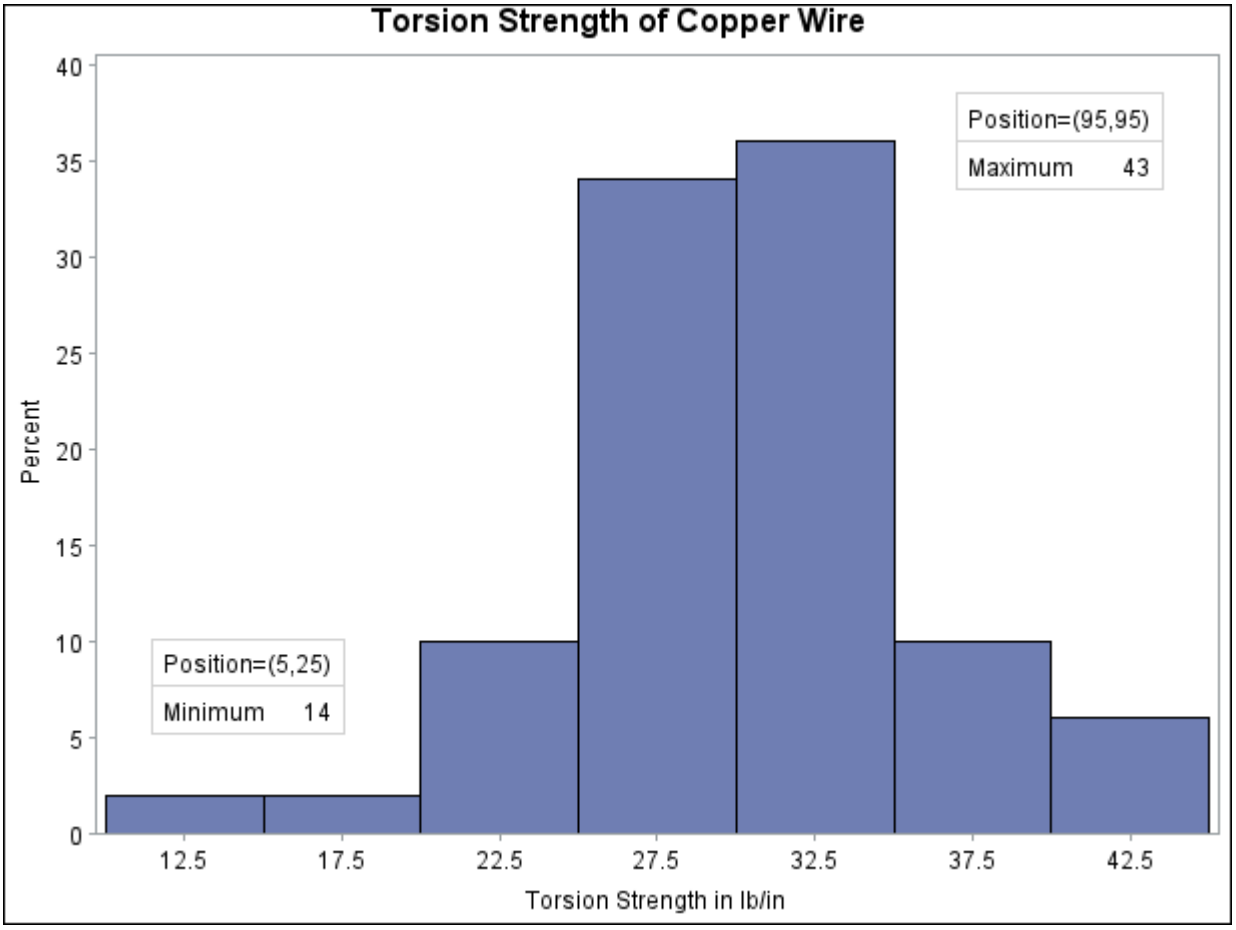
NOTE: See *Positioning the Inset* in the SAS/QC Sample Library.

If you do not use the DATA option, the inset is positioned using axis percent units. The coordinates of the bottom left corner of the display are $(0, 0)$, while the upper right corner is $(100, 100)$. For example, the following statements create a histogram with two insets, both positioned using coordinates in axis percent units:

```
title 'Torsion Strength of Copper Wire';
proc capability data=Wire;
   histogram Strength;
   inset min / position = (5,25)
               header   = 'Position=(5,25)'
               refpoint = tl;
   inset max / position = (95,95)
               header   = 'Position=(95,95)'
               refpoint = tr;
run;
```

The display is shown in Figure 6.24. Notice that the REFPOINT= option is used to determine which corner of the inset is to be placed at the coordinates specified with the POSITION= option. The first inset has REFPOINT=TL, so the top left corner of the inset is positioned 5% of the way across the horizontal axis and 25% of the way up the vertical axis. The second inset has REFPOINT=TR, so the top right corner of the inset is positioned 95% of the way across the horizontal axis and 95% of the way up the vertical axis. Note also that coordinates in axis percent units must be *between* 0 and 100.

Figure 6.24 Inset Positioned Using Axis Percent Unit Coordinates



Examples: INSET Statement

This section provides advanced examples that use the INSET statement.

Example 6.15: Inset for Goodness-of-Fit Statistics

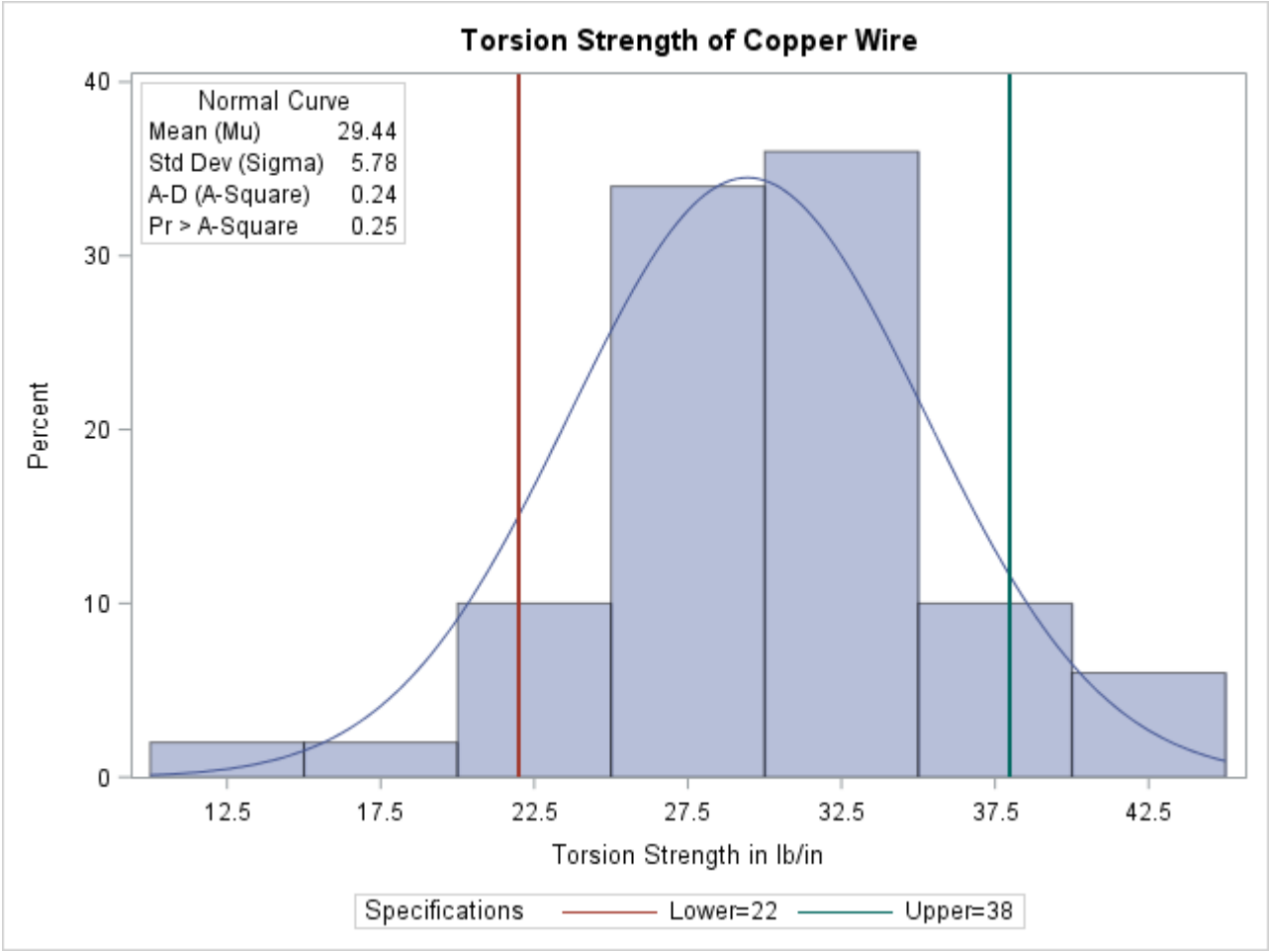
NOTE: See *Inset for Goodness-of-Fit Statistics* in the SAS/QC Sample Library.

This example fits a normal curve to the torsion strength data used in the section "Getting Started: INSET Statement" on page 385. The following statements fit a normal curve and request an inset summarizing the fitted curve with the mean, the standard deviation, and the Anderson-Darling goodness-of-fit test:

```
title 'Torsion Strength of Copper Wire';
proc capability data=Wire noprint;
   spec lsl=22 usl=38;
   histogram Strength / normal(noprint)
                        nocurvelegend
                        odstitle = title;
   inset normal(mu sigma ad adpval) / format = 7.2;
run;
```

The resulting histogram is displayed in Output 6.15.1. The NOCURVELEGEND option in the HISTOGRAM statement suppresses the default legend for curve parameters.

Output 6.15.1 Inset Table with Normal Curve Information



Example 6.16: Inset for Areas Under a Fitted Curve

NOTE: See *Inset for Areas Under a Fitted Curve* in the SAS/QC Sample Library.

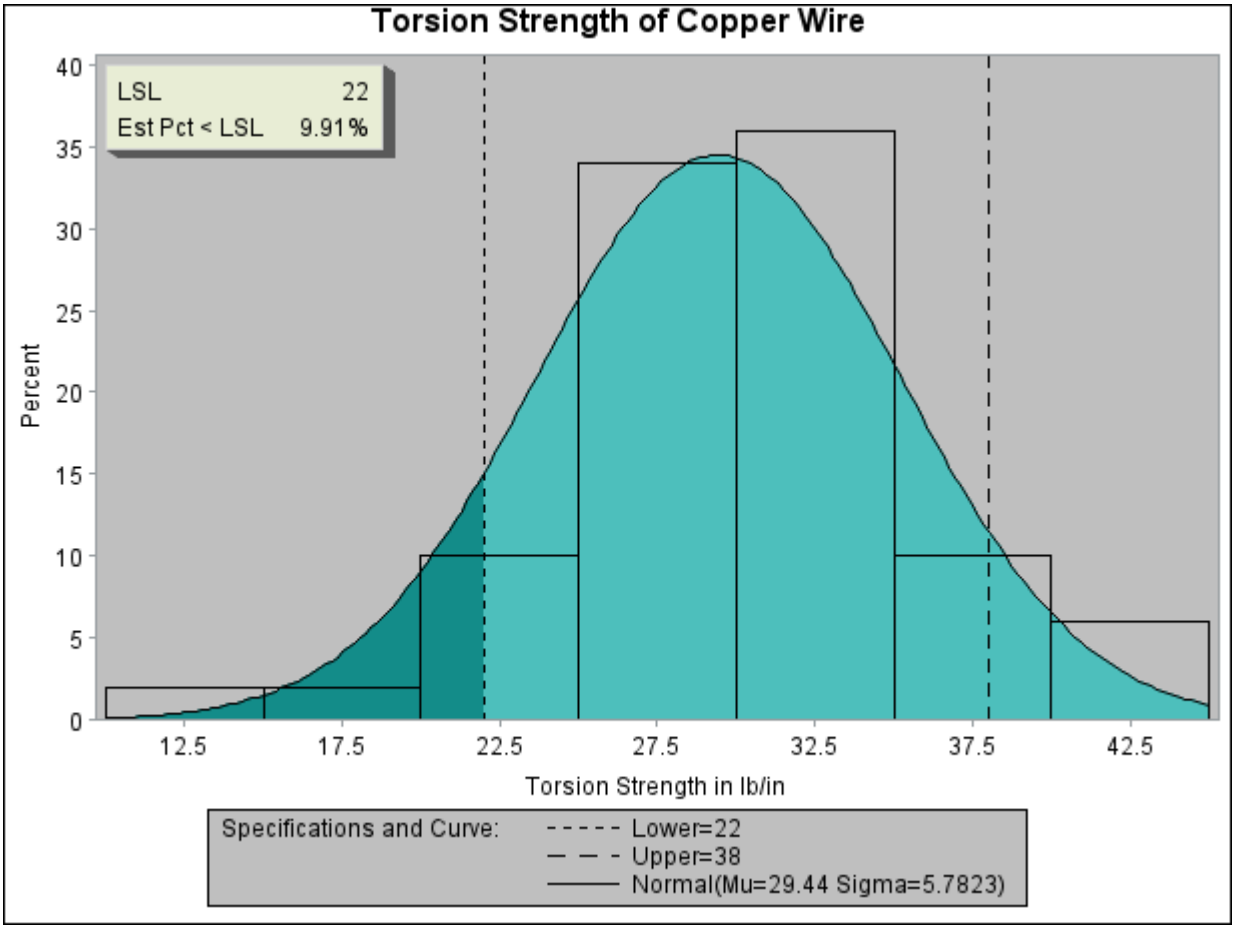
You can use the INSET keywords LSLPCT, USLPCT, and BETWEENPCT to inset legends for areas under histogram bars or fitted curves. The following statements create a histogram with an inset legend for the shaded area under the fitted normal curve to the left of the lower specification limit:

```
ods graphics off;
title 'Torsion Strength of Copper Wire';
legend2 FRAME CFRAME=ligr CBORDER=black POSITION=center;
proc capability data=Wire noprint;
   spec lsl=22 llsl=2  clsl=black cleft=vibg
        usl=38 lusl=20 cusl=black;
   histogram Strength /  cframe = ligr
                         cfill   = bibg
                         legend = legend2
                         normal(color=black noprint fill);
   inset  lsl='LSL' lslpct / cfill=ywh cshadow=dagr;
run;
```

The histogram is displayed in Output 6.16.1. The LSLPCT keyword in the INSET statement requests a legend for the area under the curve to the left of the lower specification limit. The CLEFT= option is used to fill the area under the normal curve to the left of the line, and the CFILL= color is used to fill the remaining area. If the FILL *normal-option* were not specified, the CLEFT= and CFILL= colors would be applied to the corresponding areas under the histogram, not the normal curve, and the inset box would reflect the area under the histogram bars.

You can use the USLPCT keyword in the INSET statement to request a legend for the area to the right of an upper specification limit, and you can use the BETWEENPCT keyword to request a legend for the area between the lower and upper limits. By default, the legend requested with each of the keywords LSLPCT, USLPCT, and BETWEENPCT displays a rectangle that matches the color of the corresponding area. You can substitute a customized label for each rectangle by specifying the keyword followed by an equal sign (=) and the label in quotes.

Output 6.16.1 Displaying Areas Under the Normal Curve



INTERVALS Statement: CAPABILITY Procedure

Overview: INTERVALS Statement

The INTERVALS statement tabulates various statistical intervals for selected process variables. The types of intervals you can request include

- approximate simultaneous prediction intervals for future observations
- prediction intervals for the mean of future observations
- statistical tolerance intervals that contain at least a specified proportion of the population
- confidence intervals for the population mean
- prediction intervals for the standard deviation of future observations
- confidence intervals for the population standard deviation

These intervals are computed assuming the data are sampled from a normal population. See Hahn and Meeker (1991) for a detailed discussion of these intervals.

You can use options in the INTERVALS statement to

- specify which intervals to compute
- provide probability or confidence levels for intervals
- suppress printing of output tables
- create an output data set containing interval information
- specify interval type (one-sided lower, one-sided upper, or two-sided)

Getting Started: INTERVALS Statement

This section introduces the INTERVALS statement with simple examples that illustrate commonly used options. Complete syntax for the INTERVALS statement is presented in the section "Syntax: INTERVALS Statement" on page 416.

Computing Statistical Intervals

NOTE: See *Calculating Various Statistical Intervals* in the SAS/QC Sample Library.

The following statements create the data set Cans, which contains measurements (in ounces) of the fluid weights of 100 drink cans. The filling process is assumed to be in statistical control.

```
data Cans;
   label Weight = "Fluid Weight (ounces)";
   input Weight @@;
   datalines;
12.07  12.02  12.00  12.01  11.98  11.96  12.04  12.05  12.01  11.97
12.03  12.03  12.00  12.04  11.96  12.02  12.06  12.00  12.02  11.91
12.05  11.98  11.91  12.01  12.06  12.02  12.05  11.90  12.07  11.98
12.02  12.11  12.00  11.99  11.95  11.98  12.05  12.00  12.10  12.04
12.06  12.04  11.99  12.06  11.99  12.07  11.96  11.97  12.00  11.97
12.09  11.99  11.95  11.99  11.99  11.96  11.94  12.03  12.09  12.03
11.99  12.00  12.05  12.04  12.05  12.01  11.97  11.93  12.00  11.97
12.13  12.07  12.00  11.96  11.99  11.97  12.05  11.94  11.99  12.02
11.95  11.99  11.91  12.06  12.03  12.06  12.05  12.04  12.03  11.98
12.05  12.05  12.11  11.96  12.00  11.96  11.96  12.00  12.01  11.98
;
```

Note that this data set is introduced in "Computing Descriptive Statistics" on page 197 of "PROC CAPABILITY and General Statements" on page 195. The analysis in that section provides evidence that the weight measurements are normally distributed.

By default, the INTERVALS statement computes and prints the six intervals described in the entry for the METHODS= option. The following statements tabulate these intervals for the variable Weight:

```
title 'Statistical Intervals for Fluid Weight';
proc capability data=Cans noprint;
   intervals Weight;
run;
```

The intervals are displayed in Figure 6.27.

Figure 6.25 Statistical Intervals for Weight

Statistical Intervals for Fluid Weight

The CAPABILITY Procedure
Two-Sided Statistical Intervals for Weight Assuming Normality

Approximate Prediction Interval Containing All of k Future Observations		
Confidence	k	Prediction Limits
99.00%	1	11.89 12.13
99.00%	2	11.87 12.14
99.00%	3	11.87 12.15
95.00%	1	11.92 12.10
95.00%	2	11.90 12.12
95.00%	3	11.89 12.12
90.00%	1	11.93 12.09
90.00%	2	11.92 12.10
90.00%	3	11.91 12.11

Figure 6.25 *continued*

Prediction Interval Containing the Mean of k Future Observations

Confidence	k	Prediction Limits	
99.00%	1	11.89	12.13
99.00%	2	11.92	12.10
99.00%	3	11.94	12.08
95.00%	1	11.92	12.10
95.00%	2	11.94	12.08
95.00%	3	11.95	12.06
90.00%	1	11.93	12.09
90.00%	2	11.95	12.06
90.00%	3	11.96	12.05

Tolerance Interval Containing At Least Proportion p of the Population

Confidence	p	Tolerance Limits	
99.00%	0.900	11.92	12.10
99.00%	0.950	11.90	12.12
99.00%	0.990	11.86	12.15
95.00%	0.900	11.92	12.10
95.00%	0.950	11.90	12.11
95.00%	0.990	11.87	12.15
90.00%	0.900	11.92	12.09
90.00%	0.950	11.91	12.11
90.00%	0.990	11.88	12.14

Confidence Limits Containing the Mean

Confidence	Confidence Limits	
99.00%	11.997	12.022
95.00%	12.000	12.019
90.00%	12.002	12.017

Prediction Interval Containing the Standard Deviation of k Future Observations

Confidence	k	Prediction Limits	
99.00%	2	0.0003	0.1348
99.00%	3	0.0033	0.1110
95.00%	2	0.0015	0.1069
95.00%	3	0.0075	0.0919
90.00%	2	0.0030	0.0932
90.00%	3	0.0106	0.0825

Figure 6.25 *continued*

Confidence Limits Containing the Standard Deviation		
Confidence	Confidence Limits	
99.00%	0.040	0.057
95.00%	0.041	0.055
90.00%	0.042	0.053

Computing One-Sided Lower Prediction Limits

NOTE: See *Calculating Various Statistical Intervals* in the SAS/QC Sample Library.

You can specify options after the slash (/) in the INTERVALS statement to control the computation and printing of intervals. The following statements produce a table of one-sided lower prediction limits for the mean, which is displayed in Figure 6.26:

```
title 'Statistical Intervals for Fluid Weight';
proc capability data=Cans noprint;
   intervals Weight / methods = 1 2
                      type    = lower;
run;
```

The METHODS= option specifies which intervals to compute, and the TYPE= option requests one-sided lower limits. All the options available in the INTERVALS statement are listed in "Summary of Options" on page 416 and are described in "Dictionary of Options" on page 417.

Figure 6.26 One-Sided Lower Prediction Limits for the Mean

Statistical Intervals for Fluid Weight

The CAPABILITY Procedure
One-Sided Lower Statistical Intervals for Weight Assuming Normality

Approximate Prediction Limit For All of k Future Observations		
Confidence	k	Lower Limit
99.00%	1	11.90
99.00%	2	11.89
99.00%	3	11.88
95.00%	1	11.93
95.00%	2	11.92
95.00%	3	11.91
90.00%	1	11.95
90.00%	2	11.93
90.00%	3	11.92

Figure 6.26 *continued*

Prediction Limit For the Mean of k Future Observations		
Confidence	k	Lower Limit
99.00%	1	11.90
99.00%	2	11.93
99.00%	3	11.94
95.00%	1	11.93
95.00%	2	11.95
95.00%	3	11.96
90.00%	1	11.95
90.00%	2	11.97
90.00%	3	11.97

Syntax: INTERVALS Statement

The syntax for the INTERVALS statement is as follows:

INTERVALS < *variables* > < / *options* > **;**

You can specify INTERVAL as an alias for INTERVALS. You can use any number of INTERVALS statements in the CAPABILITY procedure. The components of the INTERVALS statement are described as follows.

variables
gives a list of variables for which to compute intervals. If you specify a VAR statement, the variables must also be listed in the VAR statement. Otherwise, the variables can be any numeric variable in the input data set. If you do not specify a list of variables, then by default the INTERVALS statement computes intervals for all variables in the VAR statement (or all numeric variables in the input data set if you do not use a VAR statement).

options
alter the defaults for computing and printing intervals and for creating output data sets.

Summary of Options

The following tables list the INTERVALS statement options by function. For complete descriptions, see "Dictionary of Options" on page 417.

Table 6.49 INTERVAL Statement Options

Option	Description
ALPHA=	specifies probability or confidence levels associated with the intervals
K=	specifies values of k for prediction intervals
METHODS=	specifies which intervals are computed
NOPRINT	suppresses the output tables
OUTINTERVALS=	specifies an output data set containing interval information
P=	specifies values of p for tolerance intervals
TYPE=	specifies the type of intervals (one-sided lower, one-sided upper, or two-sided)

Dictionary of Options

The following entries provide detailed descriptions of options in the INTERVALS statement.

ALPHA=*value-list*

specifies values of α, the probability or confidence associated with the interval. For example, the following statements tabulate the default intervals at probability or confidence levels of $\alpha = 0.05$, $\alpha = 0.10$, $\alpha = 0.15$, and $\alpha = 0.20$:

```
proc capability data=steel;
   intervals width / alpha = 0.05 0.10 0.15 0.20;
run;
```

Note that some references use $\gamma = 1 - \alpha$ to denote probability or confidence levels. Values for the ALPHA= option must be between 0.00001 to 0.99999. By default, values of 0.01, 0.05, and 0.10 are used.

K=*value-list*

specifies values of k for prediction intervals. Default values of 1, 2, and 3 are used for the prediction interval for k future observations and for the prediction interval for the mean of k future observations. Default values of 2 and 3 are used for the prediction interval for the standard deviation of k future observations. The values must be integers.

METHODS=*indices*

METHOD=*indices*

specifies which intervals are computed. The indices can range from 1 to 6, and they correspond to the intervals described in Table 6.50.

Table 6.50 Intervals Computed for METHOD=*Index*

Index	Statistical Interval
1	approximate simultaneous prediction interval for k future observations
2	prediction interval for the mean of k future observations
3	statistical tolerance interval that contains at least proportion p of the population
4	confidence interval for the population mean
5	prediction interval for the standard deviation of k future observations
6	confidence interval for the population standard deviation

For example, the following statements tabulate confidence limits for the population mean (METHOD=4) and confidence limits for the population standard deviation (METHOD=6):

```
proc capability data=steel;
   intervals width / methods=4 6;
run;
```

Formulas for the intervals are given in "Methods for Computing Statistical Intervals" on page 419. By default, the procedure computes all six intervals.

NOPRINT
suppresses the tables produced by default. This option is useful when you only want to save the interval information in an OUTINTERVALS= data set.

OUTINTERVALS=*SAS-data-set*

OUTINTERVAL=*SAS-data-set*

OUTINT=*SAS-data-set*
specifies an output SAS data set containing the intervals and related information. For example, the following statements create a data set named ints containing intervals for the variable width:

```
proc capability data=steel;
   intervals width / outintervals=ints;
run;
```

See "OUTINTERVALS= Data Set" on page 422 for details.

P=*value-list*
specifies values of p for the tolerance intervals. These values must be between 0.00001 to 0.99999. Note that the P= option applies only to the tolerance intervals (METHODS=3). By default, values of 0.90, 0.95, and 0.99 are used.

TYPE=LOWER | UPPER | TWOSIDED
determines whether the intervals computed are one-sided lower, one-sided upper, or two-sided intervals, respectively. See "Computing One-Sided Lower Prediction Limits" on page 415 for an example. The default interval type is TWOSIDED.

Details: INTERVALS Statement

This section provides details on the following topics:

- formulas for statistical intervals
- OUTINTERVALS= data sets

Methods for Computing Statistical Intervals

The formulas for statistical intervals given in this section use the following notation:

Notation	Definition
n	number of nonmissing values for a variable
$\bar{X}$	mean of variable
s	standard deviation of variable
z_α	100αth percentile of the standard normal distribution
$t_\alpha(\nu)$	100αth percentile of the central t distribution with ν degrees of freedom
$t'_\alpha(\delta, \nu)$	100αth percentile of the noncentral t distribution with noncentrality parameter δ and ν degrees of freedom
$F_\alpha(\nu_1, \nu_2)$	100αth percentile of the F distribution with ν_1 degrees of freedom in the numerator and ν_2 degrees of freedom in the denominator
$\chi^2_\alpha(\nu)$	100αth percentile of the χ^2 distribution with ν degrees of freedom
$\chi^2_\alpha(\delta, \nu)$	100αth percentile of the noncentral χ^2 distribution with noncentrality parameter δ and ν degrees of freedom

The values of the variable are assumed to be independent and normally distributed. The intervals are computed using the degrees of freedom as the divisor for the standard deviation s. This divisor corresponds to the default of VARDEF=DF in the PROC CAPABILITY statement. If you specify another value for the VARDEF= option, intervals are not computed.

You select the intervals to be computed with the METHODS= option. The next six sections give computational details for each of the METHODS= options.

METHODS=1

This requests an approximate simultaneous prediction interval for k future observations. Two-sided intervals are computed using the conservative approximations

$$\text{Lower Limit} = \bar{X} - t_{1-\frac{\alpha}{2k}}(n-1)s\sqrt{1+\frac{1}{n}}$$

$$\text{Upper Limit} = \bar{X} + t_{1-\frac{\alpha}{2k}}(n-1)s\sqrt{1+\frac{1}{n}}$$

One-sided limits are computed using the conservative approximation

$$\text{Lower Limit} = \bar{X} - t_{1-\frac{\alpha}{k}}(n-1)s\sqrt{1+\frac{1}{n}}$$

$$\text{Upper Limit} = \bar{X} + t_{1-\frac{\alpha}{k}}(n-1)s\sqrt{1+\frac{1}{n}}$$

Hahn (1970b) states that these approximations are satisfactory except for combinations of small n, large k, and large α. Refer also to Hahn (1969, 1970a) and Hahn and Meeker (1991).

METHODS=2

This requests a prediction interval for the mean of k future observations. Two-sided intervals are computed as

$$\text{Lower Limit} = \bar{X} - t_{1-\frac{\alpha}{2}}(n-1)s\sqrt{\frac{1}{k}+\frac{1}{n}}$$

$$\text{Upper Limit} = \bar{X} + t_{1-\frac{\alpha}{2}}(n-1)s\sqrt{\frac{1}{k}+\frac{1}{n}}$$

One-sided limits are computed as

$$\text{Lower Limit} = \bar{X} - t_{1-\alpha}(n-1)s\sqrt{\frac{1}{k}+\frac{1}{n}}$$

$$\text{Upper Limit} = \bar{X} + t_{1-\alpha}(n-1)s\sqrt{\frac{1}{k}+\frac{1}{n}}$$

METHODS=3

This requests a statistical tolerance interval that contains at least proportion p of the population. Two-sided intervals are computed as

$$\text{Lower Limit} = \bar{X} - ks$$

$$\text{Upper Limit} = \bar{X} + ks$$

where k is the solution of the integral equation

$$\sqrt{\frac{2n}{\pi}}\int_0^\infty P\left(\chi^2_{n-1} > \frac{(n-1)\chi^2_p(z^2,1)}{k^2}\right)e^{-\frac{1}{2}nz^2}dz = 1-\alpha$$

One-sided limits are computed as

$$\text{Lower Limit} = \bar{X} - g'(p\,;n\,;1-\alpha)s$$

$$\text{Upper Limit} = \bar{X} + g'(p\,;n\,;1-\alpha)s$$

where $g'(p\,;n\,;1-\alpha) = \frac{1}{\sqrt{n}}t'_{1-\alpha}(z_p\sqrt{n}, n-1)$.

For a thorough discussion of tolerance intervals and tolerance limits, see Krishnamoorthy and Mathew (2009).

METHODS=4

This requests a confidence interval for the population mean. Two-sided intervals are computed as

$$\text{Lower Limit} = \bar{X} - t_{1-\frac{\alpha}{2}}(n-1)\frac{s}{\sqrt{n}}$$

$$\text{Upper Limit} = \bar{X} + t_{1-\frac{\alpha}{2}}(n-1)\frac{s}{\sqrt{n}}$$

One-sided limits are computed as

$$\text{Lower Limit} = \bar{X} - t_{1-\alpha}(n-1)\frac{s}{\sqrt{n}}$$

$$\text{Upper Limit} = \bar{X} + t_{1-\alpha}(n-1)\frac{s}{\sqrt{n}}$$

METHODS=5

This requests a prediction interval for the standard deviation of k future observations. Two-sided intervals are computed as

$$\text{Lower Limit} = s\left(F_{1-\frac{\alpha}{2}}(n-1,k-1)\right)^{-\frac{1}{2}}$$

$$\text{Upper Limit} = s\left(F_{1-\frac{\alpha}{2}}(k-1,n-1)\right)^{\frac{1}{2}}$$

One-sided limits are computed as

$$\text{Lower Limit} = s\left(F_{1-\alpha}(n-1,k-1)\right)^{-\frac{1}{2}}$$

$$\text{Upper Limit} = s\left(F_{1-\alpha}(k-1,n-1)\right)^{\frac{1}{2}}$$

METHODS=6

This requests a confidence interval for the population standard deviation. Two-sided intervals are computed as

$$\text{Lower Limit} = s\sqrt{\frac{n-1}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}}$$

$$\text{Upper Limit} = s\sqrt{\frac{n-1}{\chi^2_{\frac{\alpha}{2}}(n-1)}}$$

One-sided limits are computed as

$$\text{Lower Limit} = s\sqrt{\frac{n-1}{\chi^2_{1-\alpha}(n-1)}}$$

$$\text{Upper Limit} = s\sqrt{\frac{n-1}{\chi^2_{\alpha}(n-1)}}$$

OUTINTERVALS= Data Set

Each INTERVALS statement can create an output data set specified with the OUTINTERVALS= option. The OUTINTERVALS= data set contains statistical intervals and related parameters.

The number of observations in the OUTINTERVALS= data set depends on the number of variables analyzed, the number of tests specified, and the results of the tests. The OUTINTERVALS= data set is constructed as follows:

- The OUTINTERVALS= data set contains a group of observations for each variable analyzed.
- Each group contains one or more observations for each interval you specify with the METHODS= option. The actual number depends upon the number of combinations of the ALPHA=, K=, and P= values.

The following variables are saved in the OUTINTERVALS= data set:

Variable	Description
ALPHA	value of α associated with the intervals
_K_	value of K= for the prediction intervals
LOWER	lower endpoint of interval
METHOD	interval index (1–6)
_P_	value of P= for the tolerance intervals
TYPE	type of interval (ONESIDED or TWOSIDED)
UPPER	upper endpoint of interval
VAR	variable name

If you use a BY statement, the BY variables are also saved in the OUTINTERVALS= data set.

ODS Tables

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

Table 6.51 ODS Tables Produced with the INTERVALS Statement

Table Name	Description	Option
Intervals1	prediction interval for future observations	METHODS=1
Intervals2	prediction interval for mean	METHODS=2
Intervals3	tolerance interval for proportion of population	METHODS=3
Intervals4	confidence limits for mean	METHODS=4
Intervals5	prediction interval for standard deviation	METHODS=5
Intervals6	confidence limits for standard deviation	METHODS=6

OUTPUT Statement: CAPABILITY Procedure

Overview: OUTPUT Statement

You can use the OUTPUT statement to save summary statistics in a SAS data set. This information can then be used to create customized reports or to save historical information about a process.

You can use options in the OUTPUT statement to

- specify the statistics to save in the output data set
- specify the name of the output data set
- compute and save percentiles not automatically computed by the CAPABILITY procedure

Getting Started: OUTPUT Statement

This section introduces the OUTPUT statement with simple examples that illustrate commonly used options. Complete syntax for the OUTPUT statement is presented in the section "Syntax: OUTPUT Statement" on page 426, and advanced examples are given in the section "Examples: OUTPUT Statement" on page 433.

Saving Summary Statistics in an Output Data Set

NOTE: See *Saving CAPABILITY Output in a Data Set* in the SAS/QC Sample Library.

An automobile manufacturer producing seat belts saves summary information in an output data set with the CAPABILITY procedure. The following statements create the data set Belts, which contains the breaking strengths (Strength) and widths (Width) of a sample of 50 belts:

```
data Belts;
   label Strength = 'Breaking Strength (lb/in)'
         Width    = 'Width in Inches';
   input Strength Width @@;
   datalines;
1243.51  3.036  1221.95  2.995  1131.67  2.983  1129.70  3.019
1198.08  3.106  1273.31  2.947  1250.24  3.018  1225.47  2.980
1126.78  2.965  1174.62  3.033  1250.79  2.941  1216.75  3.037
1285.30  2.893  1214.14  3.035  1270.24  2.957  1249.55  2.958
1166.02  3.067  1278.85  3.037  1280.74  2.984  1201.96  3.002
1101.73  2.961  1165.79  3.075  1186.19  3.058  1124.46  2.929
1213.62  2.984  1213.93  3.029  1289.59  2.956  1208.27  3.029
1247.48  3.027  1284.34  3.073  1209.09  3.004  1146.78  3.061
1224.03  2.915  1200.43  2.974  1183.42  3.033  1195.66  2.995
1258.31  2.958  1136.05  3.022  1177.44  3.090  1246.13  3.022
1183.67  3.045  1206.50  3.024  1195.69  3.005  1223.49  2.971
1147.47  2.944  1171.76  3.005  1207.28  3.065  1131.33  2.984
1215.92  3.003  1202.17  3.058
;
```

The following statements produce two output data sets containing summary statistics:

```
proc capability data=Belts;
   var Strength Width;
   output out=Means    mean=smean wmean;
   output out=Strstats mean=smean std=sstd min=smin max=smax;
run;

proc print data=Means;
run;

proc print data=Strstats;
run;
```

Note that if you specify an OUTPUT statement, you must also specify a VAR statement. You can use multiple OUTPUT statements with a single procedure statement. Each OUTPUT statement creates a new data set. The OUT= option specifies the name of the output data set. In this case, two data sets, Means and Strstats, are created. See Figure 6.27 for a listing of Means and Figure 6.28 for a listing of Strstats.

Summary statistics are saved in an output data set by specifying *keyword=names* after the OUT= option. In the preceding statements, the first OUTPUT statement specifies the keyword MEAN followed by the names smean and wmean. The second OUTPUT statement specifies the keywords MEAN, STD, MIN, and MAX, for which the names smean, sstd, smin, and smax are given.

The keyword specifies the statistic to be saved in the output data set, and the names determine the names for the new variables. The first name listed after a keyword contains that statistic for the first variable listed in the VAR statement; the second name contains that statistic for the second variable in the VAR statement, and so on.

Thus, the data set Means contains the mean of Strength in a variable named smean and the mean of Width in a variable named wmean. The data set Strstats contains the mean, standard deviation, minimum value, and maximum value of Strength in the variables smean, sstd, smin, and smax, respectively.

Figure 6.27 Listing of the Output Data Set Means

Statistical Intervals for Fluid Weight

Obs	smean	wmean
1	1205.75	3.00584

Figure 6.28 Listing of the Output Data Set Strstats

Statistical Intervals for Fluid Weight

Obs	smean	sstd	smax	smin
1	1205.75	48.3290	1289.59	1101.73

Saving Percentiles in an Output Data Set

NOTE: See *Saving CAPABILITY Output in a Data Set* in the SAS/QC Sample Library.

The CAPABILITY procedure automatically computes the 1st, 5th, 10th, 25th, 75th, 90th, 95th, and 99th percentiles for each variable. You can save these percentiles in an output data set by specifying the appropriate keywords. For example, the following statements create an output data set named Pctlstr containing the 5th and 95th percentiles of the variable Strength:

```
proc capability data=Belts noprint;
   var Strength Width;
   output out=Pctlstr p5=p5str p95=p95str;
run;

proc print data=Pctlstr;
run;
```

The output data set Pctlstr is listed in Figure 6.29.

Figure 6.29 Listing of the Output Data Set Pctlstr

Statistical Intervals for Fluid Weight

Obs	p95str	p5str
1	1284.34	1126.78

You can use the PCTLPTS=, PCTLPRE=, and PCTLNAME= options to save percentiles not automatically computed by the CAPABILITY procedure. For example, the following statements create an output data set named Pctls containing the 20th and 40th percentiles of the variables Strength and Width:

```
proc capability data=Belts noprint;
   var Strength Width;
   output out=Pctls pctlpts  = 20 40
                    pctlpre  = S W
                    pctlname = pct20 pct40;
run;

proc print data=Pctls;
run;
```

The PCTLPTS= option specifies the percentiles to compute (in this case, the 20th and 40th percentiles). The PCTLPRE= and PCTLNAME= options build the names for the variables containing the percentiles. The PCTLPRE= option gives prefixes for the new variables, and the PCTLNAME= option gives a suffix to add to the prefix. Note that if you use the PCTLPTS= specification, you must also use the PCTLPRE= specification. For details on these options, see the section "Syntax: OUTPUT Statement" on page 426.

The preceding OUTPUT statement saves the 20th and 40th percentiles of Strength and Width in the variables spct20, wpct20, spct40, and wpct40. The output data set Pctls is listed in Figure 6.30.

Figure 6.30 Listing of the Output Data Set Pctls

Statistical Intervals for Fluid Weight

Obs	Spct20	Wpct20	Spct40	Wpct40
1	1165.91	2.9595	1199.26	2.995

Syntax: OUTPUT Statement

The syntax for the OUTPUT statement is as follows:

OUTPUT < **OUT=***SAS-data-set* > < *keyword1=names . . . keywordk=names* > < *percentile-options* > **;**

You can use any number of OUTPUT statements in the CAPABILITY procedure. Each OUTPUT statement creates a new data set containing the statistics specified in that statement. When you use the OUTPUT statement, you must also use the VAR statement. In addition, the OUTPUT statement must contain at least one of the following:

- a specification of the form *keyword=names*
- the PCTLPTS= and PCTLPRE= options

You can use the OUT= option to specify the name of the output data set:

OUT=*SAS-data-set*
specifies the name of the output data set. To create a permanent SAS data set, specify a two-level name. See *SAS DATA Step Statements: Reference* for more information on permanent SAS data sets. For example, the previous statements create an output data set named Summary. If the OUT= option is omitted, then by default the new data set is named using the DATA*n* convention.

A *keyword=names* specification selects a statistic to be included in the output data set and gives names to the new variables that contain the statistics. Specify a *keyword* for each desired statistic, an equal sign, and the *names* of the variables to contain the statistic.

In the output data set, the first variable listed after a keyword in the OUTPUT statement contains the statistic for the first variable listed in the VAR statement; the second variable contains the statistic for the second variable in the VAR statement, and so on. The list of *names* following the equal sign can be shorter than the list of variables in the VAR statement. In this case, the procedure uses the *names* in the order in which the variables are listed in the VAR statement. Consider the following example:

```
proc capability noprint;
   var length width height;
   output out=summary mean=mlength mwidth;
run;
```

The variables mlength and mwidth contain the means for length and width. The mean for height is computed by the procedure but is not saved in the output data set.

Table 6.52 lists all keywords available in the OUTPUT statement grouped by type. Formulas for selected statistics are given in the section "Details: CAPABILITY Procedure" on page 219.

Table 6.52 OUTPUT Statement Statistic Keywords

Keyword	Description
Descriptive Statistics	
CSS	Sum of squares corrected for the mean
CV	Percent coefficient of variation

Table 6.52 (continued)

Keyword	Description
GEOMEAN	Geometric mean
KURTOSIS | KURT	Kurtosis
MAX	Largest (maximum) value
MEAN	Mean
MIN	Smallest (minimum) value
MODE	Most frequent value (if not unique, the smallest mode)
N	Number of observations on which calculations are based
NMISS	Number of missing values
NOBS	Number of observations
RANGE	Range
SKEWNESS | SKEW	Skewness
STD | STDDEV	Standard deviation
STDMEAN | STDERR	Standard error of the mean
SUM	Sum
SUMWGT	Sum of weights
USS	Uncorrected sum of squares
VAR	Variance
Quantile Statistics	
MEDIAN | P50 | Q2	Median (50th percentile)
P1	1st percentile
P5	5th percentile
P10	10th percentile
P90	90th percentile
P95	95th percentile
P99	99th percentile
Q1 | P25	Lower quartile (25th percentile)
Q3 | P75	Upper quartile (75th percentile)
QRANGE	Interquartile range (Q3 – Q1)
Robust Statistics	
GINI	Gini's mean difference
MAD	Median absolute difference
QN	2nd variation of median absolute difference
SN	1st variation of median absolute difference
STD_GINI	Standard deviation for Gini's mean difference
STD_MAD	Standard deviation for median absolute difference
STD_QN	Standard deviation for the second variation of the median absolute difference
STD_QRANGE	Estimate of the standard deviation, based on interquartile range
STD_SN	Standard deviation for the first variation of the median absolute difference

Table 6.52 (continued)

Keyword	Description
Hypothesis Test Statistics	
MSIGN	Sign statistic
NORMAL	Test statistic for normality. If the sample size is less than or equal to 2000, this is the Shapiro-Wilk W statistic. Otherwise, it is the Kolmogorov D statistic.
PNORMAL | PROBN	p-value for normality test
PROBM	Probability of a greater absolute value for the sign statistic
PROBS	Probability of a greater absolute value for the signed rank statistic
PROBT	Two-tailed p-value for Student's t statistic with $n-1$ degrees of freedom
SIGNRANK	Signed rank statistic
T	Student's t statistic to test the null hypothesis that the population mean is equal to μ_0
Specification Limits and Related Statistics	
LSL	Lower specification limit
PCTGTR	Percent of nonmissing observations greater than the upper specification limit
PCTLSS	Percent of nonmissing observations less than the lower specification limit
TARGET	Target value
USL	Upper specification limit
Capability Indices and Related Statistics	
CP	Capability index C_p
CPLCL	Lower confidence limit for C_p
CPUCL	Upper confidence limit for C_p
CPK	Capability index C_{pk} (also denoted CPK)
CPKLCL	Lower confidence limit for C_{pk}
CPKUCL	Upper confidence limit for C_{pk}
CPL	Capability index CPL
CPLLCL	Lower confidence limit for CPL
CPLUCL	Upper confidence limit for CPL
CPM	Capability index C_{pm}
CPMLCL	Lower confidence limit for C_{pm}
CPMUCL	Upper confidence limit for C_{pm}
CPU	Capability index CPU
CPULCL	Lower confidence limit for CPU
CPUUCL	Upper confidence limit for CPU
K	Capability index k (also denoted K)

The CAPABILITY procedure automatically computes the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th percentiles for the data. You can save these statistics in an output data set by using *keyword=names* specifications. You can request additional percentiles by using the PCTLPTS= option. The following *percentile-options* are related to these additional percentiles:

CIPCTLDF=(*cipctl-options***)**
CIQUANTDF=(*cipctl-options***)**

requests distribution-free confidence limits for percentiles that are requested with the PCTLPTS= option. In other words, no specific parametric distribution such as the normal is assumed for the data. PROC CAPABILITY uses order statistics (ranks) to compute the confidence limits as described by Hahn and Meeker (1991). This option does not apply if you use a WEIGHT statement. You can specify the following *cipctl-options*:

ALPHA=*α*

specifies the level of significance α for $100(1-\alpha)\%$ confidence intervals. The value α must be between 0 and 1; the default value is 0.05, which results in 95% confidence intervals. The default value is the value of ALPHA= given in the PROC statement.

LOWERPRE=*prefixes*

specifies one or more prefixes that are used to create names for variables that contain the lower confidence limits. To save lower confidence limits for more than one analysis variable, specify a list of prefixes. The order of the prefixes corresponds to the order of the analysis variables in the VAR statement.

LOWERNAME=*suffixes*

specifies one or more suffixes that are used to create names for variables that contain the lower confidence limits. PROC CAPABILITY creates a variable name by combining the LOWERPRE= value and suffix name. Because the suffixes are associated with the requested percentiles, list the suffixes in the same order as the PCTLPTS= percentiles.

TYPE=*keyword*

specifies the type of confidence limit, where *keyword* is LOWER, UPPER, SYMMETRIC, or ASYMMETRIC. The default value is SYMMETRIC.

UPPERPRE=*prefixes*

specifies one or more prefixes that are used to create names for variables that contain the upper confidence limits. To save upper confidence limits for more than one analysis variable, specify a list of prefixes. The order of the prefixes corresponds to the order of the analysis variables in the VAR statement.

UPPERNAME=*suffixes*

specifies one or more suffixes that are used to create names for variables that contain the upper confidence limits. PROC CAPABILITY creates a variable name by combining the UPPERPRE= value and suffix name. Because the suffixes are associated with the requested percentiles, list the suffixes in the same order as the PCTLPTS= percentiles.

NOTE: See the entries for the PCTLPTS=, PCTLPRE=, and PCTLNAME= options for a detailed description of how variable names are created using prefixes, percentile values, and suffixes.

CIPCTLNORMAL=(*cipctl-options***)**
CIQUANTNORMAL=(*cipctl-options***)**

requests confidence limits based on the assumption that the data are normally distributed for percentiles that are requested with the PCTLPTS= option. The computational method is described in Section 4.4.1 of Hahn and Meeker (1991) and uses the noncentral *t* distribution as given by Odeh and Owen (1980). This option does not apply if you use a WEIGHT statement. You can specify the following *cipctl-options*:

ALPHA=α
specifies the level of significance α for $100(1-\alpha)\%$ confidence intervals. The value α must be between 0 and 1; the default value is 0.05, which results in 95% confidence intervals. The default value is the value of ALPHA= given in the PROC statement.

LOWERPRE=*prefixes*
specifies one or more prefixes that are used to create names for variables that contain the lower confidence limits. To save lower confidence limits for more than one analysis variable, specify a list of prefixes. The order of the prefixes corresponds to the order of the analysis variables in the VAR statement.

LOWERNAME=*suffixes*
specifies one or more suffixes that are used to create names for variables that contain the lower confidence limits. PROC CAPABILITY creates a variable name by combining the LOWERPRE= value and suffix name. Because the suffixes are associated with the requested percentiles, list the suffixes in the same order as the PCTLPTS= percentiles.

TYPE=*keyword*
specifies the type of confidence limit, where *keyword* is LOWER, UPPER, or TWOSIDED. The default is TWOSIDED.

UPPERPRE=*prefixes*
specifies one or more prefixes that are used to create names for variables that contain the upper confidence limits. To save upper confidence limits for more than one analysis variable, specify a list of prefixes. The order of the prefixes corresponds to the order of the analysis variables in the VAR statement.

UPPERNAME=*suffixes*
specifies one or more suffixes that are used to create names for variables that contain the upper confidence limits. PROC CAPABILITY creates a variable name by combining the UPPERPRE= value and suffix name. Because the suffixes are associated with the requested percentiles, list the suffixes in the same order as the PCTLPTS= percentiles.

NOTE: See the entries for the PCTLPTS=, PCTLPRE=, and PCTLNAME= options for a detailed description of how variable names are created using prefixes, percentile values, and suffixes.

PCTLGROUP=BYSTAT | BYVAR
specifies the order in which variables that you request with the PCTLPTS= option are added to the OUT= data set when the VAR statement lists more than one analysis variable. By default (or if you specify PCTLGROUP=BYSTAT), all variables that are associated with a percentile value are created consecutively. If you specify PCTLGROUP=BYVAR, all variables that are associated with an analysis variable are created consecutively.

Consider the following statements:

```
proc univariate data=Score;
   var PreTest PostTest;
   output out=ByStat pctlpts=20 40 pctlpre=Pre_ Post_;
   output out=ByVar pctlgroup=byvar pctlpts=20 40 pctlpre=Pre_ Post_;
run;
```

The order of variables in the data set ByStat is Pre_20, Post_20, Pre_40, Post_40. The order of variables in the data set ByVar is Pre_20, Pre_40, Post_20, Post_40.

PCTLNAME=*suffixes*

provides name suffixes for the new variables created by the PCTLPTS= option. These suffixes are appended to the prefixes you specify with the PCTLPRE= option, replacing the percentile values that are used as suffixes by default. List the suffixes in the same order in which you specify the percentiles. If you specify n suffixes with the PCTLNAME= option and m percentile values with the PCTLPTS= option, where $m > n$, the suffixes are used to name the first n percentiles, and the default names are used for the remaining $m - n$ percentiles. For example, consider the following statements:

```
proc capability;
   var length width height;
   output pctlpts  = 20 40
          pctlpre  = pl pw ph
          pctlname = twenty;
run;
```

The value "twenty" in the PCTLNAME= option is used for only the first percentile in the PCTLPTS= list. This suffix is appended to the values in the PCTLPRE= option to generate the new variable names pltwenty, pwtwenty, and phtwenty, which contain the 20th percentiles for length, width, and height, respectively. Because a second PCTLNAME= suffix is not specified, variable names for the 40th percentiles for length, width, and height are generated using the prefixes and percentile values. Thus, the output data set contains the variables pltwenty, pl40, pwtwenty, pw40, phtwenty, and ph40.

PCTLNDEC=*value*

specifies the number of decimal places in percentile values that are incorporated into percentile variable names. The default value is 1. For example, the following statements create two output data sets, each containing one percentile variable. The variable in data set short is named pwid85_1, while the one in data set long is named pwid85_125.

```
proc capability;
   var width;
   output out=short pctlpts=85.125 pctlpre=pwid;
   output out=long  pctlpts=85.125 pctlpre=pwid pctlndec=3;
run;
```

PCTLPRE=*prefixes*

specifies prefixes used to create variable names for percentiles requested with the PCTLPTS= option. The PCTLPRE= and PCTLPTS= options must be used together.

The procedure generates new variable names by using the prefix and the percentile values. If the specified percentile is an integer, the variable name is simply the prefix followed by the value. For noninteger percentiles, an underscore replaces the decimal point in the variable name, and decimal values are truncated to one decimal place. For example, the following statements create the variables pwid20, pwid33_3, pwid66_6, and pwid80 for the 20th, 33.33rd, 66.67th, and 80th percentiles of width, respectively:

```
proc capability noprint;
   var width;
   output pctlpts=20 33.33 66.67 80 pctlpre=pwid;
run;
```

If you request percentiles for more than one variable, you should list prefixes in the same order in which the variables appear in the VAR statement. For example, the following statements compute the 80th and 87.5th percentiles for length and width and save the new variables plength80, plength87_5, pwidth80, and pwidth87_5 in the output data set:

```
proc capability noprint;
   var length width;
   output pctlpts=80 87.5 pctlpre=plength pwidth;
run;
```

PCTLPTS=*percentiles*

specifies *percentiles* that are not automatically computed by the procedure. The CAPABILITY procedure automatically computes the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th percentiles for the data. These can be saved in an output data set by using keyword=names specifications. The PCTLPTS= option generates additional percentiles and outputs them to a data set; these additional percentiles are not printed.

If you use the PCTLPTS= option, you must also use the PCTLPRE= option to provide a prefix for the new variable names. For example, to create variables that contain the 20th, 40th, 60th, and 80th percentiles of length, use the following statements:

```
proc capability noprint;
   var length;
   output pctlpts=20 40 60 80 pctlpre=plen;
run;
```

This creates the variables plen20, plen40, plen60, and plen80, whose values are the corresponding percentiles of length. In addition to specifying name prefixes with the PCTLPRE= option, you can also use the PCTLNAME= option to create name suffixes for the new variables created by the PCTLPTS= option.

Details: OUTPUT Statement

OUT= Data Set

The CAPABILITY procedure creates an OUT= data set for each OUTPUT statement. The new data set contains an observation for each combination of levels of the variables in the BY and CLASS statements, or a single observation if you do not specify a BY or CLASS statement. Thus, the number of observations in the new data set corresponds to the number of groups for which statistics are calculated. The variables in the new data set are as follows:

- variables in the BY statement. The values of these variables match the values in the corresponding BY group in the DATA= data set.
- variables in the CLASS statement. The values of these variables identify the CLASS level within a BY group in the DATA= data set that from which statistics are computed.
- variables in the ID statement. The values of these variables match those for the first observation in each BY group, or for the first observation in the data set if you do not specify a BY statement.
- variables created by selecting statistics in the OUTPUT statement. The values of the statistics are computed using all the nonmissing data, or statistics are computed for each BY group if you use a BY statement.
- variables created by requesting new percentiles with the PCTLPTS= option. The names of these new variables depend on the values of the PCTLPRE= and PCTLNAME= options.

If the output data set contains a percentile variable or a quartile variable, the percentile definition assigned with the PCTLDEF= option in the PROC CAPABILITY statement is recorded on the output data set label.

The values of variables requested with the statistics keywords CP, CPK, CPL, CPM, CPU, K, PCTGTR, and PCTLSS are missing unless you identify specification limits in a SPEC statement or in a SPEC= data set.

As an alternative to OUT= data sets, you can create an OUTTABLE= data set. The structure of the OUTTABLE= data set may be more appropriate when you are computing summary statistics and capability indices for multiple process variables. See "OUTTABLE= Data Set" on page 222.

Examples: OUTPUT Statement

This section provides additional examples of the OUTPUT statement.

Example 6.17: Computing Nonstandard Capability Indices

NOTE: See *Computing Nonstandard Capability Indices* in the SAS/QC Sample Library.

In recent years, a number of process capability indices that have been proposed in the research literature are gradually being introduced in applications. As shown in this example, you can compute such indices in the DATA step after using the OUTPUT statement in the CAPABILITY procedure to save various summary statistics.

Hardness measurements (in scaled units) for 50 titanium samples are saved as values of the variable Hardness in the following SAS data set:

```
data Titanium;
   label Hardness = 'Hardness Measurement';
   input Hardness @@;
   datalines;
1.38  1.49  1.43  1.60  1.59
1.34  1.44  1.64  1.83  1.57
1.45  1.74  1.61  1.39  1.63
```

```
1.73  1.61  1.35  1.51  1.47
1.46  1.41  1.56  1.40  1.58
1.43  1.53  1.53  1.58  1.62
1.58  1.46  1.26  1.57  1.41
1.53  1.36  1.63  1.36  1.66
1.49  1.55  1.67  1.41  1.39
1.75  1.37  1.36  1.86  1.49
;
```

The target value for hardness is 1.6, and the lower and upper specification limits are 0.8 and 2.4, respectively. The samples are produced by an in-control process, and the measurements are assumed to be normally distributed.

The following statements use the OUTPUT statement to save various descriptive statistics and an estimate of the index C_{pm} in a data set named Indices:

```
proc capability data=Titanium noprint;
   var Hardness;
   specs lsl=0.8 target=1.6 usl=2.4;
   output out=Indices
      n       = n
      mean    = avg
      std     = std
      var     = var
      lsl     = lsl
      target  = t
      usl     = usl
      pnormal = pnormal
      cpm     = cpm ;
run;
```

In addition to C_{pm}, you want to report an estimate for the index C_{pmk}, which is defined as follows:

$$C_{pmk} = \frac{d - |\mu - m|}{3\sqrt{\sigma^2 + (\mu - T)^2}}$$

where $d = (\text{USL} - \text{LSL})/2$, $m = (\text{USL} + \text{LSL})/2$, and μ and σ are the mean and standard deviation of the normal distribution. Refer to Section 3.6 of Kotz and Johnson (1993). A natural estimator for C_{pmk} is

$$\hat{C}_{pmk} = \frac{d - |\overline{X} - m|}{3\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - T)^2}}$$

The following statements compute this estimate:

```
data Indices;
   set Indices;
   d    = 0.5*( USL - LSL );
   m    = 0.5*( USL + LSL );
   num  = d - abs( avg - m );
   den  = 3 * sqrt( (n-1)*var/n + (avg-t)*(avg-t) );
   cpmk = num/den;
run;
```

```
title "Capability Analysis of Titanium Hardness";
proc print data=Indices noobs;
   var n avg std lsl t usl cpm cpmk pnormal;
run;
```

The results are listed in Output 6.17.1.

Output 6.17.1 Computation of C_{pmk}

Capability Analysis of Titanium Hardness

n	avg	std	lsl	t	usl	cpm	cpmk	pnormal
50	1.5212	0.13295	0.8	1.6	2.4	1.72545	1.56713	0.25111

Note that the p-value for the Kolmogorov-Smirnov test of normality is 0.25111, indicating that the assumption of normality is justified.

The following statements also compute an estimate of the index C_{pm} by using the SPECIALINDICES option:

```
proc capability data=Titanium specialindices;
   var Hardness;
   specs lsl=0.8 target=1.6 usl=2.4;
run;
```

Output 6.17.2 Computation of C_{pmk} by Using the SPECIALINDICES Option

Capability Analysis of Titanium Hardness

The CAPABILITY Procedure
Variable: Hardness (Hardness Measurement)

Process Capability Indices			
Index	Value	95% Confidence Limits	
Cp	2.005745	1.609575	2.401129
CPL	1.808179	1.438675	2.175864
CPU	2.203311	1.757916	2.646912
Cpk	1.808179	1.438454	2.177904
Cpm	1.725446	1.410047	2.066027

Example 6.18: Approximate Confidence Limits for Cpk

NOTE: See *Approximate Confidence Limits for Cpk* in the SAS/QC Sample Library.

This example illustrates how you can use the OUTPUT statement to compute confidence limits for the capability index C_{pk}.

You can request the approximate confidence limits given by Bissell (1990) with the keywords CPKLCL and CPKUCL in the OUTPUT statement. However, this is not the only method that has been proposed for computing confidence limits for C_{pk}. Zhang, Stenback, and Wardrop (1990), referred to here as ZSW,

proposed approximate confidence limits of the form

$$\widehat{C}_{pk} \pm k\widehat{\sigma}_{pk}$$

where $\widehat{\sigma}_{pk}$ is an estimator of the standard deviation of $\widehat{C}_{pk}$. Equation (8) of ZSW provides an approximation to the variance of $\widehat{C}_{pk}$ from which one can obtain $100\gamma\%$ confidence limits for C_{pk} as

$$\begin{aligned}
\text{LCL} &= \widehat{C}_{pk}\left[1 - \Phi^{-1}((1-\gamma)/2)\sqrt{\frac{n-1}{n-3} - \frac{(n-1)\Gamma^2((n-2)/2)}{2\Gamma^2((n-1)/2)}}\right] \\
\text{UCL} &= \widehat{C}_{pk}\left[1 + \Phi^{-1}(1-(1-\gamma)/2)\sqrt{\frac{n-1}{n-3} - \frac{(n-1)\Gamma^2((n-2)/2)}{2\Gamma^2((n-1)/2)}}\right]
\end{aligned}$$

This assumes that $\widehat{C}_{pk}$ is normally distributed. You can also compute approximate confidence limits based on equation (6) of ZSW, which provides an exact expression for the variance of $\widehat{C}_{pk}$.

The following program uses the methods of Bissell (1990) and ZSW to compute approximate confidence limits for C_{pk} for the variable Hardness in the data set Titanium (see Example 6.17).

```
proc capability data=Titanium noprint;
   var Hardness;
   specs lsl=0.8 usl=2.4;
   output out=Summary
       n    = n
       mean = mean
       std  = std
       lsl  = lsl
       usl  = usl
       cpk  = cpk
       cpklcl = cpklcl
       cpkucl = cpkucl
       cpl  = cpl
       cpu  = cpu ;

data Summary;
   set Summary;
   length Method $ 16;

   Method = "Bissell";
   lcl = cpklcl;
   ucl = cpkucl;
   output;

   * Assign confidence level;
   level = 0.95;
   aux   = probit( 1 - (1-level)/2 );

   Method = "ZSW Equation 6";
```

```
   zsw = log(0.5*n-0.5)
         + ( 2*(lgamma(0.5*n-1)-lgamma(0.5*n-0.5)) );
   zsw = sqrt((n-1)/(n-3)-exp(zsw));
   lcl = cpk*(1-aux*zsw);
   ucl = cpk*(1+aux*zsw);
   output;

   Method = "ZSW Equation 8";
   ds = 3*(cpu+cpl)/2;
   ms = 3*(cpl-cpu)/2;
   f1 = (1/3)*sqrt((n-1)/2)*gamma((n-2)/2)*(1/gamma((n-1)/2));
   f2 = sqrt(2/n)*(1/gamma(0.5))*exp(-n*0.5*ms*ms);
   f3 = ms*(1-(2*probnorm(-sqrt(n)*ms)));
   ex = f1*(ds-f2-f3);
   sd = ((n-1)/(9*(n-3)))*(ds**2-(2*ds*(f2+f3))+ms**2+(1/n));
   sd = sd-(ex*ex);
   sd = sqrt(sd);
   lcl = cpk-aux*sd;
   ucl = cpk+aux*sd;
   output;
run;

title "Approximate 95% Confidence Limits for Cpk";
proc print data = Summary noobs;
   var Method lcl cpk ucl;
run;
```

The results are shown in Output 6.18.1.

Output 6.18.1 Approximate Confidence Limits for C_{pk}

Approximate 95% Confidence Limits for Cpk

Method	lcl	cpk	ucl
Bissell	1.43845	1.80818	2.17790
ZSW Equation 6	1.43596	1.80818	2.18040
ZSW Equation 8	1.42419	1.80818	2.19217

Note that there is fairly close agreement in the three methods.

You can display the confidence limits computed using Bissell's approach on plots produced by the CAPABILITY procedure by specifying the keywords CPKLCL and CPKUCL in the INSET statement.

The following statements also compute an estimate of the index C_{pk} along with approximate limits by using the SPECIALINDICES option:

```
proc capability data=Titanium specialindices;
   var Hardness;
   specs lsl=0.8 usl=2.4;
run;
```

Output 6.18.2 Approximate Confidence Limits for C_{pk} using the SPECIALINDICES option

Approximate 95% Confidence Limits for Cpk

The CAPABILITY Procedure
Variable: Hardness (Hardness Measurement)

Process Capability Indices			
Index	**Value**	**95% Confidence Limits**	
Cp	2.005745	1.609575	2.401129
CPL	1.808179	1.438675	2.175864
CPU	2.203311	1.757916	2.646912
Cpk	1.808179	1.438454	2.177904

PPPLOT Statement: CAPABILITY Procedure

Overview: PPPLOT Statement

The PPPLOT statement creates a probability-probability plot (also referred to as a P-P plot or percent plot), which compares the empirical cumulative distribution function (ecdf) of a variable with a specified theoretical cumulative distribution function such as the normal. If the two distributions match, the points on the plot form a linear pattern that passes through the origin and has unit slope. Thus, you can use a P-P plot to determine how well a theoretical distribution models a set of measurements.

You can specify one of the following theoretical distributions with the PPPLOT statement:

- beta
- exponential
- gamma
- Gumbel
- inverse Gaussian
- lognormal
- normal
- generalized Pareto
- power function
- Rayleigh
- Weibull

You can use options in the PPPLOT statement to do the following:

- specify or estimate parameters for the theoretical distribution
- request graphical enhancements

You can also create a comparative P-P plot by using the PPPLOT statement in conjunction with a CLASS statement.

You have three alternatives for producing P-P plots with the PPPLOT statement:

- ODS Graphics output is produced if ODS Graphics is enabled, for example by specifying the ODS GRAPHICS ON statement prior to the PROC statement.
- Otherwise, traditional graphics are produced by default if SAS/GRAPH is licensed.
- Legacy line printer charts are produced when you specify the LINEPRINTER option in the PROC statement.

See Chapter 4, "SAS/QC Graphics," for more information about producing these different kinds of graphs.

NOTE: Probability-probability plots should not be confused with probability plots, which compare a set of ordered measurements with *percentiles* from a specified distribution. You can create probability plots with the PROBPLOT statement.

Getting Started: PPPLOT Statement

The following example illustrates the basic syntax of the PPPLOT statement. For complete details of the PPPLOT statement, see the section "Syntax: PPPLOT Statement" on page 441.

Creating a Normal Probability-Probability Plot

NOTE: See *Creating P-P Plots* in the SAS/QC Sample Library.

The distances between two holes cut into 50 steel sheets are measured and saved as values of the variable Distance in the following data set: [3]

[3] These data are also used to create Q-Q plots in "QQPLOT Statement: CAPABILITY Procedure" on page 492.

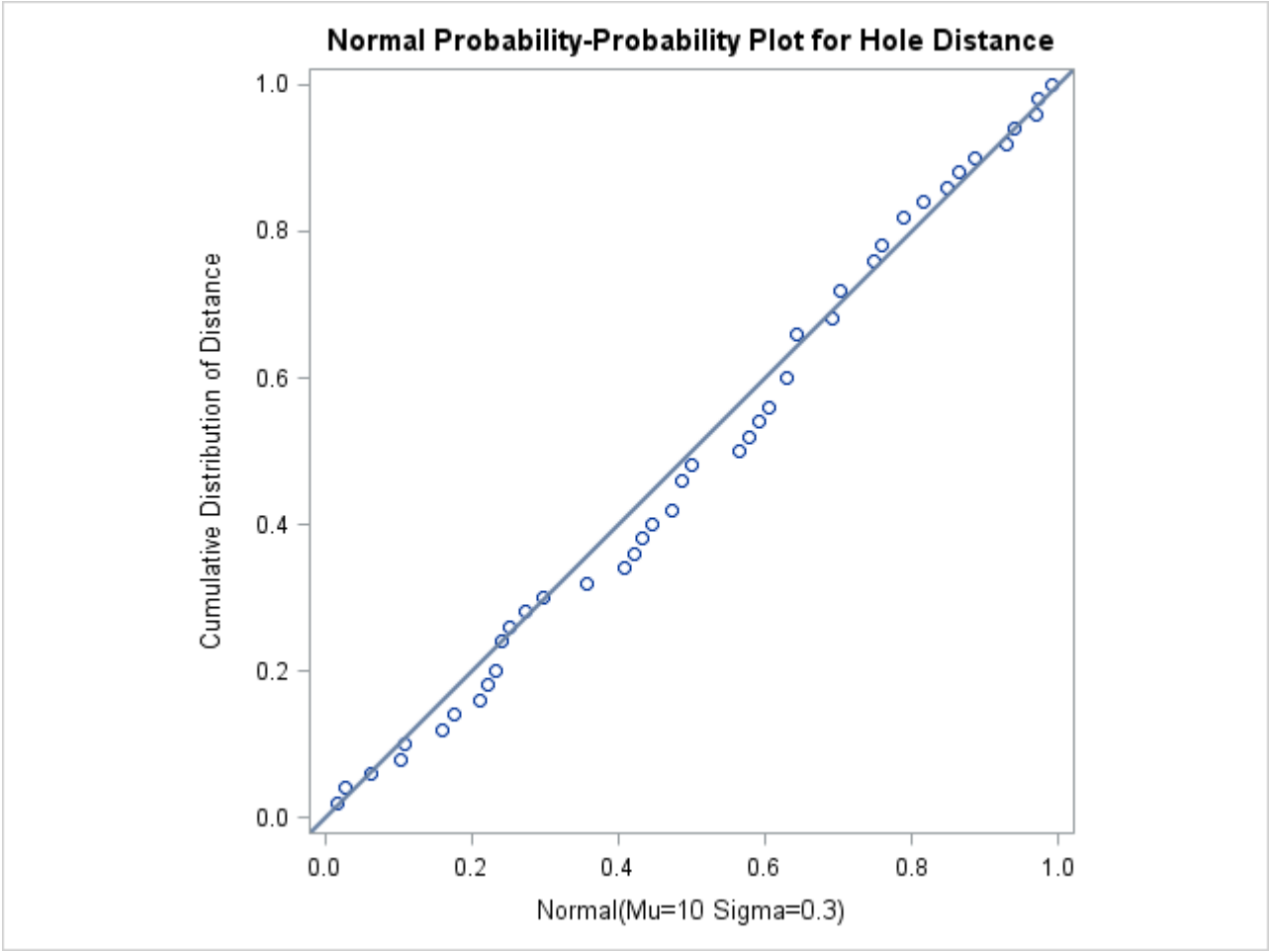
```
data Sheets;
   input Distance @@;
   label Distance='Hole Distance in cm';
   datalines;
 9.80 10.20 10.27  9.70  9.76
10.11 10.24 10.20 10.24  9.63
 9.99  9.78 10.10 10.21 10.00
 9.96  9.79 10.08  9.79 10.06
10.10  9.95  9.84 10.11  9.93
10.56 10.47  9.42 10.44 10.16
10.11 10.36  9.94  9.77  9.36
 9.89  9.62 10.05  9.72  9.82
 9.99 10.16 10.58 10.70  9.54
10.31 10.07 10.33  9.98 10.15
;
```

The cutting process is in statistical control. As a preliminary step in a capability analysis of the process, it is decided to check whether the distances are normally distributed. The following statements create a P-P plot, shown in Figure 6.31, which is based on the normal distribution with mean $\mu = 10$ and standard deviation $\sigma = 0.3$:

```
title 'Normal Probability-Probability Plot for Hole Distance';
proc capability data=Sheets noprint;
   ppplot Distance / normal(mu=10 sigma=0.3)
                     square
                     odstitle=title;
run;
```

The NORMAL option in the PPPLOT statement requests a P-P plot based on the normal cumulative distribution function, and the MU= and SIGMA= *normal-options* specify μ and σ. Note that a P-P plot is always based on a *completely specified* distribution, in other words, a distribution with specific parameters. In this example, if you did not specify the MU= and SIGMA= *normal-options*, the sample mean and sample standard deviation would be used for μ and σ.

Figure 6.31 Normal P-P Plot with Diagonal Reference Line



The linearity of the pattern in Figure 6.31 is evidence that the measurements are normally distributed with mean 10 and standard deviation 0.3. The SQUARE option displays the plot in a square format.

Syntax: PPPLOT Statement

The syntax for the PPPLOT statement is as follows:

PPPLOT < *variables* > < / *options* > **;**

You can specify the keyword PP as an alias for PPPLOT, and you can use any number of PPPLOT statements in the CAPABILITY procedure. The components of the PPPLOT statement are described as follows.

variables

are the process variables for which to create P-P plots. If you specify a VAR statement, the variables must also be listed in the VAR statement. Otherwise, the variables can be any numeric variables in the input data set. If you do not specify a list of variables, then by default, the procedure creates a P-P plot for each variable listed in the VAR statement or for each numeric variable in the input data set if you do not specify a VAR statement. For example, each of the following PPPLOT statements produces two P-P plots, one for length and one for width:

```
proc capability data=measures;
   var length width;
   ppplot;
run;

proc capability data=measures;
   ppplot length width;
run;
```

options

specify the theoretical distribution for the plot or add features to the plot. If you specify more than one variable, the options apply equally to each variable. Specify all options after the slash (/) in the PPPLOT statement. You can specify only one option naming a distribution, but you can specify any number of other options. The distributions available are the beta, exponential, gamma, Gumbel, inverse Gaussian, lognormal, normal, generalized Pareto, power function, Rayleigh, and Weibull. By default, the procedure produces a P-P plot based on the normal distribution.

In the following example, the NORMAL, MU= and SIGMA= options request a P-P plot based on the normal distribution with mean 10 and standard deviation 0.3. The SQUARE option displays the plot in a square frame, and the CTEXT= option specifies the text color.

```
proc capability data=measures;
   ppplot length width / normal(mu=10 sigma=0.3)
                         square
                         ctext=blue;
run;
```

Summary of Options

The following tables list the PPPLOT statement options by function. For complete descriptions, see the section "Dictionary of Options" on page 446.

Distribution Options

Table 6.53 summarizes the options for requesting a specific theoretical distribution.

Table 6.53 Options for Specifying a Theoretical Distribution

Option	Description
BETA(*beta-options*)	specifies beta P-P plot
EXPONENTIAL(*exponential-options*)	specifies exponential P-P plot
GAMMA(*gamma-options*)	specifies gamma P-P plot
GUMBEL(*Gumbel-options*)	specifies Gumbel P-P plot
IGAUSS(*iGauss-options*)	specifies inverse Gaussian P-P plot
LOGNORMAL(*lognormal-options*)	specifies lognormal P-P plot
NORMAL(*normal-options*)	specifies normal P-P plot
PARETO(*Pareto-options*)	specifies generalized Pareto P-P plot
POWER(*power-options*)	specifies power function P-P plot
RAYLEIGH(*Rayleigh-options*)	specifies Rayleigh P-P plot
WEIBULL(*Weibull-options*)	specifies Weibull P-P plot

Table 6.54 summarizes options that specify distribution parameters and control the display of the diagonal distribution reference line. Specify these options in parentheses after the distribution option. For example, the following statements use the NORMAL option to request a normal P-P plot:

```
proc capability data=measures;
   ppplot length / normal(mu=10 sigma=0.3 color=red);
run;
```

The MU= and SIGMA= *normal-options* specify μ and σ for the normal distribution, and the COLOR= *normal-option* specifies the color for the line.

Table 6.54 Distribution Options

Option	Description
Distribution Reference Line Options	
COLOR=	specifies color of distribution reference line
L=	specifies line type of distribution reference line
NOLINE	suppresses the distribution reference line
SYMBOL=	specifies plotting character for line printer plots
W=	specifies width of distribution reference line
Beta-Options	
ALPHA=	specifies shape parameter α
BETA=	specifies shape parameter β
SIGMA=	specifies scale parameter σ
THETA=	specifies lower threshold parameter θ
Exponential-Options	
SIGMA=	specifies scale parameter σ
THETA=	specifies threshold parameter θ
Gamma-Options	
ALPHA=	specifies shape parameter α
SIGMA=	specifies scale parameter σ
THETA=	specifies threshold parameter θ

Table 6.54 (*continued*)

Option	Description
Gumbel-Options	
MU=	specifies location parameter μ
SIGMA=	specifies scale parameter σ
IGauss-Options	
LAMBDA=	specifies shape parameter λ
MU=	specifies mean μ
Lognormal-Options	
SIGMA=	specifies shape parameter σ
THETA=	specifies threshold parameter θ
ZETA=	specifies scale parameter ζ
Normal-Options	
MU=	specifies mean μ
SIGMA=	specifies standard deviation σ
Pareto-Options	
ALPHA=	specifies shape parameter α
SIGMA=	specifies scale parameter σ
THETA=	specifies threshold parameter θ
Power-Options	
ALPHA=	specifies shape parameter α
SIGMA=	specifies scale parameter σ
THETA=	specifies threshold parameter θ
Rayleigh-Options	
SIGMA=	specifies scale parameter σ
THETA=	specifies threshold parameter θ
Weibull-Options	
C=	specifies shape parameter c
SIGMA=	specifies scale parameter σ
THETA=	specifies threshold parameter θ

General Options

Table 6.55 lists options that control the appearance of the plots.

Table 6.55 General PPPLOT Statement Options

Option	Description
General Plot Layout Options	
CONTENTS=	specifies table of contents entry for P-P plot grouping
HREF=	specifies reference lines perpendicular to the horizontal axis
HREFLABELS=	specifies line labels for HREF= lines
NOFRAME	suppresses frame around plotting area
SQUARE	displays P-P plot in square format
VREF=	specifies reference lines perpendicular to the vertical axis

Table 6.55 (*continued*)

Option	Description
VREFLABELS=	specifies line labels for VREF= lines
Graphics Options	
ANNOTATE=	provides an annotate data set
CAXIS=	specifies color for axis
CFRAME=	specifies color for frame
CHREF=	specifies colors for HREF= lines
CTEXT=	specifies color for text
CVREF=	specifies colors for VREF= lines
DESCRIPTION=	specifies description for plot in graphics catalog
FONT=	specifies software font for text
HAXIS=	specifies AXIS statement for horizontal axis
HEIGHT=	specifies height of text used outside framed areas
HMINOR=	specifies number of minor tick marks on horizontal axis
HREFLABPOS=	specifies position for HREF= line labels
INFONT=	specifies software font for text inside framed areas
INHEIGHT=	specifies height of text inside framed areas
LHREF=	specifies line styles for HREF= lines
LVREF=	specifies line styles for VREF= lines
NAME=	specifies name for plot in graphics catalog
NOHLABEL	suppresses label for horizontal axis
NOVLABEL	suppresses label for vertical axis
NOVTICK	suppresses tick marks and tick mark labels for vertical axis
TURNVLABELS	turns and vertically strings out characters in labels for vertical axis
VAXIS=	specifies AXIS statement for vertical axis
VAXISLABEL=	specifies label for vertical axis
VMINOR=	specifies number of minor tick marks on vertical axis
VREFLABPOS=	specifies position for VREF= line labels
WAXIS=	specifies line thickness for axes and frame
Options for ODS Graphics Output	
ODSFOOTNOTE=	specifies footnote displayed on P-P plot
ODSFOOTNOTE2=	specifies secondary footnote displayed on P-P plot
ODSTITLE=	specifies title displayed on P-P plot
ODSTITLE2=	specifies secondary title displayed on P-P plot
Options for Comparative Plots	
ANNOKEY	applies annotation requested in ANNOTATE= data set to key cell only
CFRAMESIDE=	specifies color for filling row label frames
CFRAMETOP=	specifies color for filling column label frames
CPROP=	specifies color for proportion of frequency bar
CTEXTSIDE=	specifies color for row labels
CTEXTTOP=	specifies color for column labels
INTERTILE=	specifies distance between tiles in comparative plot
NCOLS=	specifies number of columns in comparative plot

Table 6.55 (*continued*)

Option	Description
NROWS=	specifies number of rows in comparative plot
OVERLAY	overlays plots for different class levels (ODS Graphics only)
Options for Line Printer Charts	
HREFCHAR=	specifies line character for HREF= lines
NOOBSLEGEND	suppresses legend for hidden points
PPSYMBOL=	specifies character for plotted points
VREFCHAR=	specifies line character for VREF= lines

Dictionary of Options

The following entries provide detailed descriptions of the options specific to the PPPLOT statement. See "Dictionary of Common Options: CAPABILITY Procedure" on page 533 for detailed descriptions of options common to all the plot statements.

ALPHA=*value*

specifies the shape parameter α ($\alpha > 0$) for P-P plots requested with the BETA, GAMMA, PARETO, and POWER options. For examples, see the entries for the distribution options.

BETA<(*beta-options***)>**

creates a beta P-P plot. To create the plot, the n nonmissing observations are ordered from smallest to largest:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

The y-coordinate of the ith point is the empirical cdf value $\frac{i}{n}$. The x-coordinate is the theoretical beta cdf value

$$B_{\alpha\beta}\left(\frac{x_{(i)}-\theta}{\sigma}\right) = \int_{\theta}^{x_{(i)}} \frac{(t-\theta)^{\alpha-1}(\theta+\sigma-t)^{\beta-1}}{B(\alpha,\beta)\sigma^{(\alpha+\beta-1)}}dt$$

where $B_{\alpha\beta}(\cdot)$ is the normalized incomplete beta function, $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, and

θ = lower threshold parameter
σ = scale parameter ($\sigma > 0$)
α = first shape parameter ($\alpha > 0$)
β = second shape parameter ($\beta > 0$)

You can specify α, β, σ, and θ with the ALPHA=, BETA=, SIGMA=, and THETA= *beta-options*, as illustrated in the following example:

```
proc capability data=measures;
   ppplot width / beta(theta=1 sigma=2 alpha=3 beta=4);
run;
```

If you do not specify values for these parameters, then by default, $\theta = 0$, $\sigma = 1$, and maximum likelihood estimates are calculated for α and β.

IMPORTANT: If the default unit interval (0,1) does not adequately describe the range of your data, then you should specify THETA=θ and SIGMA=σ so that your data fall in the interval $(\theta, \theta + \sigma)$.

If the data are beta distributed with parameters α, β, σ, and θ, then the points on the plot for ALPHA=α, BETA=β, SIGMA=σ, and THETA=θ tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified beta distribution is a good fit. You can specify the SCALE= option as an alias for the SIGMA= option and the THRESHOLD= option as an alias for the THETA= option.

BETA=*value*

specifies the shape parameter β ($\beta > 0$) for P-P plots requested with the BETA distribution option. See the preceding entry for the BETA distribution option for an example.

C=*value*

specifies the shape parameter c ($c > 0$) for P-P plots requested with the WEIBULL option. See the entry for the WEIBULL option for examples.

EXPONENTIAL< (*exponential-options***) >**

EXP< (*exponential-options***) >**

creates an exponential P-P plot. To create the plot, the n nonmissing observations are ordered from smallest to largest:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

The y-coordinate of the ith point is the empirical cdf value $\frac{i}{n}$. The x-coordinate is the theoretical exponential cdf value

$$F(x_{(i)}) = 1 - \exp\left(-\frac{x_{(i)} - \theta}{\sigma}\right)$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)

You can specify σ and θ with the SIGMA= and THETA= *exponential-options*, as illustrated in the following example:

```
proc capability data=measures;
   ppplot width / exponential(theta=1 sigma=2);
run;
```

If you do not specify values for these parameters, then by default, $\theta = 0$ and a maximum likelihood estimate is calculated for σ.

IMPORTANT: Your data must be greater than or equal to the lower threshold θ. If the default $\theta = 0$ is not an adequate lower bound for your data, specify θ with the THETA= option.

If the data are exponentially distributed with parameters σ and θ, the points on the plot for SIGMA=σ and THETA=θ tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified exponential distribution is a good fit. You can specify the SCALE= option as an alias for the SIGMA= option and the THRESHOLD= option as an alias for the THETA= option.

GAMMA< (*gamma-options*) >

creates a gamma P-P plot. To create the plot, the n nonmissing observations are ordered from smallest to largest:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

The y-coordinate of the ith point is the empirical cdf value $\frac{i}{n}$. The x-coordinate is the theoretical gamma cdf value

$$G_\alpha\left(\frac{x_{(i)}-\theta}{\sigma}\right) = \int_\theta^{x_{(i)}} \frac{1}{\sigma\Gamma(\alpha)}\left(\frac{t-\theta}{\sigma}\right)^{\alpha-1} \exp\left(-\frac{t-\theta}{\sigma}\right) dt$$

where $G_\alpha(\cdot)$ is the normalized incomplete gamma function, and

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)
α = shape parameter ($\alpha > 0$)

You can specify α, σ, and θ with the ALPHA=, SIGMA=, and THETA= *gamma-options*, as illustrated in the following example:

```
proc capability data=measures;
   ppplot width / gamma(alpha=1 sigma=2 theta=3);
run;
```

If you do not specify values for these parameters, then by default, $\theta = 0$ and maximum likelihood estimates are calculated for α and σ.

IMPORTANT: Your data must be greater than or equal to the lower threshold θ. If the default $\theta = 0$ is not an adequate lower bound for your data, specify θ with the THETA= option.

If the data are gamma distributed with parameters α, σ, and θ, the points on the plot for ALPHA=α, SIGMA=σ, and THETA=θ tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified gamma distribution is a good fit. You can specify the SHAPE= option as an alias for the ALPHA= option, the SCALE= option as an alias for the SIGMA= option, and the THRESHOLD= option as an alias for the THETA= option.

GUMBEL< (*Gumbel-options*) >

creates a Gumbel P-P plot. To create the plot, the n nonmissing observations are ordered from smallest to largest:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

The y-coordinate of the ith point is the empirical cdf value $\frac{i}{n}$. The x-coordinate is the theoretical Gumbel cdf value

$$F(x_{(i)}) = \exp\left(-e^{-(x_{(i)}-\mu)/\sigma}\right)$$

where

μ = location parameter
σ = scale parameter ($\sigma > 0$)

You can specify μ and σ with the MU= and SIGMA= *Gumbel-options*. By default, maximum likelihood estimates are computed for μ and σ.

If the data are Gumbel distributed with parameters μ and σ, the points on the plot for MU=μ and SIGMA=σ tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified Gumbel distribution is a good fit.

IGAUSS< (*iGauss-options*) >

creates an inverse Gaussian P-P plot. To create the plot, the n nonmissing observations are ordered from smallest to largest:

$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$

The y-coordinate of the ith point is the empirical cdf value $\frac{i}{n}$. The x-coordinate is the theoretical inverse Gaussian cdf value

$$F(x_{(i)}) = \Phi\left\{\sqrt{\frac{\lambda}{x_{(i)}}}\left(\frac{x_{(i)}}{\mu} - 1\right)\right\} + e^{2\lambda/\mu}\Phi\left\{-\sqrt{\frac{\lambda}{x_{(i)}}}\left(\frac{x_{(i)}}{\mu} + 1\right)\right\}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function, and

μ = mean parameter ($\mu > 0$)
λ = shape parameter ($\lambda > 0$)

You can specify known values for μ and λ with the MU= and LAMBDA= *iGauss-options*. By default, the sample mean is calculated for μ and a maximum likelihood estimate is computed for and λ.

If the data are inverse Gaussian distributed with parameters μ and λ, the points on the plot for MU=μ and LAMBDA=λ tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified inverse Gaussian distribution is a good fit.

LAMBDA=*value*

specifies the shape parameter λ ($\lambda > 0$) for P-P plots requested with the IGAUSS option. Enclose the LAMBDA= option in parentheses after the IGAUSS distribution keyword. If you do not specify a value for λ, the procedure calculates a maximum likelihood estimate.

LOGNORMAL< (*lognormal-options*) >

LNORM< (*lognormal-options*) >

creates a lognormal P-P plot. To create the plot, the n nonmissing observations are ordered from smallest to largest:

$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$

The y-coordinate of the ith point is the empirical cdf value $\frac{i}{n}$. The x-coordinate is the theoretical lognormal cdf value

$\Phi\left(\frac{\log(x_{(i)}-\theta)-\zeta}{\sigma}\right)$

where $\Phi(\cdot)$ is the cumulative standard normal distribution function, and

θ = threshold parameter

ζ = scale parameter
σ = shape parameter ($\sigma > 0$)

You can specify θ, ζ, and σ with the THETA=, ZETA=, and SIGMA= *lognormal-options*, as illustrated in the following example:

```
proc capability data=measures;
   ppplot width / lognormal(theta=1 zeta=2);
run;
```

If you do not specify values for these parameters, then by default, $\theta = 0$ and maximum likelihood estimates are calculated for σ and ζ.

IMPORTANT: Your data must be greater than the lower threshold θ. If the default $\theta = 0$ is not an adequate lower bound for your data, specify θ with the THETA= option.

If the data are lognormally distributed with parameters σ, θ, and ζ, the points on the plot for SIGMA=σ, THETA=θ, and ZETA=ζ tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified lognormal distribution is a good fit. You can specify the SHAPE= option as an alias for the SIGMA= option, the SCALE= option as an alias for the ZETA= option, and the THRESHOLD= option as an alias for the THETA= option.

MU=*value*
specifies the parameter μ for a P-P plot requested with the GUMBEL, IGAUSS, and NORMAL options. For examples, see Figure 6.31, or Figure 6.32 and Figure 6.33. For the normal and inverse Gaussian distributions, the default value of μ is the sample mean. If you do not specify a value for μ for the Gumbel distribution, the procedure calculates a maximum likelihood estimate.

NOLINE
suppresses the diagonal reference line.

NOOBSLEGEND
NOOBSL
suppresses the legend that indicates the number of hidden observations in a legacy line printer plot. This option is ignored unless you specify the LINEPRINTER option in the PROC CAPABILITY statement.

NORMAL<(*normal-options***)>**
NORM<(*normal-options* **)>**
creates a normal P-P plot. By default, if you do not specify a distribution option, the procedure displays a normal P-P plot. To create the plot, the n nonmissing observations are ordered from smallest to largest:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

The y-coordinate of the ith point is the empirical cdf value $\frac{i}{n}$. The x-coordinate is the theoretical normal cdf value

$$\Phi\left(\frac{x_{(i)}-\mu}{\sigma}\right) = \int_{-\infty}^{x_{(i)}} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt$$

where $\Phi(\cdot)$ is the cumulative standard normal distribution function, and

μ = location parameter or mean
σ = scale parameter or standard deviation ($\sigma > 0$)

You can specify μ and σ with the MU= and SIGMA= *normal-options*, as illustrated in the following example:

```
proc capability data=measures;
   ppplot width / normal(mu=1 sigma=2);
run;
```

By default, the sample mean and sample standard deviation are used for μ and σ.

If the data are normally distributed with parameters μ and σ, the points on the plot for MU=μ and SIGMA=σ tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified normal distribution is a good fit. For an example, see Figure 6.31.

PARETO< (*Pareto-options*) >

creates a generalized Pareto P-P plot. To create the plot, the n nonmissing observations are ordered from smallest to largest:

$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$

The y-coordinate of the ith point is the empirical cdf value $\frac{i}{n}$. The x-coordinate is the theoretical generalized Pareto cdf value

$$F(x_{(i)}) = 1 - \left(1 - \frac{\alpha(x_{(i)} - \theta)}{\sigma}\right)^{\frac{1}{\alpha}}$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)
α = shape parameter

The parameter θ for the generalized Pareto distribution must be less than the minimum data value. You can specify θ with the THETA= *Pareto-option*. The default value for θ is 0. In addition, the generalized Pareto distribution has a shape parameter α and a scale parameter σ. You can specify these parameters with the ALPHA= and SIGMA= *Pareto-options*. By default, maximum likelihood estimates are computed for α and σ.

If the data are generalized Pareto distributed with parameters θ, σ, and α, the points on the plot for THETA=θ, SIGMA=σ, and ALPHA=α tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified generalized Pareto distribution is a good fit.

POWER< (*power-options*) >

creates a power function P-P plot. To create the plot, the n nonmissing observations are ordered from smallest to largest:

$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$

The y-coordinate of the ith point is the empirical cdf value $\frac{i}{n}$. The x-coordinate is the theoretical power function cdf value

$$F(x_{(i)}) = \left(\frac{x_{(i)} - \theta}{\sigma}\right)^{\alpha}$$

where

θ = lower threshold parameter (lower endpoint)
σ = scale parameter ($\sigma > 0$)
α = shape parameter ($\alpha > 0$)

The power function distribution is bounded below by the parameter θ and above by the value $\theta + \sigma$. You can specify θ and σ by using the THETA= and SIGMA= *power-options*. The default values for θ and σ are 0 and 1, respectively.

You can specify a value for the shape parameter, α, with the ALPHA= *power-option*. If you do not specify a value for α, the procedure calculates a maximum likelihood estimate.

The power function distribution is a special case of the beta distribution with its second shape parameter, $\beta = 1$.

If the data are power function distributed with parameters θ, σ, and α, the points on the plot for THETA=θ, SIGMA=σ, and ALPHA=α tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified power function distribution is a good fit.

PPSYMBOL='*character*'
specifies the character used to plot the points in a legacy line printer plot. The default is the plus sign (+). This option is ignored unless you specify the LINEPRINTER option in the PROC CAPABILITY statement.

RAYLEIGH<(*Rayleigh-options***)>**
creates a Rayleigh P-P plot. To create the plot, the n nonmissing observations are ordered from smallest to largest:

$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$

The y-coordinate of the ith point is the empirical cdf value $\frac{i}{n}$. The x-coordinate is the theoretical Rayleigh cdf value

$$F(x_{(i)}) = 1 - e^{-(x_{(i)}-\theta)^2/(2\sigma^2)}$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)

The parameter θ for the Rayleigh distribution must be less than the minimum data value. You can specify θ with the THETA= *Rayleigh-option*. The default value for θ is 0. You can specify σ with the SIGMA= *Rayleigh-option*. By default, a maximum likelihood estimate is computed for σ.

If the data are Rayleigh distributed with parameters θ and σ, the points on the plot for THETA=θ and SIGMA=σ tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified Rayleigh distribution is a good fit.

SIGMA=*value*

specifies the parameter σ, where $\sigma > 0$. When used with the BETA, EXPONENTIAL, GAMMA, GUMBEL, NORMAL, PARETO, POWER, RAYLEIGH, and WEIBULL options, the SIGMA= option specifies the scale parameter. When used with the LOGNORMAL option, the SIGMA= option specifies the shape parameter. Enclose the SIGMA= option in parentheses after the distribution keyword. For an example of the SIGMA= option used with the NORMAL option, see Figure 6.31.

SQUARE

displays the P-P plot in a square frame. The default is a rectangular frame. See Figure 6.31 for an example.

SYMBOL='*character*'

specifies the character used for the diagonal reference line in legacy line printer plots. The default character is the first letter of the distribution option keyword. This option is ignored unless you specify the LINEPRINTER option in the PROC CAPABILITY statement.

THETA=*value*

THRESHOLD=*value*

specifies the lower threshold parameter θ for plots requested with the BETA, EXPONENTIAL, GAMMA, LOGNORMAL, PARETO, POWER, RAYLEIGH, and WEIBULL options.

WEIBULL< (*Weibull-options***) >**

WEIB< (*Weibull-options***) >**

creates a Weibull P-P plot. To create the plot, the n nonmissing observations are ordered from smallest to largest:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

The y-coordinate of the ith point is the empirical cdf value $\frac{i}{n}$. The x-coordinate is the theoretical Weibull cdf value

$$F(x_{(i)}) = 1 - \exp\left(-\left(\frac{x_{(i)}-\theta}{\sigma}\right)^c\right)$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)
c = shape parameter ($c > 0$)

You can specify c, σ, and θ with the C=, SIGMA=, and THETA= *Weibull-options*, as illustrated in the following example:

```
proc capability data=measures;
   ppplot width / weibull(theta=1 sigma=2);
run;
```

If you do not specify values for these parameters, then by default $\theta = 0$ and maximum likelihood estimates are calculated for σ and c.

IMPORTANT: Your data must be greater than or equal to the lower threshold θ. If the default $\theta = 0$ is not an adequate lower bound for your data, you should specify θ with the THETA= option.

If the data are Weibull distributed with parameters c, σ, and θ, the points on the plot for C=c, SIGMA=σ, and THETA=θ tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified Weibull distribution is a good fit. You can specify the SHAPE= option as an alias for the C= option, the SCALE= option as an alias for the SIGMA= option, and the THRESHOLD= option as an alias for the THETA= option.

ZETA=*value*

specifies a value for the scale parameter ζ for lognormal P-P plots requested with the LOGNORMAL option.

Details: PPPLOT Statement

This section provides details on the following topics:

- construction and interpretation of P-P plots
- comparison of P-P plots with Q-Q plots
- distributions supported by the PPPLOT statement
- graphical enhancements of P-P plots

Construction and Interpretation of P-P Plots

A P-P plot compares the empirical cumulative distribution function (ecdf) of a variable with a specified theoretical cumulative distribution function $F(\cdot)$. The ecdf, denoted by $F_n(x)$, is defined as the proportion of nonmissing observations less than or equal to x, so that $F_n(x_{(i)}) = \frac{i}{n}$.

To construct a P-P plot, the n nonmissing values are first sorted in increasing order:

$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$

Then the ith ordered value $x_{(i)}$ is represented on the plot by the point whose x-coordinate is $F(x_{(i)})$ and whose y-coordinate is $\frac{i}{n}$.

Like Q-Q plots and probability plots, P-P plots can be used to determine how well a theoretical distribution models a data distribution. If the theoretical cdf reasonably models the ecdf in all respects, including location and scale, the point pattern on the P-P plot is linear through the origin and has unit slope.

NOTE: See *Interpreting P-P Plots* in the SAS/QC Sample Library.

Unlike Q-Q and probability plots, P-P plots are not invariant to changes in location and scale. For example, the data in the section "Getting Started: PPPLOT Statement" on page 439 are reasonably described by a normal distribution with mean 10 and standard deviation 0.3. It is instructive to display these data on normal P-P plots with a different mean and standard deviation, as created by the following statements:

```
data Sheets;
   input Distance @@;
   label Distance='Hole Distance in cm';
   datalines;
 9.80 10.20 10.27  9.70  9.76
10.11 10.24 10.20 10.24  9.63
 9.99  9.78 10.10 10.21 10.00
 9.96  9.79 10.08  9.79 10.06
10.10  9.95  9.84 10.11  9.93
10.56 10.47  9.42 10.44 10.16
10.11 10.36  9.94  9.77  9.36
 9.89  9.62 10.05  9.72  9.82
 9.99 10.16 10.58 10.70  9.54
10.31 10.07 10.33  9.98 10.15
;


proc capability data=Sheets noprint;
   ppplot Distance / normal(mu=9.5 sigma=0.3) square;
   ppplot Distance / normal(mu=10 sigma=0.5) square;
run;
```

The ODS GRAPHICS ON statement specified before the PROC CAPABILITY statement enables ODS Graphics, so the P-P plots are created using ODS Graphics instead of traditional graphics. The resulting plots are shown in Figure 6.32 and Figure 6.33.

Figure 6.32 Normal P-P Plot with Mean Specified Incorrectly

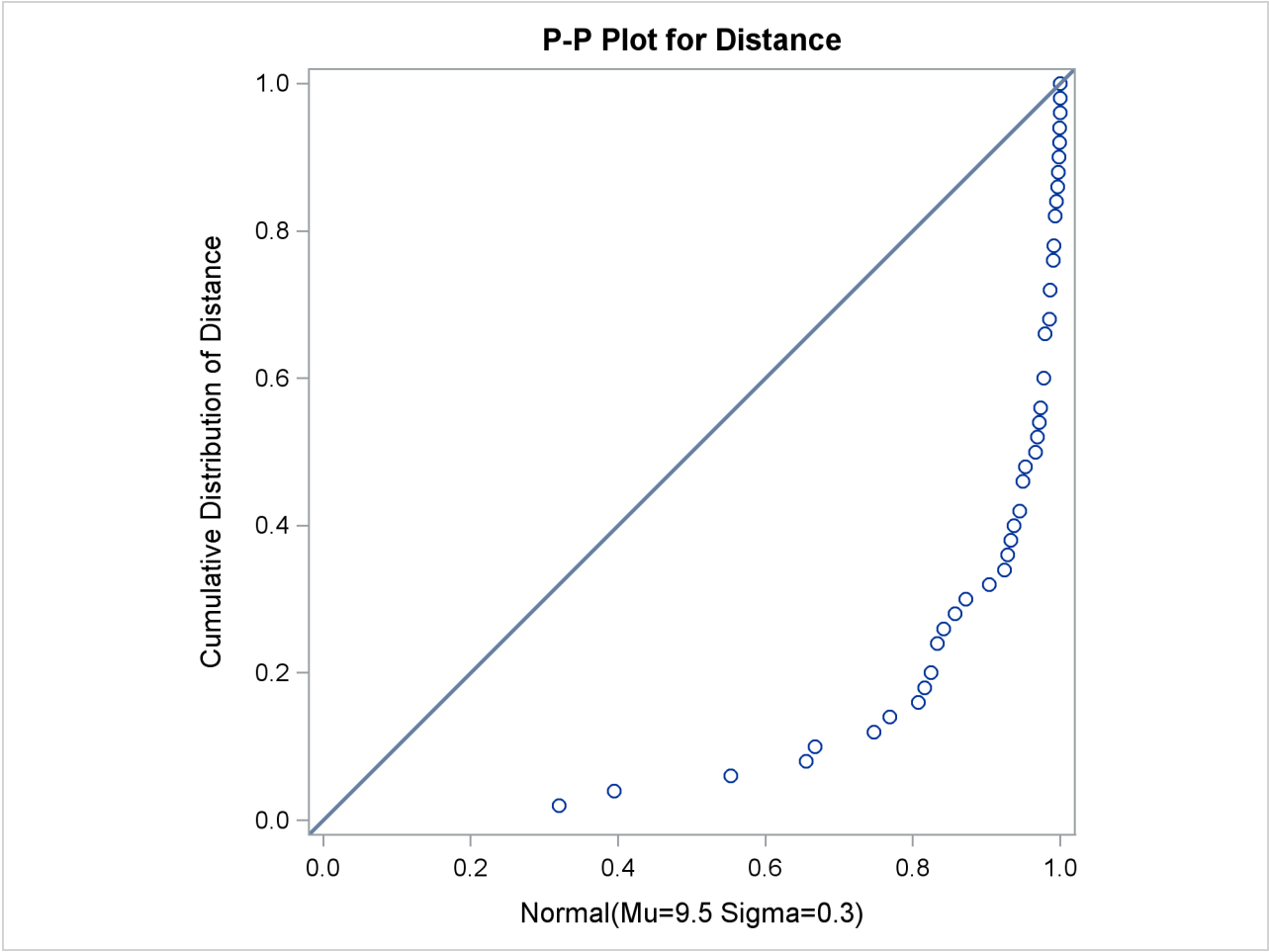
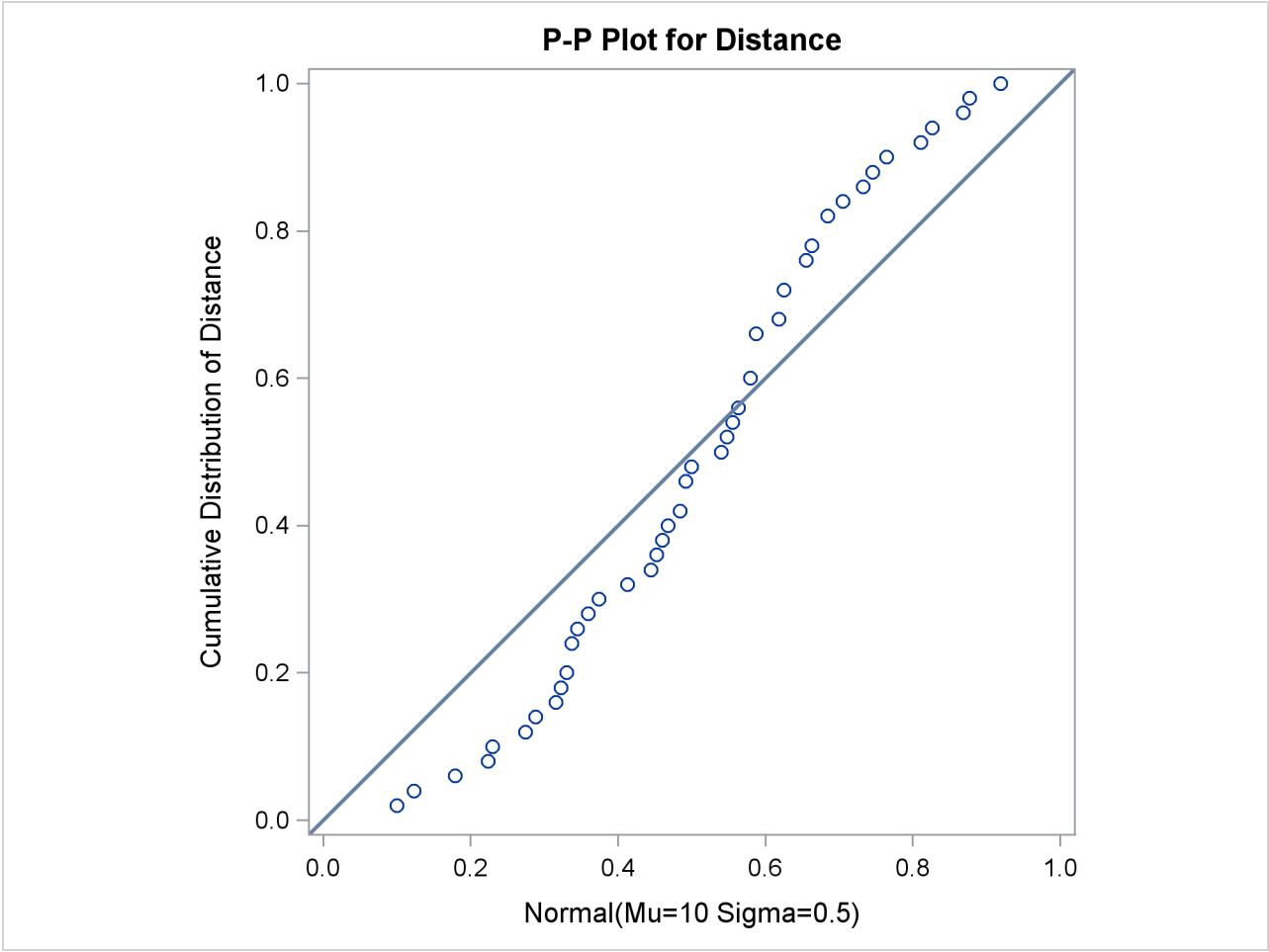


Figure 6.33 Normal P-P Plot with Standard Deviation Specified Incorrectly



Specifying a mean of 9.5 instead of 10 results in the plot shown in Figure 6.32, while specifying a standard deviation of 0.5 instead of 0.3 results in the plot shown in Figure 6.33. Both plots clearly reveal the model misspecification.

Comparison of P-P Plots and Q-Q Plots

A P-P plot compares the empirical cumulative distribution function of a data set with a specified theoretical cumulative distribution function $F(\cdot)$. A Q-Q plot compares the quantiles of a data distribution with the quantiles of a standardized theoretical distribution from a specified family of distributions. There are three important differences in the way P-P plots and Q-Q plots are constructed and interpreted:

- The construction of a Q-Q plot does not require that the location or scale parameters of $F(\cdot)$ be specified. The theoretical quantiles are computed from a standard distribution within the specified family. A linear point pattern indicates that the specified family reasonably describes the data distribution, and the location and scale parameters can be estimated visually as the intercept and slope of the linear pattern. In contrast, the construction of a P-P plot requires the location and scale parameters of $F(\cdot)$ to evaluate the cdf at the ordered data values.

- The linearity of the point pattern on a Q-Q plot is unaffected by changes in location or scale. On a P-P plot, changes in location or scale do not necessarily preserve linearity.
- On a Q-Q plot, the reference line representing a particular theoretical distribution depends on the location and scale parameters of that distribution, having intercept and slope equal to the location and scale parameters. On a P-P plot, the reference line for any distribution is always the diagonal line $y = x$.

Consequently, you should use a Q-Q plot if your objective is to compare the data distribution with a family of distributions that vary only in location and scale, particularly if you want to estimate the location and scale parameters from the plot.

An advantage of P-P plots is that they are discriminating in regions of high probability density, because in these regions the empirical and theoretical cumulative distributions change more rapidly than in regions of low probability density. For example, if you compare a data distribution with a particular normal distribution, differences in the middle of the two distributions are more apparent on a P-P plot than on a Q-Q plot.

For further details on P-P plots, refer to Gnanadesikan (1997) and Wilk and Gnanadesikan (1968).

Summary of Theoretical Distributions

You can use the PPPLOT statement to request P-P plots based on the theoretical distributions summarized in the following table:

Table 6.56 Distributions and Parameters

Family	Distribution Function $F(x)$	Range	Parameters		
			Location	Scale	Shape
Beta	$\int_{\theta}^{x} \frac{(t-\theta)^{\alpha-1}(\theta+\sigma-t)^{\beta-1}}{B(\alpha,\beta)\sigma^{(\alpha+\beta-1)}} dt$	$\theta < x < \theta+\sigma$	θ	σ	α, β
Exponential	$1 - \exp\left(-\frac{x-\theta}{\sigma}\right)$	$x \geq \theta$	θ	σ	
Gamma	$\int_{\theta}^{x} \frac{1}{\sigma\Gamma(\alpha)} \left(\frac{t-\theta}{\sigma}\right)^{\alpha-1} \exp\left(-\frac{t-\theta}{\sigma}\right) dt$	$x > \theta$	θ	σ	α
Gumbel	$\exp\left(-e^{(x-\mu)/\sigma}\right)$	all x	μ	σ	
Inverse Gaussian	$\Phi\left\{\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\mu}-1\right)\right\} + e^{2\lambda/\mu}\Phi\left\{-\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\mu}+1\right)\right\}$	$x > 0$	μ		λ
Lognormal	$\int_{\theta}^{x} \frac{1}{\sigma\sqrt{2\pi}(t-\theta)} \exp\left(-\frac{(\log(t-\theta)-\zeta)^2}{2\sigma^2}\right) dt$	$x > \theta$	θ	ζ	σ
Normal	$\int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt$	all x	μ	σ	
Generalized Pareto	$1 - \left(1 - \frac{\alpha(x-\theta)}{\sigma}\right)^{1/\alpha}$	all x	θ	σ	α
Power Function	$\left(\frac{x-\theta}{\sigma}\right)^{\alpha}$	$\theta < x < \theta+\sigma$	θ	σ	α

Table 6.56 (*continued*)

Family	Distribution Function $F(x)$	Range	Parameters		
			Location	Scale	Shape
Rayleigh	$1 - e^{-(x-\theta)^2/(2\sigma^2)}$	$x \geq \theta$	θ	σ	
Weibull	$1 - \exp\left(-\left(\frac{x-\theta}{\sigma}\right)^c\right)$	$x > \theta$	θ	σ	c

You can request these distributions with the BETA, EXPONENTIAL, GAMMA, GUMBEL, IGAUSS, NORMAL, LOGNORMAL, PARETO, POWER, RAYLEIGH, and WEIBULL options, respectively. If you do not specify a distribution option, a normal P-P plot is created.

To create a P-P plot, you must provide all of the parameters for the theoretical distribution. If you do not specify parameters, then default values or estimates are substituted, as summarized by the following table:

Table 6.57 Defaults for Parameters

Family	Default Values	Estimated Values
Beta	$\theta = 0, \sigma = 1$	maximum likelihood estimates for α and β
Exponential	$\theta = 0$	maximum likelihood estimate for σ
Gamma	$\theta = 0$	maximum likelihood estimates for σ and α
Gumbel	None	maximum likelihood estimates for μ and σ
Inverse Gaussian	None	sample estimate for μ, maximum likelihood estimate for λ
Lognormal	$\theta = 0$	maximum likelihood estimates for σ and ζ
Normal	None	sample estimates for μ and σ
Generalized Pareto	$\theta = 0$	maximum likelihood estimates for σ and α
Power Function	$\theta = 0, \sigma = 1$	maximum likelihood estimate for α
Rayleigh	$\theta = 0$	maximum likelihood estimate for σ
Weibull	$\theta = 0$	maximum likelihood estimates for σ and c

Specification of Symbol Markers

If you produce traditional graphics, you can use options in the SYMBOL1 statement to specify the appearance of the symbol marker for the points. The V= option specifies the symbol, the C= option specifies the color, and the H= option specifies the height. Refer to *SAS/GRAPH: Help* for details concerning these options. If you produce a line printer plot, you can use the PPSYMBOL= option in the PPPLOT statement to specify the character used to plot the points.

Specification of the Distribution Reference Line

If you produce traditional graphics, you can control the color, type, and width of the diagonal distribution reference line by specifying the COLOR=, L=, and W= options in parentheses after the distribution option in the PPPLOT statement. Alternatively, you can control these features with the C=, L=, and W= options in the SYMBOL4 statement. Refer to *SAS/GRAPH: Help* for details concerning these options. If you produce

a line printer plot, you can specify the character used for the line with the SYMBOL= option enclosed in parentheses after the distribution option in the PPPLOT statement.

ODS Graphics

Before you create ODS Graphics output, ODS Graphics must be enabled (for example, by using the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" (Chapter 21, *SAS/STAT User's Guide*).

The appearance of a graph produced with ODS Graphics is determined by the style associated with the ODS destination where the graph is produced. PPPLOT options used to control the appearance of traditional graphics are ignored for ODS Graphics output.

When ODS Graphics is in effect, the PPPLOT statement assigns a name to the graph it creates. You can use this name to reference the graph when using ODS. The name is listed in Table 6.58.

Table 6.58 ODS Graphics Produced by the PPPLOT Statement

ODS Graph Name	Plot Description
PPPlot	P-P plot

See Chapter 4, "SAS/QC Graphics," for more information about ODS Graphics and other methods for producing charts.

PROBPLOT Statement: CAPABILITY Procedure

Overview: PROBPLOT Statement

The PROBPLOT statement creates a probability plot, which compares ordered values of a variable with percentiles of a specified theoretical distribution such as the normal. If the data distribution matches the theoretical distribution, the points on the plot form a linear pattern. Thus, you can use a probability plot to determine how well a theoretical distribution models a set of measurements.

You can specify one of the following theoretical distributions with the PROBPLOT statement:

- beta
- exponential
- gamma
- Gumbel
- three-parameter lognormal
- normal

- generalized Pareto
- power function
- Rayleigh
- two-parameter Weibull
- three-parameter Weibull

You can use options in the PROBPLOT statement to do the following:

- specify or estimate shape parameters for the theoretical distribution
- display a reference line corresponding to specified or estimated location and scale parameters for the theoretical distribution
- request graphical enhancements

You can also create a comparative probability plot by using the PROBPLOT statement in conjunction with a CLASS statement.

You have three alternatives for producing probability plots the PROBPLOT statement:

- ODS Graphics output is produced if ODS Graphics is enabled, for example by specifying the ODS GRAPHICS ON statement prior to the PROC statement.
- Otherwise, traditional graphics are produced by default if SAS/GRAPH is licensed.
- Legacy line printer charts are produced when you specify the LINEPRINTER option in the PROC statement.

See Chapter 4, "SAS/QC Graphics," for more information about producing these different kinds of graphs.

NOTE: Probability plots are similar to Q-Q plots, which you can create with the QQPLOT statement (see "QQPLOT Statement: CAPABILITY Procedure" on page 492). Probability plots are preferable for graphical estimation of percentiles, whereas Q-Q plots are preferable for graphical estimation of distribution parameters and capability indices.

Getting Started: PROBPLOT Statement

The following examples illustrate the basic syntax of the PROBPLOT statement. For complete details of the PROBPLOT statement, see the section "Syntax: PROBPLOT Statement" on page 467. Advanced examples are provided on the section "Examples: PROBPLOT Statement" on page 489.

Creating a Normal Probability Plot

NOTE: See *Creating a Normal Probability Plot* in the SAS/QC Sample Library.

The diameters of 50 steel rods are measured and saved as values of the variable Diameter in the following data set:[4]

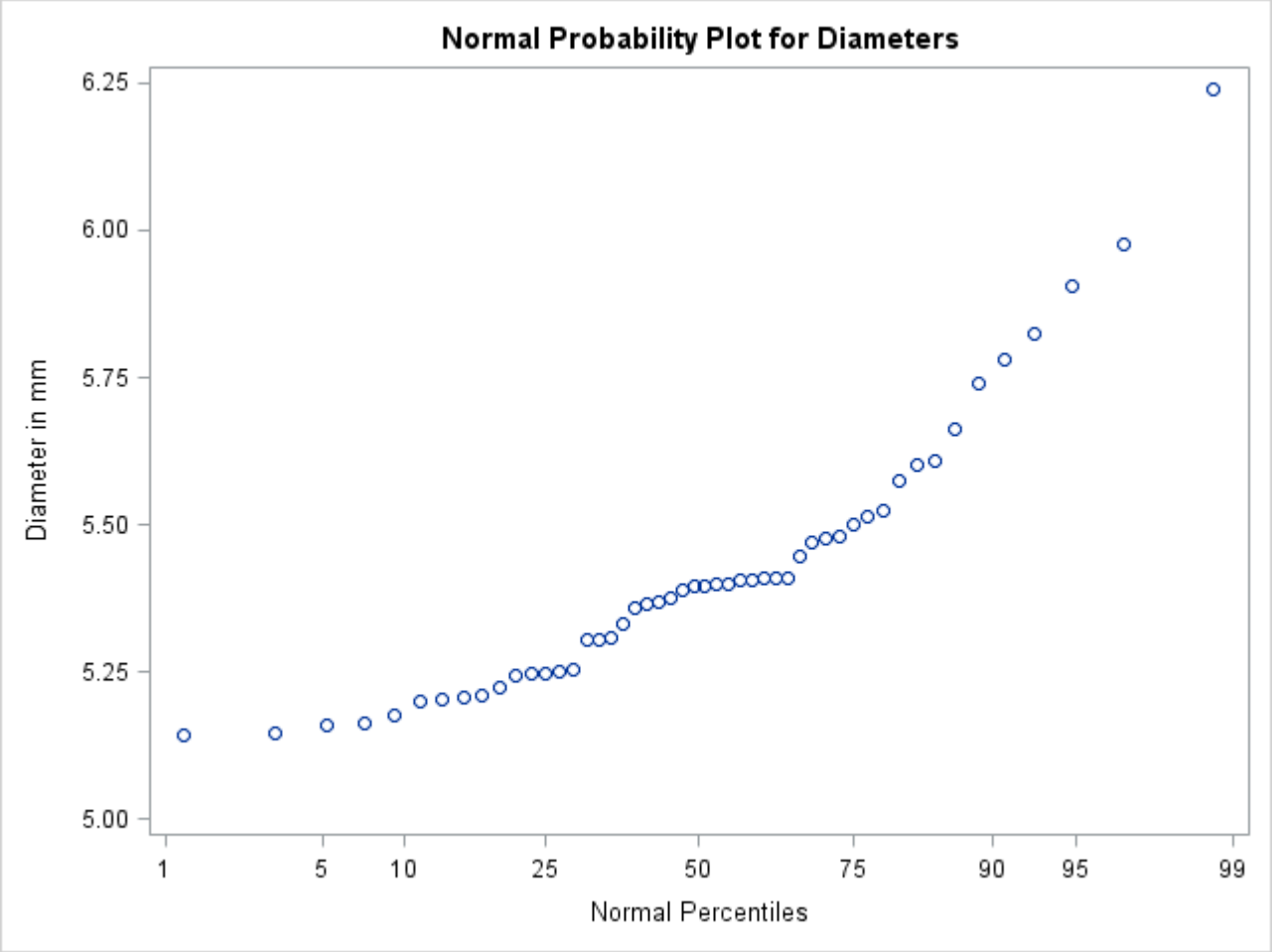
```
data Rods;
   input Diameter @@;
   label Diameter='Diameter in mm';
   datalines;
5.501  5.251  5.404  5.366  5.445
5.576  5.607  5.200  5.977  5.177
5.332  5.399  5.661  5.512  5.252
5.404  5.739  5.525  5.160  5.410
5.823  5.376  5.202  5.470  5.410
5.394  5.146  5.244  5.309  5.480
5.388  5.399  5.360  5.368  5.394
5.248  5.409  5.304  6.239  5.781
5.247  5.907  5.208  5.143  5.304
5.603  5.164  5.209  5.475  5.223
;
```

The process producing the rods is in statistical control, and as a preliminary step in a capability analysis of the process, you decide to check whether the diameters are normally distributed. The following statements create the normal probability plot shown in Figure 6.34:

```
title 'Normal Probability Plot for Diameters';
proc capability data=Rods noprint;
   probplot Diameter / odstitle=title;
run;
```

[4] This data set is analyzed using quantile-quantile plots in Example 6.21 and Example 6.22.

Figure 6.34 Normal Probability Plot Created with Traditional Graphics



Note that the PROBPLOT statement creates a normal probability plot for Diameter by default.

The nonlinearity of the point pattern indicates a departure from normality. Because the point pattern is curved with slope increasing from left to right, a theoretical distribution that is skewed to the right, such as a lognormal distribution, should provide a better fit than the normal distribution. This possibility is explored in the next example.

Creating Lognormal Probability Plots

NOTE: See *Creating Lognormal Probability Plots* in the SAS/QC Sample Library.

When you request a lognormal probability plot, you must specify the shape parameter σ for the lognormal distribution (see Table 6.62 for the equation). The value of σ must be positive, and typical values of σ range from 0.1 to 1.0. Alternatively, you can specify that σ is to be estimated from the data.

The following statements illustrate the first approach by creating a series of three lognormal probability plots for the variable Diameter introduced in the preceding example:

```
proc capability data=Rods noprint;
   probplot Diameter / lognormal(sigma=0.2 0.5 0.8)
                       href = 95
                       square;
run;
```

The LOGNORMAL option requests plots based on the lognormal family of distributions, and the SIGMA= option requests plots for σ equal to 0.2, 0.5, and 0.8. The SQUARE option displays the probability plot in a square format and the HREF= option requests a reference line at the 95th percentile.

The resulting plots are displayed in Figure 6.35, Figure 6.36, and Figure 6.37, respectively. The value $\sigma = 0.5$ in Figure 6.36 produces the most linear pattern.

Figure 6.35 Probability Plot Based on Lognormal Distribution with $\sigma = 0.2$

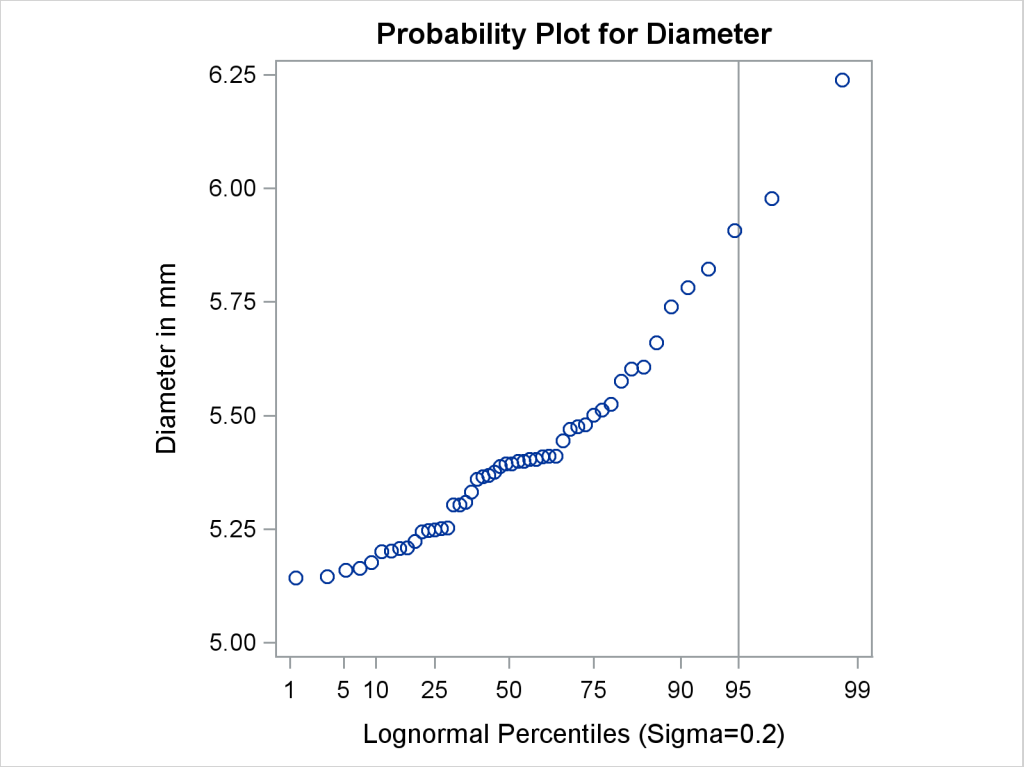


Figure 6.36 Probability Plot Based on Lognormal Distribution with $\sigma = 0.5$

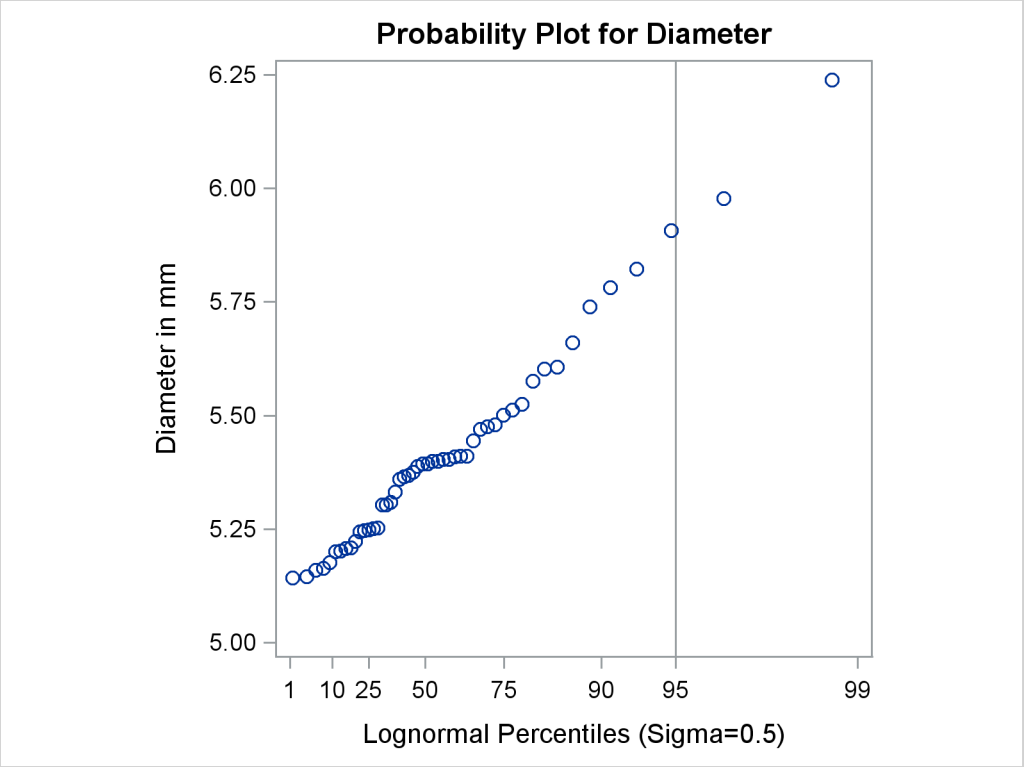
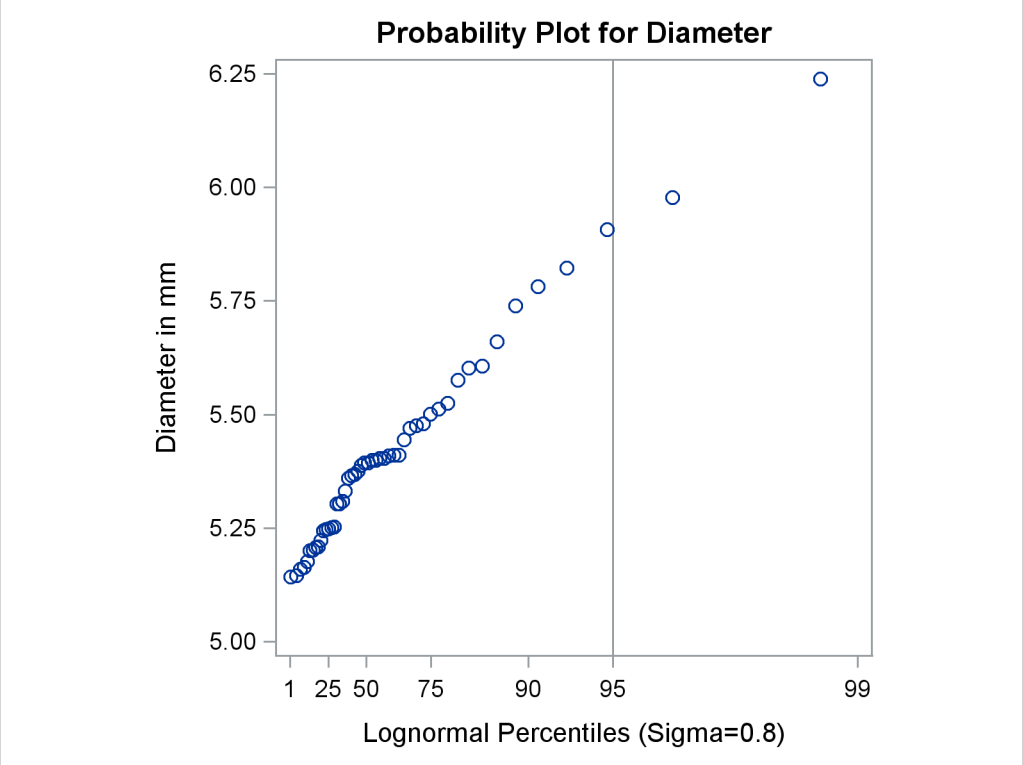


Figure 6.37 Probability Plot Based on Lognormal Distribution with $\sigma = 0.8$



Based on Figure 6.36, the 95th percentile of the diameter distribution is approximately 5.9 mm, because this is the value corresponding to the intersection of the point pattern with the reference line.

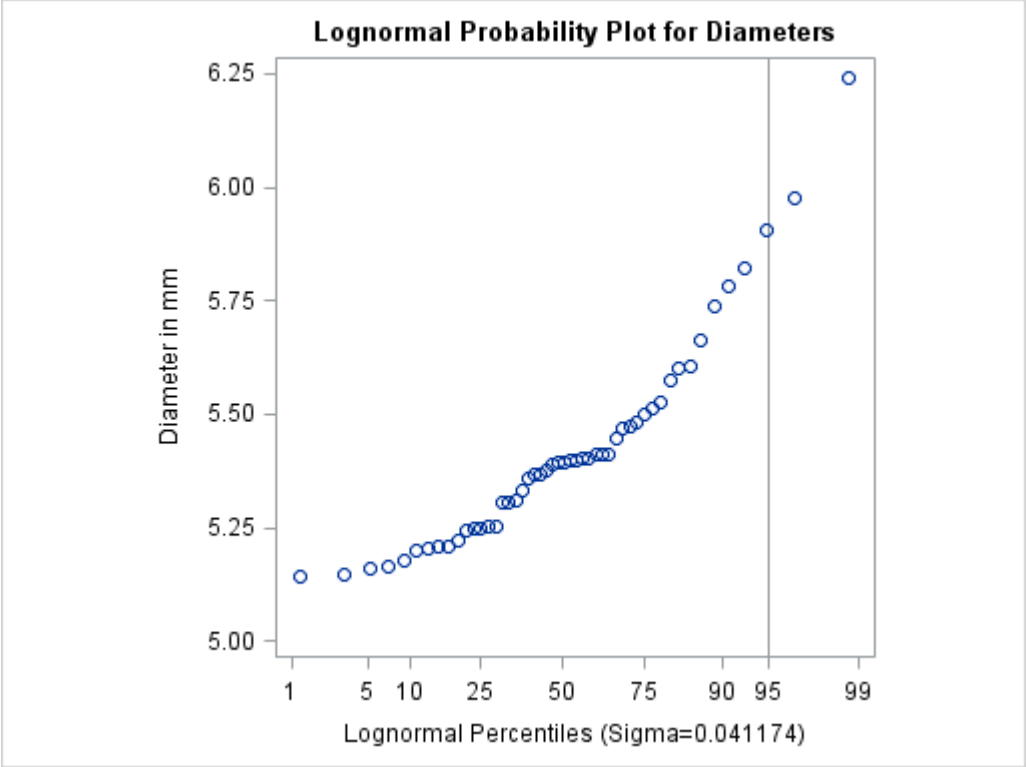
The following statements illustrate how you can create a lognormal probability plot for Diameter using a local maximum likelihood estimate for σ.

```
title 'Lognormal Probability Plot for Diameters';
proc capability data=Rods noprint;
   probplot Diameter / lognormal(sigma=est)
                       href     = 95
                       odstitle = title
                       square;
run;
```

The plot is displayed in Figure 6.38.

Note that the maximum likelihood estimate of σ (in this case 0.041) does not necessarily produce the most linear point pattern. This example is continued in Example 6.20.

Figure 6.38 Probability Plot Based on Lognormal Distribution with Estimated σ



Syntax: PROBPLOT Statement

The syntax for the PROBPLOT statement is as follows:

PROBPLOT < *variables* > < / *options* > **;**

You can specify the keyword PROB as an alias for PROBPLOT, and you can use any number of PROBPLOT statements in the CAPABILITY procedure. The components of the PROBPLOT statement are described as follows.

variables

are the process variables for which to create probability plots. If you specify a VAR statement, the variables must also be listed in the VAR statement. Otherwise, the variables can be any numeric variables in the input data set. If you do not specify a list of variables, then by default the procedure creates a probability plot for each variable listed in the VAR statement, or for each numeric variable in the DATA= data set if you do not specify a VAR statement. For example, each of the following PROBPLOT statements produces two probability plots, one for length and one for width:

```
proc capability data=measures;
   var length width;
   probplot;
run;

proc capability data=measures;
   probplot length width;
run;
```

options

specify the theoretical distribution for the plot or add features to the plot. If you specify more than one variable, the options apply equally to each variable. Specify all options after the slash (/) in the PROBPLOT statement. You can specify only one option naming the distribution in each PROBPLOT statement, but you can specify any number of other options. The distributions available are the beta, exponential, gamma, Gumbel, lognormal, normal, generalized Pareto, power function, Rayleigh, two-parameter Weibull, and three-parameter Weibull. By default, the procedure produces a plot for the normal distribution.

In the following example, the NORMAL option requests a normal probability plot for each variable, while the MU= and SIGMA= *normal-options* request a distribution reference line corresponding to the normal distribution with $\mu = 10$ and $\sigma = 0.3$. The SQUARE option displays the plot in a square frame, and the CTEXT= option specifies the text color.

```
proc capability data=measures;
   probplot length1 length2 / normal(mu=10 sigma=0.3)
                              square
                              ctext=blue;
run;
```

Summary of Options

The following tables list the PROBPLOT statement options by function. For complete descriptions, see the section "Dictionary of Options" on page 472.

Distribution Options

Table 6.59 summarizes the options for requesting a specific theoretical distribution.

Table 6.59 Options for Specifying a Theoretical Distribution

Option	Description
BETA(*beta-options*)	specifies beta probability plot for shape parameters α, β specified with mandatory ALPHA= and BETA= *beta-options*
EXPONENTIAL(*exponential-options*)	specifies exponential probability plot
GAMMA(*gamma-options*)	specifies gamma probability plot for shape parameter α specified with mandatory ALPHA= *gamma-option*
GUMBEL(*Gumbel-options*)	specifies Gumbel probability plot
LOGNORMAL(*lognormal-options*)	specifies lognormal probability plot for shape parameter σ specified with mandatory SIGMA= *lognormal-option*
NORMAL(*normal-options*)	specifies normal probability plot
PARETO(*Pareto-options*)	specifies generalized Pareto probability plot for shape parameter α specified with mandatory ALPHA= *Pareto-option*
POWER(*power-options*)	specifies power function probability plot for shape parameter α specified with mandatory ALPHA= *power-option*
RAYLEIGH(*Rayleigh-options*)	specifies Rayleigh probability plot
WEIBULL(*Weibull-options*)	specifies three-parameter Weibull probability plot for shape parameter c specified with mandatory C= *Weibull-option*
WEIBULL2(*Weibull2-options*)	specifies two-parameter Weibull probability plot

Table 6.60 summarizes options that specify distribution parameters and control the display of a distribution reference line. Specify these options in parentheses after the distribution option. For example, the following statements use the NORMAL option to request a normal probability plot with a distribution reference line:

```
proc capability data=measures;
   probplot length / normal(mu=10 sigma=0.3 color=red);
run;
```

The MU= and SIGMA= *normal-options* display a distribution reference line that corresponds to the normal distribution with mean $\mu_0 = 10$ and standard deviation $\sigma_0 = 0.3$, and the COLOR= *normal-option* specifies the color for the line.

Table 6.60 Distribution Options

Option	Description
Distribution Reference Line Options	
COLOR=	specifies color of distribution reference line
L=	specifies line type of distribution reference line
SYMBOL=	specifies plotting character for line printer plots
W=	specifies width of distribution reference line
Beta-Options	
ALPHA=	specifies mandatory shape parameter α
BETA=	specifies mandatory shape parameter β
SIGMA=	specifies σ_0 for distribution reference line
THETA=	specifies θ_0 for distribution reference line
Exponential-Options	
SIGMA=	specifies σ_0 for distribution reference line
THETA=	specifies θ_0 for distribution reference line
Gamma-Options	
ALPHA=	specifies mandatory shape parameter α
SIGMA=	specifies σ_0 for distribution reference line
THETA=	specifies θ_0 for distribution reference line
Gumbel-Options	
MU=	specifies location parameter μ
SIGMA=	specifies scale parameter σ
Lognormal-Options	
SIGMA=	specifies mandatory shape parameter σ
SLOPE=	specifies slope of distribution reference line
THETA=	specifies θ_0 for distribution reference line
ZETA=	specifies ζ_0 for distribution reference line (slope is $\exp(\zeta_0)$)
Normal-Options	
MU=	specifies μ_0 for distribution reference line
SIGMA=	specifies σ_0 for distribution reference line
Pareto-Options	
ALPHA=	specifies mandatory shape parameter α
SIGMA=	specifies scale parameter σ
THETA=	specifies threshold parameter θ
Power-Options	
ALPHA=	specifies mandatory shape parameter α
SIGMA=	specifies scale parameter σ
THETA=	specifies threshold parameter θ

Table 6.60 (*continued*)

Option	Description
Rayleigh-Options	
SIGMA=	specifies scale parameter σ
THETA=	specifies threshold parameter θ
Weibull-Options	
C=	specifies mandatory shape parameter c
SIGMA=	specifies σ_0 for distribution reference line
THETA=	specifies θ_0 for distribution reference line
Weibull2-Options	
C=	specifies c_0 for distribution reference line (slope is $1/c_0$)
SIGMA=	specifies σ_0 for distribution reference line (intercept is $\log(\sigma_0)$)
SLOPE=	specifies slope of distribution reference line
THETA=	specifies known lower threshold θ_0

General Options

Table 6.61 lists options that control the appearance of the plots.

Table 6.61 General PROBPLOT Statement Options

Option	Description
General Plot Layout Options	
CONTENTS=	specifies table of contents entry for probability plot grouping
GRID	draws grid lines perpendicular to the percentile axis
HREF=	specifies reference lines perpendicular to the horizontal axis
HREFLABELS=	specifies line labels for HREF= lines
LEGEND=	identifies LEGEND statement
NADJ=	adjusts sample size (N) when computing percentiles
NOFRAME	suppresses frame around plotting area
NOLEGEND	suppresses legend
NOLINELEGEND	suppresses distribution reference line information in legend
NOSPECLEGEND	suppresses specifications information in legend
PCTLMINOR	requests minor tick marks for percentile axis
PCTLORDER=	specifies tick mark labels for percentile axis
RANKADJ=	adjusts ranks when computing percentiles
ROTATE	switches horizontal and vertical axes
SQUARE	displays plot in square format
VREF=	specifies reference lines perpendicular to the vertical axis
VREFLABELS=	specifies line labels for VREF= lines
Graphics Options	
ANNOTATE=	specifies annotate data set
CAXIS=	specifies color for axis
CFRAME=	specifies color for frame
CGRID=	specifies color for grid lines

Table 6.61 (*continued*)

Option	Description
CHREF=	specifies colors for HREF= lines
CTEXT=	specifies color for text
CSTATREF=	specifies colors for STATREF= lines
CVREF=	specifies colors for VREF= lines
DESCRIPTION=	specifies description for plot in graphics catalog
FONT=	specifies software font for text
HAXIS=	specifies AXIS statement for horizontal axis
HEIGHT=	specifies height of text used outside framed areas
HMINOR=	specifies number of horizontal minor tick marks
HREFLABPOS=	specifies position for HREF= line labels
INFONT=	specifies software font for text inside framed areas
INHEIGHT=	specifies height of text inside framed areas
LGRID=	specifies a line type for grid lines
LHREF=	specifies line styles for HREF= lines
LSTATREF=	specifies line styles for STATREF= lines
LVREF=	specifies line styles for VREF= lines
NAME=	specifies name for plot in graphics catalog
NOHLABEL	suppresses label for horizontal axis
NOVLABEL	suppresses label for vertical axis
NOVTICK	suppresses tick marks and tick mark labels for vertical axis
STATREF=	specifies reference lines at values of summary statistics
STATREFLABELS=	specifies labels for STATREF= lines
STATREFSUBCHAR=	specifies substitution character for displaying statistic values in STATREFLABELS= labels
TURNVLABELS	turns and vertically strings out characters in labels for vertical axis
VAXIS=	specifies AXIS statement for vertical axis
VAXISLABEL=	specifies label for vertical axis
VMINOR=	specifies number of vertical minor tick marks
VREFLABPOS=	specifies horizontal position of labels for VREF= lines
WAXIS=	specifies line thickness for axes and frame
WGRID=	specifies line thickness for grid
Options for ODS Graphics Output	
ODSFOOTNOTE=	specifies footnote displayed on probability plot
ODSFOOTNOTE2=	specifies secondary footnote displayed on probability plot
ODSTITLE=	specifies title displayed on probability plot
ODSTITLE2=	specifies secondary title displayed on probability plot
Options for Comparative Plots	
ANNOKEY	applies annotation to key cell only
CFRAMESIDE=	specifies color for filling frame for row labels
CFRAMETOP=	specifies color for filling frame for column labels
CPROP=	specifies color for proportion of frequency bar
CTEXTSIDE=	specifies color for row labels
CTEXTTOP=	specifies color for column labels

Table 6.61 (*continued*)

Option	Description
INTERTILE=	specifies distance between tiles
NCOLS=	specifies number of columns in comparative probability plot
NROWS=	specifies number of rows in comparative probability plot
OVERLAY	overlays plots for different class levels (ODS Graphics only)
Options to Enhance Line Printer Plots	
GRIDCHAR=	specifies character for GRID lines
HREFCHAR=	specifies character for HREF= lines
NOOBSLEGEND	suppresses legend for hidden points
PROBSYMBOL=	specifies character for plotted points
VREFCHAR=	specifies character for VREF= lines

Dictionary of Options

The following sections provide detailed descriptions of options specific to the PROBPLOT statement. See "Dictionary of Common Options: CAPABILITY Procedure" on page 533 for detailed descriptions of options common to all the plot statements.

General Options

You can specify the following options whether you are producing ODS Graphics output or traditional graphics:

ALPHA=*value-list*|**EST**

specifies values for a mandatory shape parameter α ($\alpha > 0$) for probability plots requested with the BETA, GAMMA, PARETO, and POWER options. A plot is created for each value specified. For examples, see the entries for the distribution options. If you specify ALPHA=EST, a maximum likelihood estimate is computed for α.

BETA(ALPHA=*value-list*|**EST BETA=***value-list*|**EST <** *beta-options* **>)**

creates a beta probability plot for each combination of the shape parameters α and β given by the mandatory ALPHA= and BETA= options. If you specify ALPHA=EST and BETA=EST, a plot is created based on maximum likelihood estimates for α and β. In the following examples, the first PROBPLOT statement produces one plot, the second statement produces four plots, the third statement produces six plots, and the fourth statement produces one plot:

```
proc capability data=measures;
   probplot width / beta(alpha=2 beta=2);
   probplot width / beta(alpha=2 3 beta=1 2);
   probplot width / beta(alpha=2 to 3 beta=1 to 2 by 0.5);
   probplot width / beta(alpha=est beta=est);
run;
```

To create the plot, the observations are ordered from smallest to largest, and the ith ordered observation is plotted against the quantile $B^{-1}_{\alpha\beta}\left(\frac{i-0.375}{n+0.25}\right)$, where $B^{-1}_{\alpha\beta}(\cdot)$ is the inverse normalized incomplete beta

function, n is the number of nonmissing observations, and α and β are the shape parameters of the beta distribution. The horizontal axis is scaled in percentile units.

The point pattern on the plot for ALPHA=α and BETA=β tends to be linear with intercept θ and slope σ if the data are beta distributed with the specific density function

$$p(x) = \begin{cases} \frac{(x-\theta)^{\alpha-1}(\theta+\sigma-x)^{\beta-1}}{B(\alpha,\beta)\sigma^{(\alpha+\beta-1)}} & \text{for } \theta < x < \theta+\sigma \\ 0 & \text{for } x \le \theta \text{ or } x \ge \theta+\sigma \end{cases}$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and

θ = lower threshold parameter
σ = scale parameter ($\sigma > 0$)
α = first shape parameter ($\alpha > 0$)
β = second shape parameter ($\beta > 0$)

The intercept and slope are based on the quantile scale for the horizontal axis, which is displayed on a Q-Q plot; see "QQPLOT Statement: CAPABILITY Procedure" on page 492.

To obtain graphical estimates of α and β, specify lists of values for the ALPHA= and BETA= options, and select the combination of α and β that most nearly linearizes the point pattern.

To assess the point pattern, you can add a diagonal distribution reference line corresponding to θ_0 and σ_0 with the *beta-options* THETA=θ_0 and SIGMA=σ_0. Alternatively, you can add a line corresponding to estimated values of θ_0 and σ_0 with the *beta-options* THETA=EST and SIGMA=EST. Specify these options in parentheses, as in the following example:

```
proc capability data=measures;
   probplot width / beta(alpha=2 beta=3 theta=4 sigma=5);
run;
```

Agreement between the reference line and the point pattern indicates that the beta distribution with parameters α, β, θ_0 and σ_0 is a good fit. You can specify the SCALE= option as an alias for the SIGMA= option and the THRESHOLD= option as an alias for the THETA= option.

BETA=*value-list*|**EST**
specifies values for the shape parameter β ($\beta > 0$) for probability plots requested with the BETA distribution option. A plot is created for each value specified with the BETA= option. If you specify BETA=EST, a maximum likelihood estimate is computed for β. For examples, see the preceding entry for the BETA option.

C=*value(-list)*|**EST**
specifies the shape parameter c ($c > 0$) for probability plots requested with the WEIBULL and WEIBULL2 options. You must specify C= as a *Weibull-option* with the WEIBULL option; in this situation it accepts a list of values, or if you specify C=EST, a maximum likelihood estimate is computed for c. You can optionally specify C=*value* or C=EST as a *Weibull2-option* with the WEIBULL2 option to request a distribution reference line; in this situation, you must also specify SIGMA=*value* or SIGMA=EST.

For example, the first PROBPLOT statement below creates three three-parameter Weibull plots corresponding to the shape parameters $c = 1$, $c = 2$, and $c = 3$. The second PROBPLOT statement

creates a single three-parameter Weibull plot corresponding to an estimated value of c. The third PROBPLOT statement creates a single two-parameter Weibull plot with a distribution reference line corresponding to $c_0 = 2$ and $\sigma_0 = 3$.

```
proc capability data=measures;
   probplot width / weibull(c=1 2 3);
   probplot width / weibull(c=est);
   probplot width / weibull2(c=2 sigma=3);
run;
```

EXPONENTIAL< (*exponential-options*) >

EXP(< *exponential-options* >)

creates an exponential probability plot. To create the plot, the observations are ordered from smallest to largest, and the ith ordered observation is plotted against the quantile $-\log\left(1 - \frac{i-0.375}{n+0.25}\right)$, where n is the number of nonmissing observations. The horizontal axis is scaled in percentile units.

The point pattern on the plot tends to be linear with intercept θ and slope σ if the data are exponentially distributed with the specific density function

$$p(x) = \begin{cases} \frac{1}{\sigma}\exp\left(-\frac{x-\theta}{\sigma}\right) & \text{for } x \geq \theta \\ 0 & \text{for } x < \theta \end{cases}$$

where θ is a threshold parameter, and σ is a positive scale parameter.

The intercept and slope are based on the quantile scale for the horizontal axis, which is displayed on a Q-Q plot; see "QQPLOT Statement: CAPABILITY Procedure" on page 492.

To assess the point pattern, you can add a diagonal distribution reference line corresponding to θ_0 and σ_0 with the *exponential-options* THETA=θ_0 and SIGMA=σ_0. Alternatively, you can add a line corresponding to estimated values of θ_0 and σ_0 with the *exponential-options* THETA=EST and SIGMA=EST. Specify these options in parentheses, as in the following example:

```
proc capability data=measures;
   probplot width / exponential(theta=4 sigma=5);
run;
```

Agreement between the reference line and the point pattern indicates that the exponential distribution with parameters θ_0 and σ_0 is a good fit. You can specify the SCALE= option as an alias for the SIGMA= option and the THRESHOLD= option as an alias for the THETA= option.

GAMMA(ALPHA=*value-list*|EST < *gamma-options* >)

creates a gamma probability plot for each value of the shape parameter α given by the mandatory ALPHA= option. If you specify ALPHA=EST, a plot is created based on a maximum likelihood estimate for α.

For example, the first PROBPLOT statement below creates three plots corresponding to $\alpha = 0.4$, $\alpha = 0.5$, and $\alpha = 0.6$. The second PROBPLOT statement creates a single plot.

```
proc capability data=measures;
   probplot width / gamma(alpha=0.4 to 0.6 by 0.2);
   probplot width / gamma(alpha=est);
run;
```

To create the plot, the observations are ordered from smallest to largest, and the *i*th ordered observation is plotted against the quantile $G_\alpha^{-1}\left(\frac{i-0.375}{n+0.25}\right)$, where $G_\alpha^{-1}(\cdot)$ is the inverse normalized incomplete gamma function, n is the number of nonmissing observations, and α is the shape parameter of the gamma distribution. The horizontal axis is scaled in percentile units.

The point pattern on the plot for ALPHA=α tends to be linear with intercept θ and slope σ if the data are gamma distributed with the specific density function

$$p(x) = \begin{cases} \frac{1}{\sigma\Gamma(\alpha)}\left(\frac{x-\theta}{\sigma}\right)^{\alpha-1}\exp\left(-\frac{x-\theta}{\sigma}\right) & \text{for } x > \theta \\ 0 & \text{for } x \le \theta \end{cases}$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)
α = shape parameter ($\alpha > 0$)

The intercept and slope are based on the quantile scale for the horizontal axis, which is displayed on a Q-Q plot; see "QQPLOT Statement: CAPABILITY Procedure" on page 492.

To obtain a graphical estimate of α, specify a list of values for the ALPHA= option, and select the value that most nearly linearizes the point pattern.

To assess the point pattern, you can add a diagonal distribution reference line corresponding to θ_0 and σ_0 with the *gamma-options* THETA=θ_0 and SIGMA=σ_0. Alternatively, you can add a line corresponding to estimated values of θ_0 and σ_0 with the *gamma-options* THETA=EST and SIGMA=EST. Specify these options in parentheses, as in the following example:

```
proc capability data=measures;
   probplot width / gamma(alpha=2 theta=3 sigma=4);
run;
```

Agreement between the reference line and the point pattern indicates that the gamma distribution with parameters α, θ_0 and σ_0 is a good fit. You can specify the SCALE= option as an alias for the SIGMA= option and the THRESHOLD= option as an alias for the THETA= option.

GRID

draws reference lines perpendicular to the percentile axis at major tick marks.

GUMBEL(< *Gumbel-options* **>)**

creates a Gumbel probability plot. To create the plot, the observations are ordered from smallest to largest, and the *i*th ordered observation is plotted against the quantile $-\log\left(-\log\left(\frac{i-0.375}{n+0.25}\right)\right)$, where n is the number of nonmissing observations. The horizontal axis is scaled in percentile units.

The point pattern on the plot tends to be linear with intercept μ and slope σ if the data are Gumbel distributed with the specific density function

$$p(x) = \frac{e^{-(x-\mu)/\sigma}}{\sigma} \exp\left(-e^{-(x-\mu)/\sigma}\right)$$

where μ is a location parameter and σ is a positive scale parameter.

The intercept and slope are based on the quantile scale for the horizontal axis, which is displayed on a Q-Q plot; see "QQPLOT Statement: CAPABILITY Procedure" on page 492.

To assess the point pattern, you can add a diagonal distribution reference line corresponding to μ_0 and σ_0 with the *Gumbel-options* MU=μ_0 and SIGMA=σ_0. Alternatively, you can add a line corresponding to estimated values of μ_0 and σ_0 with the *Gumbel-options* MU=EST and SIGMA=EST. Specify these options in parentheses following the GUMBEL option.

Agreement between the reference line and the point pattern indicates that the Gumbel distribution with parameters μ_0 and σ_0 is a good fit.

LOGNORMAL(SIGMA=*value-list***|EST <** *lognormal-options* **>)**

LNORM(SIGMA=*value-list***|EST <** *lognormal-options* **>)**

creates a lognormal probability plot for each value of the shape parameter σ given by the mandatory SIGMA= option or its alias, the SHAPE= option. If you specify SIGMA=EST, a plot is created based on a maximum likelihood estimate for σ.

For example, the first PROBPLOT statement below produces two plots, and the second PROBPLOT statement produces a single plot:

```
proc capability data=measures;
   probplot width / lognormal(sigma=1.5 2.5 l=2);
   probplot width / lognormal(sigma=est);
run;
```

To create the plot, the observations are ordered from smallest to largest, and the *i*th ordered observation is plotted against the quantile $\exp\left(\sigma\,\Phi^{-1}\left(\frac{i-0.375}{n+0.25}\right)\right)$, where $\Phi^{-1}(\cdot)$ is the inverse standard cumulative normal distribution, n is the number of nonmissing observations, and σ is the shape parameter of the lognormal distribution. The horizontal axis is scaled in percentile units.

The point pattern on the plot for SIGMA=σ tends to be linear with intercept θ and slope $\exp(\zeta)$ if the data are lognormally distributed with the specific density function

$$p(x) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}(x-\theta)} \exp\left(-\frac{(\log(x-\theta)-\zeta)^2}{2\sigma^2}\right) & \text{for } x > \theta \\ 0 & \text{for } x \le \theta \end{cases}$$

where

θ = threshold parameter
ζ = scale parameter
σ = shape parameter ($\sigma > 0$)

The intercept and slope are based on the quantile scale for the horizontal axis, which is displayed on a Q-Q plot; see "QQPLOT Statement: CAPABILITY Procedure" on page 492.

To obtain a graphical estimate of σ, specify a list of values for the SIGMA= option, and select the value that most nearly linearizes the point pattern.

To assess the point pattern, you can add a diagonal distribution reference line corresponding to θ_0 and ζ_0 with the *lognormal-options* THETA=θ_0 and ZETA=ζ_0. Alternatively, you can add a line corresponding to estimated values of θ_0 and ζ_0 with the *lognormal-options* THETA=EST and ZETA=EST.

Specify these options in parentheses, as in the following example:

```
proc capability data=measures;
   probplot width / lognormal(sigma=2 theta=3 zeta=0);
run;
```

Agreement between the reference line and the point pattern indicates that the lognormal distribution with parameters σ, θ_0, and ζ_0 is a good fit. See Example 6.20 for an example.

You can specify the THRESHOLD= option as an alias for the THETA= option and the SCALE= option as an alias for the ZETA= option.

MU=*value*|**EST**
specifies the mean μ_0 for a probability plot requested with the GUMBEL and NORMAL options. If you specify MU=EST, μ_0 is equal to the sample mean for the normal distribution. For the Gumbel distribution, a maximum likelihood estimate is calculated. See Example 6.19.

NADJ=*value*
specifies the adjustment value added to the sample size in the calculation of theoretical percentiles. The default is $\frac{1}{4}$, as recommended by Blom (1958). Also refer to Chambers et al. (1983) for additional information.

NOLEGEND
suppresses legends for specification limits, fitted curves, distribution lines, and hidden observations.

NOLINELEGEND
NOLINEL
suppresses the legend for the optional distribution reference line.

NORMAL<(*normal-options***)>**
NORM<(*normal-options***)>**
creates a normal probability plot. This is the default if you do not specify a distribution option. To create the plot, the observations are ordered from smallest to largest, and the *i*th ordered observation is plotted against the quantile $\Phi^{-1}\left(\frac{i-0.375}{n+0.25}\right)$, where $\Phi^{-1}(\cdot)$ is the inverse cumulative standard normal distribution, and n is the number of nonmissing observations. The horizontal axis is scaled in percentile units.

The point pattern on the plot tends to be linear with intercept μ and slope σ if the data are normally distributed with the specific

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad \text{for all } x$$

where μ is the mean and σ is the standard deviation ($\sigma > 0$).

The intercept and slope are based on the quantile scale for the horizontal axis, which is displayed on a Q-Q plot; see "QQPLOT Statement: CAPABILITY Procedure" on page 492.

To assess the point pattern, you can add a diagonal distribution reference line corresponding to μ_0 and σ_0 with the *normal-options* MU=μ_0 and SIGMA=σ_0. Alternatively, you can add a line corresponding to estimated values of μ_0 and σ_0 with the *normal-options* MU=EST and SIGMA=EST; the estimates of μ_0 and σ_0 are the sample mean and sample standard deviation.

Specify these options in parentheses, as in the following example:

```
proc capability data=measures;
   probplot length / normal(mu=10 sigma=0.3);
   probplot length / normal(mu=est sigma=est);
run;
```

Agreement between the reference line and the point pattern indicates that the normal distribution with parameters μ_0 and σ_0 is a good fit.

NOSPECLEGEND
NOSPECL
suppresses the legend for specification limit reference lines.

PARETO(< *Pareto-options* **>)**
creates a generalized Pareto probability plot for each value of the shape parameter α given by the mandatory ALPHA= option. If you specify ALPHA=EST, a plot is created based on a maximum likelihood estimate for α.

To create the plot, the observations are ordered from smallest to largest, and the ith ordered observation is plotted against the quantile $(1-(1-\frac{i-0.375}{n+0.25})^{\alpha})/\alpha$ $(\alpha \neq 0)$ or $-\log(1-\frac{i-0.375}{n+0.25})$ $(\alpha = 0)$, where n is the number of nonmissing observations and α is the shape parameter of the generalized Pareto distribution. The horizontal axis is scaled in percentile units.

The point pattern on the plot for ALPHA=α tends to be linear with intercept θ and slope σ if the data are generalized Pareto distributed with the specific density function

$$p(x) = \begin{cases} \frac{1}{\sigma}(1-\alpha(x-\theta)/\sigma)^{1/\alpha-1} & \text{if } \alpha \neq 0 \\ \frac{1}{\sigma}\exp(-(x-\theta)/\sigma) & \text{if } \alpha = 0 \end{cases}$$

where

θ = threshold parameter
σ = scale parameter $(\sigma > 0)$
α = shape parameter $(\alpha > 0)$

The intercept and slope are based on the quantile scale for the horizontal axis, which is displayed on a Q-Q plot; see "QQPLOT Statement: CAPABILITY Procedure" on page 492.

To obtain a graphical estimate of α, specify a list of values for the ALPHA= option, and select the value that most nearly linearizes the point pattern.

To assess the point pattern, you can add a diagonal distribution reference line corresponding to θ_0 and σ_0 with the *Pareto-options* THETA=θ_0 and SIGMA=σ_0. Alternatively, you can add a line corresponding

to estimated values of θ_0 and σ_0 with the *Pareto-options* THETA=EST and SIGMA=EST. Specify these options in parentheses following the PARETO option.

Agreement between the reference line and the point pattern indicates that the generalized Pareto distribution with parameters α, θ_0, and σ_0 is a good fit.

PCTLORDER=*value-list*

specifies the tick mark values labeled on the theoretical percentile axis. Because the values are percentiles, the labels must be between 0 and 100, exclusive. The values must be listed in increasing order and must cover the plotted percentile range. Otherwise, a default list is used. For example, consider the following:

```
proc capability data=measures;
   probplot length / pctlorder=1 10 25 50 75 90 99;
run;
```

Note that the ORDER= option in the AXIS statement is not supported by the PROBPLOT statement.

POWER(< *power-options* **>)**

creates a power function probability plot for each value of the shape parameter α given by the mandatory ALPHA= option. If you specify ALPHA=EST, a plot is created based on a maximum likelihood estimate for α.

To create the plot, the observations are ordered from smallest to largest, and the ith ordered observation is plotted against the quantile $B^{-1}_{\alpha(1)}\left(\frac{i-0.375}{n+0.25}\right)$, where $B^{-1}_{\alpha(1)}(\cdot)$ is the inverse normalized incomplete beta function, n is the number of nonmissing observations, α is one shape parameter of the beta distribution, and the second shape parameter, $\beta = 1$. The horizontal axis is scaled in percentile units.

The point pattern on the plot for ALPHA=α tends to be linear with intercept θ and slope σ if the data are power function distributed with the specific density function

$$p(x) = \begin{cases} \frac{\alpha}{\sigma}\left(\frac{x-\theta}{\sigma}\right)^{\alpha-1} & \text{for } \theta < x < \theta + \sigma \\ 0 & \text{for } x \le \theta \text{ or } x \ge \theta + \sigma \end{cases}$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)
α = shape parameter ($\alpha > 0$)

The intercept and slope are based on the quantile scale for the horizontal axis, which is displayed on a Q-Q plot; see "QQPLOT Statement: CAPABILITY Procedure" on page 492.

To obtain a graphical estimate of α, specify a list of values for the ALPHA= option, and select the value that most nearly linearizes the point pattern.

To assess the point pattern, you can add a diagonal distribution reference line corresponding to θ_0 and σ_0 with the *power-options* THETA=θ_0 and SIGMA=σ_0. Alternatively, you can add a line corresponding to estimated values of θ_0 and σ_0 with the *power-options* THETA=EST and SIGMA=EST. Specify these options in parentheses following the POWER option.

Agreement between the reference line and the point pattern indicates that the power function distribution with parameters α, θ_0, and σ_0 is a good fit.

RANKADJ=*value*

specifies the adjustment value added to the ranks in the calculation of theoretical percentiles. The default is $-\frac{3}{8}$, as recommended by Blom (1958). Also refer to Chambers et al. (1983) for additional information.

RAYLEIGH(< *Rayleigh-options* **>)**

creates a Rayleigh probability plot. To create the plot, the observations are ordered from smallest to largest, and the *i*th ordered observation is plotted against the quantile $\sqrt{-2\log\left(1-\frac{i-0.375}{n+0.25}\right)}$, where n is the number of nonmissing observations. The horizontal axis is scaled in percentile units.

The point pattern on the plot tends to be linear with intercept θ and slope σ if the data are Rayleigh distributed with the specific density function

$$p(x) = \begin{cases} \frac{x-\theta}{\sigma^2}\exp(-(x-\theta)^2/(2\sigma^2)) & \text{for } x \geq \theta \\ 0 & \text{for } x < \theta \end{cases}$$

where θ is a threshold parameter, and σ is a positive scale parameter.

The intercept and slope are based on the quantile scale for the horizontal axis, which is displayed on a Q-Q plot; see "QQPLOT Statement: CAPABILITY Procedure" on page 492.

To assess the point pattern, you can add a diagonal distribution reference line corresponding to θ_0 and σ_0 with the *Rayleigh-options* THETA=θ_0 and SIGMA=σ_0. Alternatively, you can add a line corresponding to estimated values of θ_0 and σ_0 with the *Rayleigh-options* THETA=EST and SIGMA=EST. Specify these options in parentheses after the RAYLEIGH option.

Agreement between the reference line and the point pattern indicates that the Rayleigh distribution with parameters θ_0 and σ_0 is a good fit.

ROTATE

switches the horizontal and vertical axes so that the theoretical percentiles are plotted vertically while the data are plotted horizontally. Regardless of whether the plot has been rotated, horizontal axis options (such as HAXIS=) still refer to the horizontal axis, and vertical axis options (such as VAXIS=) still refer to the vertical axis. All other options that depend on axis placement adjust to the rotated axes.

SIGMA=*value-list*|**EST**

specifies the value of the parameter σ, where $\sigma > 0$. Alternatively, you can specify SIGMA=EST to request a maximum likelihood estimate for σ_0. The interpretation and use of the SIGMA= option depend on the distribution option with which it is specified, as indicated by the following table.

Distribution Option	Use of the SIGMA= Option
BETA EXPONENTIAL GAMMA PARETO POWER RAYLEIGH WEIBULL	THETA=θ_0 and SIGMA=σ_0 request a distribution reference line corresponding to θ_0 and σ_0.
GUMBEL	MU=μ_0 and SIGMA=σ_0 request a distribution reference line corresponding to μ_0 and σ_0.
LOGNORMAL	SIGMA=$\sigma_1 \ldots \sigma_n$ requests n probability plots with shape parameters $\sigma_1 \ldots \sigma_n$. The SIGMA= option must be specified.
NORMAL	MU=μ_0 and SIGMA=σ_0 request a distribution reference line corresponding to μ_0 and σ_0. SIGMA=EST requests a line with σ_0 equal to the sample standard deviation.
WEIBULL2	SIGMA=σ_0 and C=c_0 request a distribution reference line corresponding to σ_0 and c_0.

In the following example, the first PROBPLOT statement requests a normal plot with a distribution reference line corresponding to $\mu_0 = 5$ and $\sigma_0 = 2$, and the second PROBPLOT statement requests a lognormal plot with shape parameter $\sigma = 3$:

```
proc capability data=measures;
   probplot length / normal(mu=5 sigma=2);
   probplot width  / lognormal(sigma=3);
run;
```

SLOPE=*value*|**EST**

specifies the slope for a distribution reference line requested with the LOGNORMAL and WEIBULL2 options. The intercept and slope are based on the quantile scale for the horizontal axis, which is displayed on a Q-Q plot; see "QQPLOT Statement: CAPABILITY Procedure" on page 492.

When you use the SLOPE= option with the LOGNORMAL option, you must also specify a threshold parameter value θ_0 with the THETA= *lognormal-option* to request the line. The SLOPE= option is an alternative to the ZETA= *lognormal-option* for specifying ζ_0, because the slope is equal to $\exp(\zeta_0)$.

When you use the SLOPE= option with the WEIBULL2 option, you must also specify a scale parameter value σ_0 with the SIGMA= *Weibull2-option* to request the line. The SLOPE= option is an alternative to the C= *Weibull2-option* for specifying c_0, because the slope is equal to $1/c_0$. See "Location and Scale Parameters" on page 486.

For example, the first and second PROBPLOT statements below produce the same set of probability plots as the third and fourth PROBPLOT statements:

```
proc capability data=measures;
   probplot width / lognormal(sigma=2 theta=0 zeta=0);
   probplot width / weibull2(sigma=2 theta=0 c=0.25);
   probplot width / lognormal(sigma=2 theta=0 slope=1);
   probplot width / weibull2(sigma=2 theta=0 slope=4);
run;
```

SQUARE

displays the probability plot in a square frame. For an example, see Output 6.20.1. The default is a rectangular frame.

THETA=*value*|**EST**

THRESHOLD=*value*

specifies the lower threshold parameter θ for probability plots requested with the BETA, EXPONENTIAL, GAMMA, LOGNORMAL, PARETO, POWER, RAYLEIGH, WEIBULL, and WEIBULL2 options. When used with the WEIBULL2 option, the THETA= option specifies the known lower threshold θ_0, for which the default is 0. When used with the other distribution options, the THETA= option specifies θ_0 for a distribution reference line; alternatively in this situation, you can specify THETA=EST to request a maximum likelihood estimate for θ_0. To request the line, you must also specify a scale parameter. See Output 6.20.1 for an example of the THETA= option with a lognormal probability plot.

WEIBULL(C=*value-list*|**EST** < *Weibull-options* >**)**

WEIB(C=*value-list* < *Weibull-options* >**)**

creates a three-parameter Weibull probability plot for each value of the shape parameter c given by the mandatory C= option or its alias, the SHAPE= option. If you specify C=EST, a plot is created based on a maximum likelihood estimate for c. In the following example, the first PROBPLOT statement creates four plots, and the second PROBPLOT statement creates a single plot:

```
proc capability data=measures;
   probplot width / weibull(c=1.8 to 2.4 by 0.2 w=2);
   probplot width / weibull(c=est);
run;
```

To create the plot, the observations are ordered from smallest to largest, and the ith ordered observation is plotted against the quantile $\left(-\log\left(1-\frac{i-0.375}{n+0.25}\right)\right)^{\frac{1}{c}}$, where n is the number of nonmissing observations, and c is the Weibull distribution shape parameter. The horizontal axis is scaled in percentile units.

The point pattern on the plot for C=c tends to be linear with intercept θ and slope σ if the data are Weibull distributed with the specific density function

$$p(x) = \begin{cases} \frac{c}{\sigma}\left(\frac{x-\theta}{\sigma}\right)^{c-1}\exp\left(-\left(\frac{x-\theta}{\sigma}\right)^{c}\right) & \text{for } x > \theta \\ 0 & \text{for } x \le \theta \end{cases}$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)
c = shape parameter ($c > 0$)

The intercept and slope are based on the quantile scale for the horizontal axis, which is displayed on a Q-Q plot; see "QQPLOT Statement: CAPABILITY Procedure" on page 492.

To obtain a graphical estimate of c, specify a list of values for the C= option, and select the value that most nearly linearizes the point pattern.

To assess the point pattern, you can add a diagonal distribution reference line corresponding to θ_0 and σ_0 with the *Weibull-options* THETA=θ_0 and SIGMA=σ_0. Alternatively, you can add a line corresponding to estimated values of θ_0 and σ_0 with the *Weibull-options* THETA=EST and SIGMA=EST. Specify these options in parentheses, as in the following example:

```
proc capability data=measures;
   probplot width / weibull(c=2 theta=3 sigma=4);
run;
```

Agreement between the reference line and the point pattern indicates that the Weibull distribution with parameters c, θ_0, and σ_0 is a good fit. You can specify the SCALE= option as an alias for the SIGMA= option and the THRESHOLD= option as an alias for the THETA= option.

WEIBULL2<(*Weibull2-options***)>**
W2<(*Weibull2-options***)>**

creates a two-parameter Weibull probability plot. You should use the WEIBULL2 option when your data have a *known* lower threshold θ_0. You can specify the threshold value θ_0 with the THETA= *Weibull2-option* or its alias, the THRESHOLD= *Weibull2-option*. The default is $\theta_0 = 0$.

To create the plot, the observations are ordered from smallest to largest, and the log of the shifted ith ordered observation $x_{(i)}$, denoted by $\log(x_{(i)} - \theta_0)$, is plotted against the quantile $\log\left(-\log\left(1 - \frac{i-0.375}{n+0.25}\right)\right)$, where n is the number of nonmissing observations. The horizontal axis is scaled in percentile units. Note that the C= shape parameter option is not mandatory with the WEIBULL2 option.

The point pattern on the plot for THETA=θ_0 tends to be linear with intercept $\log(\sigma)$ and slope $\frac{1}{c}$ if the data are Weibull distributed with the specific density function

$$p(x) = \begin{cases} \frac{c}{\sigma}\left(\frac{x-\theta_0}{\sigma}\right)^{c-1} \exp\left(-\left(\frac{x-\theta_0}{\sigma}\right)^{c}\right) & \text{for } x > \theta_0 \\ 0 & \text{for } x \le \theta_0 \end{cases}$$

where

θ_0 = known lower threshold
σ = scale parameter ($\sigma > 0$)
c = shape parameter ($c > 0$)

An advantage of the two-parameter Weibull plot over the three-parameter Weibull plot is that the parameters c and σ can be estimated from the slope and intercept of the point pattern. A disadvantage is that the two-parameter Weibull distribution applies only in situations where the threshold parameter is known.

To assess the point pattern, you can add a diagonal distribution reference line corresponding to σ_0 and c_0 with the *Weibull2-options* SIGMA=σ_0 and C=c_0. Alternatively, you can add a distribution reference line corresponding to estimated values of σ_0 and c_0 with the *Weibull2-options* SIGMA=EST and C=EST. Specify these options in parentheses, as in the following example:

```
proc capability data=measures;
   probplot width / weibull2(theta=3 sigma=4 c=2);
run;
```

Agreement between the distribution reference line and the point pattern indicates that the Weibull distribution with parameters c_0, θ_0 and σ_0 is a good fit. You can specify the SCALE= option as an alias for the SIGMA= option and the SHAPE= option as an alias for the C= option.

ZETA=*value*|**EST**
specifies a value for the scale parameter ζ for lognormal probability plots requested with the LOGNORMAL option. Specify THETA=θ_0 and ZETA=ζ_0 to request a distribution reference line with intercept θ_0 and slope $\exp(\zeta_0)$. See Output 6.20.1 for an example.

Options for Traditional Graphics

You can specify the following options if you are producing traditional graphics:

CGRID=*color*
specifies the color for the grid lines requested by the GRID option.

LEGEND=*name* | **NONE**
specifies the name of a LEGEND statement describing the legend for specification limit reference lines and fitted curves. Specifying LEGEND=NONE is equivalent to specifying the NOLEGEND option.

LGRID=*linetype*
specifies the line type for the grid lines requested by the GRID option.

PCTLMINOR
requests minor tick marks for the percentile axis. See Output 6.20.1 for an example.

WGRID=*n*
specifies the width of the grid lines requested with the GRID option. If you use the WGRID= option, you do not need to specify the GRID option.

Options for Legacy Line Printer Plots

You can specify the following options if you are producing legacy line printer plots:

GRIDCHAR='*character*'
specifies the character used for the lines requested by the GRID option for a line printer plot. The default is the vertical bar (|).

NOOBSLEGEND
NOOBSL
suppresses the legend that indicates the number of hidden observations.

PROBSYMBOL='*character*'
specifies the character used to mark the points in a line printer plot. The default is the plus sign (+).

SYMBOL='*character*'
specifies the character used to display the distribution reference line in a line printer plot. The default character is the first letter of the distribution option keyword.

Details: PROBPLOT Statement

This section provides details on the following topics:

- distributions supported by the PROBPLOT statement
- SYMBOL statement options

Summary of Theoretical Distributions

You can use the PROBPLOT statement to request probability plots based on the theoretical distributions summarized in Table 6.62.

Table 6.62 Distributions and Parameters

Distribution	Density Function $p(x)$	Range	Parameters		
			Location	Scale	Shape
Beta	$\frac{(x-\theta)^{\alpha-1}(\theta+\sigma-x)^{\beta-1}}{B(\alpha,\beta)\sigma^{(\alpha+\beta-1)}}$	$\theta < x < \theta+\sigma$	θ	σ	α, β
Exponential	$\frac{1}{\sigma}\exp\left(-\frac{x-\theta}{\sigma}\right)$	$x \geq \theta$	θ	σ	
Gamma	$\frac{1}{\sigma\Gamma(\alpha)}\left(\frac{x-\theta}{\sigma}\right)^{\alpha-1}\exp\left(-\frac{x-\theta}{\sigma}\right)$	$x > \theta$	θ	σ	α
Gumbel	$\frac{e^{-(x-\mu)/\sigma}}{\sigma}\exp\left(-e^{-(x-\mu)/\sigma}\right)$	all x	μ	σ	
Lognormal (3-parameter)	$\frac{1}{\sigma\sqrt{2\pi}(x-\theta)}\exp\left(-\frac{(\log(x-\theta)-\zeta)^2}{2\sigma^2}\right)$	$x > \theta$	θ	ζ	σ
Normal	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	all x	μ	σ	
Generalized Pareto	$\alpha \neq 0$ $\frac{1}{\sigma}(1-\alpha(x-\theta)/\sigma)^{1/\alpha-1}$ $\alpha = 0$ $\frac{1}{\sigma}\exp(-(x-\theta)/\sigma)$	$x > \theta$	θ	σ	α
Power Function	$\frac{\alpha}{\sigma}\left(\frac{x-\theta}{\sigma}\right)^{\alpha-1}$	$x > \theta$	θ	σ	α
Rayleigh	$\frac{x-\theta}{\sigma^2}\exp(-(x-\theta)^2/(2\sigma^2))$	$x \geq \theta$	θ	σ	
Weibull (3-parameter)	$\frac{c}{\sigma}\left(\frac{x-\theta}{\sigma}\right)^{c-1}\exp\left(-\left(\frac{x-\theta}{\sigma}\right)^{c}\right)$	$x > \theta$	θ	σ	c

Table 6.62 (*continued*)

Distribution	Density Function $p(x)$	Range	Parameters		
			Location	Scale	Shape
Weibull (2-parameter)	$\frac{c}{\sigma}\left(\frac{x-\theta_0}{\sigma}\right)^{c-1}\exp\left(-\left(\frac{x-\theta_0}{\sigma}\right)^{c}\right)$	$x > \theta_0$ (known)	θ_0	σ	c

You can request these distributions with the BETA, EXPONENTIAL, GAMMA, LOGNORMAL, NORMAL, WEIBULL, and WEIBULL2 options, respectively. If you do not specify a distribution option, a normal probability plot is created.

Shape Parameters

Some of the distribution options in the PROBPLOT statement require you to specify one or two shape parameters in parentheses after the distribution keyword. These are summarized in Table 6.63.

Table 6.63 Shape Parameter Options for the PROBPLOT Statement

Distribution Keyword	Mandatory Shape Parameter Option	Range
BETA	ALPHA=α, BETA=β	$\alpha > 0, \beta > 0$
EXPONENTIAL	None	
GAMMA	ALPHA=α	$\alpha > 0$
GUMBEL	None	
LOGNORMAL	SIGMA=σ	$\sigma > 0$
NORMAL	None	
PARETO	ALPHA=α	$\alpha > 0$
POWER	ALPHA=α	$\alpha > 0$
RAYLEIGH	None	
WEIBULL	C=c	$c > 0$
WEIBULL2	None	

You can visually estimate the value of a shape parameter by specifying a list of values for the shape parameter option. The PROBPLOT statement produces a separate plot for each value. You can then use the value of the shape parameter producing the most nearly linear point pattern. Alternatively, you can request that the plot be created using an estimated shape parameter. For an example, see "Creating Lognormal Probability Plots" on page 463.

Location and Scale Parameters

If you specify the location and scale parameters for a distribution (or if you request estimates for these parameters), a diagonal distribution reference line is displayed on the plot. (An exception is the two-parameter Weibull distribution, for which a line is displayed when you specify or estimate the scale and shape parameters.) Agreement between this line and the point pattern indicates that the distribution with these parameters is a good fit. For illustrations, see Example 6.19 and Example 6.20.

The following table shows how the specified parameters determine the intercept[5] and slope of the line:

Table 6.64 Intercept and Slope of Distribution Reference Line

	Parameters			Linear Pattern	
Distribution	**Location**	**Scale**	**Shape**	**Intercept**	**Slope**
Beta	θ	σ	α , β	θ	σ
Exponential	θ	σ		θ	σ
Gamma	θ	σ	α	θ	σ
Gumbel	μ	σ		μ	σ
Lognormal	θ	ζ	σ	θ	$\exp(\zeta)$
Normal	μ	σ		μ	σ
Generalized Pareto	θ	σ	α	θ	σ
Power Function	θ	σ	α	θ	σ
Rayleigh	θ	σ		θ	σ
Weibull (3-parameter)	θ	σ	c	θ	σ
Weibull (2-parameter)	θ_0 (known)	σ	c	$\log(\sigma)$	$\frac{1}{c}$

For the LOGNORMAL and WEIBULL2 options, you can specify the slope directly with the SLOPE= option. That is, for the LOGNORMAL option, specifying THETA=θ_0 and SLOPE=$\exp(\zeta_0)$ displays the same line as specifying THETA=θ_0 and ZETA=ζ_0. For the WEIBULL2 option, specifying SIGMA=σ_0 and SLOPE=$\frac{1}{c_0}$ displays the same line as specifying SIGMA=σ_0 and C=c_0.

SYMBOL Statement Options

In earlier releases of SAS/QC software, graphical features of lower and upper specification lines and diagonal distribution reference lines were controlled with options in the SYMBOL2, SYMBOL3, and SYMBOL4 statements, respectively. These options are still supported, although they have been superseded by options in the PROBPLOT and SPEC statements. Table 6.65 summarizes the two sets of options. **NOTE:** These statements have no effect on ODS Graphics output.

[5] The intercept and slope are based on the quantile scale for the horizontal axis, which is displayed on a Q-Q plot; see "QQPLOT Statement: CAPABILITY Procedure" on page 492.

Table 6.65 SYMBOL Statement Options

Feature	Statement and Options	Alternative Statement and Options
Symbol markers	**SYMBOL1 Statement**	
character	VALUE=*special-symbol*	
color	COLOR=*color*	
font	FONT=*font*	
height	HEIGHT=*value*	
Lower specification line	**SPEC Statement**	**SYMBOL2 Statement**
position	LSL=*value*	
color	CLSL=*color*	COLOR=*color*
line type	LLSL=*linetype*	LINE=*linetype*
width	WLSL=*value*	WIDTH=*value*
Upper specification line	**SPEC Statement**	**SYMBOL3 Statement**
position	USL=*value*	
color	CUSL=*color*	COLOR=*color*
line type	LUSL=*linetype*	LINE=*linetype*
width	WUSL=*value*	WIDTH=*value*
Target reference line	**SPEC Statement**	
position	TARGET=*value*	
color	CTARGET=*color*	
line type	LTARGET=*linetype*	
width	WTARGET=*value*	
Distribution reference line	**PROBPLOT Statement**	**SYMBOL4 Statement**
color	COLOR=*color*	COLOR=*color*
line type	LINE=*linetype*	LINE=*linetype*
width	WIDTH=*value*	WIDTH=*value*

For an illustration of these options, see Example 6.19.

ODS Graphics

Before you create ODS Graphics output, ODS Graphics must be enabled (for example, by using the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" (Chapter 21, *SAS/STAT User's Guide*).

The appearance of a graph produced with ODS Graphics is determined by the style associated with the ODS destination where the graph is produced. PROBPLOT options used to control the appearance of traditional graphics are ignored for ODS Graphics output.

When ODS Graphics is in effect, the PROBPLOT statement assigns a name to the graph it creates. You can use this name to reference the graph when using ODS. The name is listed in Table 6.66.

Table 6.66 ODS Graphics Produced by the PROBPLOT Statement

ODS Graph Name	Plot Description
ProbPlot	probability plot

See Chapter 4, "SAS/QC Graphics," for more information about ODS Graphics and other methods for producing charts.

Examples: PROBPLOT Statement

This section provides advanced examples of the PROBPLOT statement.

Example 6.19: Displaying a Normal Reference Line

NOTE: See *Probability Plot with Normal Reference Line* in the SAS/QC Sample Library.

Measurements of the distance between two holes cut into 50 steel sheets are saved as values of the variable Distance in the following data set:

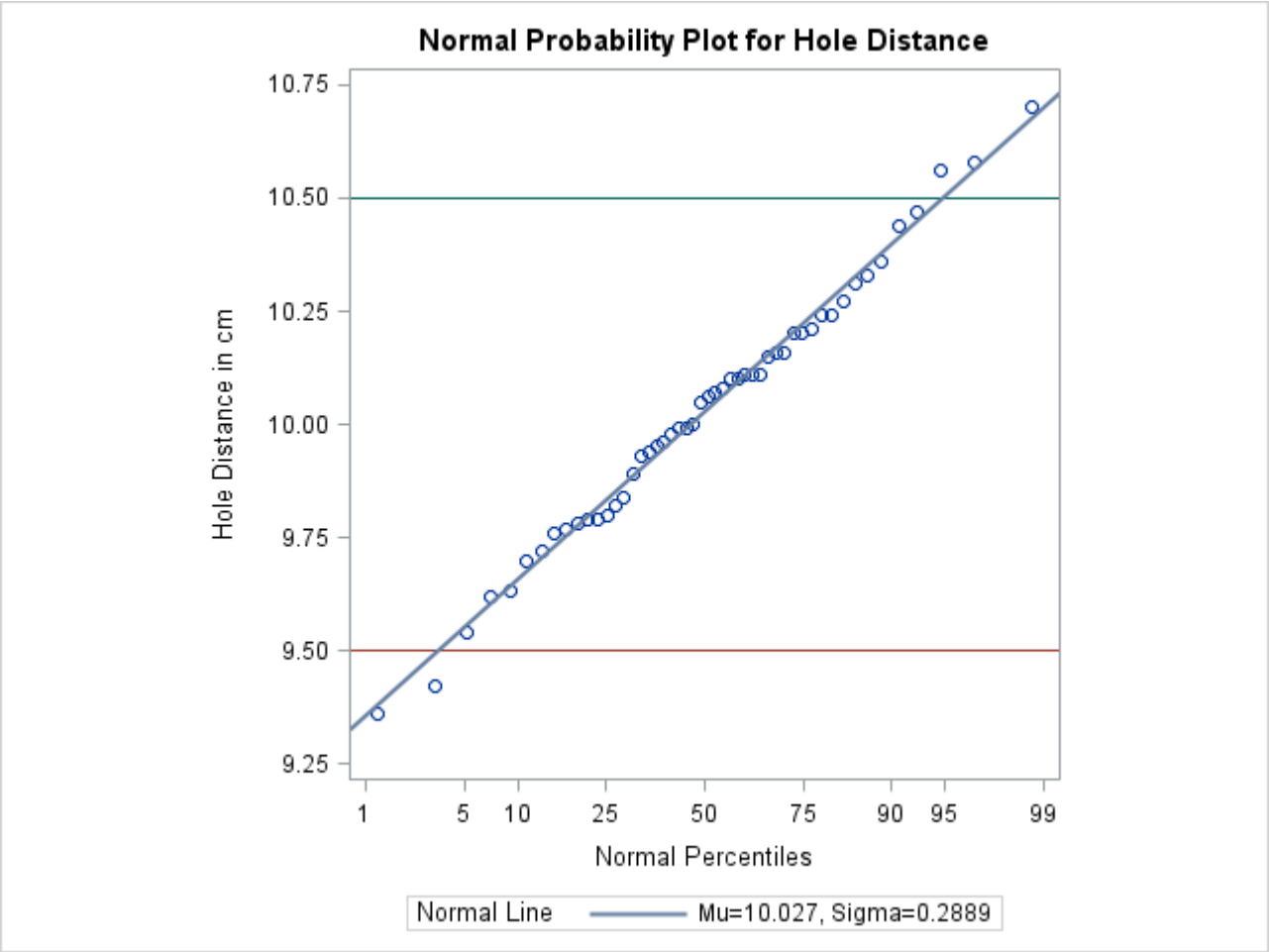
```
data Sheets;
   input Distance @@;
   label Distance='Hole Distance in cm';
   datalines;
 9.80 10.20 10.27  9.70  9.76
10.11 10.24 10.20 10.24  9.63
 9.99  9.78 10.10 10.21 10.00
 9.96  9.79 10.08  9.79 10.06
10.10  9.95  9.84 10.11  9.93
10.56 10.47  9.42 10.44 10.16
10.11 10.36  9.94  9.77  9.36
 9.89  9.62 10.05  9.72  9.82
 9.99 10.16 10.58 10.70  9.54
10.31 10.07 10.33  9.98 10.15
;
```

The cutting process is in control, and you decide to check whether the process distribution is normal. The following statements create a normal probability plot for Distance with lower and upper specification lines at 9.5 cm and 10.5 cm:

```
title 'Normal Probability Plot for Hole Distance';
proc capability data=Sheets noprint;
   spec lsl=9.5 usl=10.5;
   probplot Distance / normal(mu=est sigma=est)
                       square
                       odstitle=title
                       nospeclegend;
run;
```

The plot is shown in Output 6.19.1. The MU= and SIGMA= *normal-options* request the diagonal reference line that corresponds to the normal distribution with estimated parameters $\hat{\mu} = 10.027$ and $\hat{\sigma} = 0.2889$. The LSL= and USL= SPEC statement options request the lower and upper specification lines. The SYMBOL statement specifies the symbol marker for the plotted points.

Output 6.19.1 Normal Reference Line



Example 6.20: Displaying a Lognormal Reference Line

NOTE: See *Creating Lognormal Probability Plots* in the SAS/QC Sample Library.

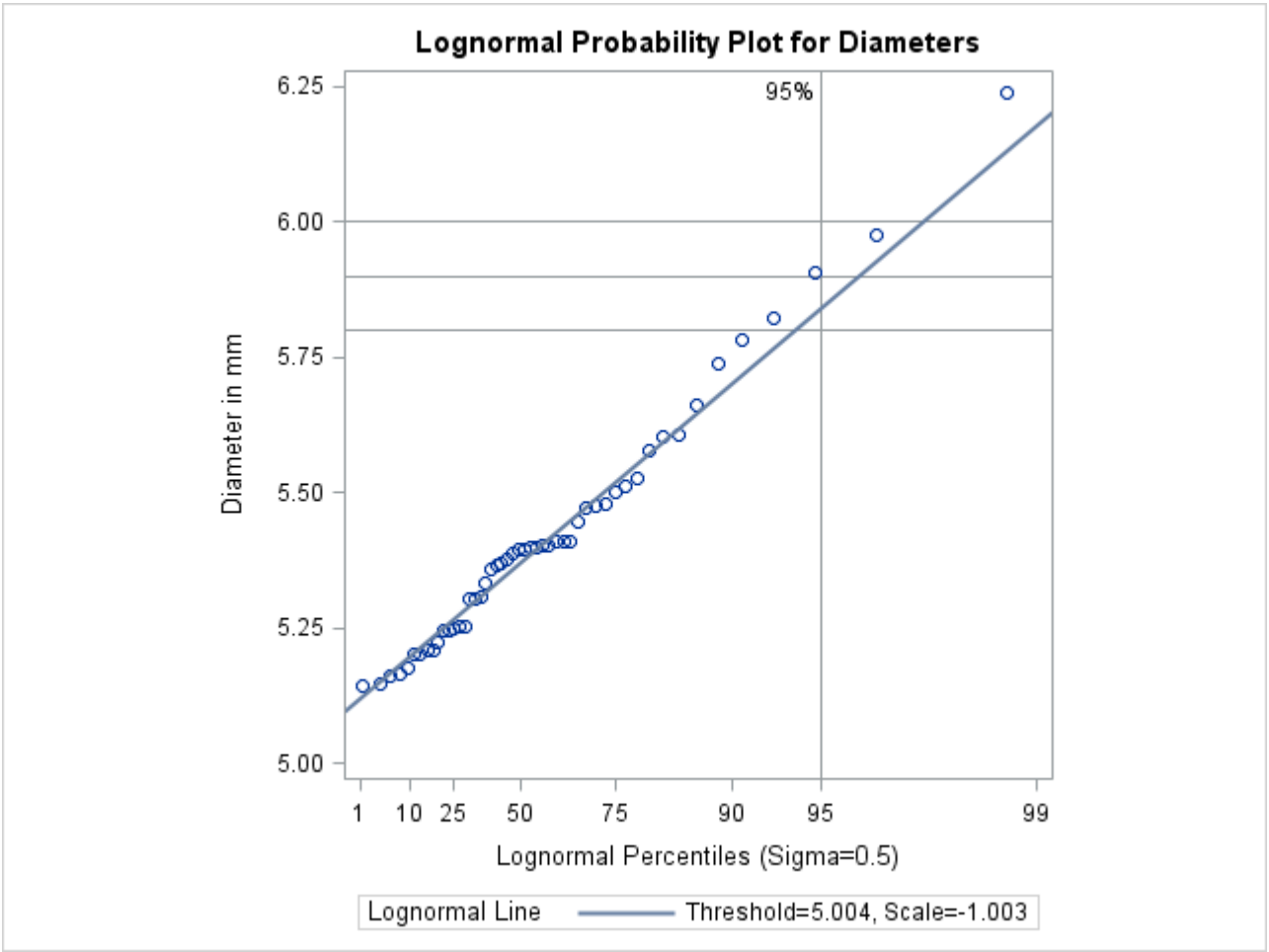
This example is a continuation of "Creating Lognormal Probability Plots" on page 463. Figure 6.36 shows that a lognormal distribution with shape parameter $\sigma = 0.5$ is a good fit for the distribution of Diameter in the data set Rods.

The lognormal distribution involves two other parameters: a threshold parameter θ and a scale parameter ζ. See Table 6.62 for the equation of the lognormal density function. The following statements illustrate how you can request a diagonal distribution reference line whose slope and intercept are determined by estimates of θ and ζ.

```
title 'Lognormal Probability Plot for Diameters';
proc capability data=Rods noprint;
   probplot Diameter / lognormal(sigma=0.5 theta=est zeta=est)
                       square
                       pctlminor
                       href      = 95
                       hreflabel = '95%'
                       vref      = 5.8 to 6.0 by 0.1
                       odstitle  = title;
run;
```

The plot is shown in Output 6.20.1.

Output 6.20.1 Lognormal Reference Line



The close agreement between the diagonal reference line and the point pattern indicates that the specific lognormal distribution with $\hat{\sigma} = 0.5$, $\hat{\theta} = 5.004$, and $\hat{\zeta} = -1.003$ is a good fit for the diameter measurements.

Specifying HREF=95 adds a reference line indicating the 95th percentile of the lognormal distribution. The HREFLABEL= option specifies a label for this line. The PCTLMINOR option displays minor tick marks on

the percentile axis. The VREF= option adds reference lines indicating diameter values of 5.8, 5.9, and 6.0, and the CHREF= and CVREF= options specify colors for the horizontal and vertical reference lines.

Based on the intersection of the diagonal reference line with the HREF= line, the estimated 95th percentile of the diameter distribution is 5.85 mm.

Note that you could also construct a similar plot in which all three parameters are estimated by substituting SIGMA=EST for SIGMA=0.5 in the preceding statements.

QQPLOT Statement: CAPABILITY Procedure

Overview: QQPLOT Statement

The QQPLOT statement creates a quantile-quantile plot (Q-Q plot), which compares ordered values of a variable with quantiles of a specified theoretical distribution such as the normal. If the data distribution matches the theoretical distribution, the points on the plot form a linear pattern. Thus, you can use a Q-Q plot to determine how well a theoretical distribution models a set of measurements.

You can specify one of the following theoretical distributions with the QQPLOT statement:

- beta
- exponential
- gamma
- Gumbel
- three-parameter lognormal
- normal
- generalized Pareto
- power function
- Rayleigh
- two-parameter Weibull
- three-parameter Weibull

You can use options in the QQPLOT statement to do the following:

- specify or estimate parameters for the theoretical distribution
- display a reference line corresponding to specific location and scale parameters for the theoretical distribution
- request graphical enhancements

You can also create a comparative Q-Q plot by using the QQPLOT statement in conjunction with a CLASS statement.

You have three alternatives for producing Q-Q plots with the QQPLOT statement:

- ODS Graphics output is produced if ODS Graphics is enabled, for example by specifying the ODS GRAPHICS ON statement prior to the PROC statement.
- Otherwise, traditional graphics are produced by default if SAS/GRAPH is licensed.
- Legacy line printer charts are produced when you specify the LINEPRINTER option in the PROC statement.

See Chapter 4, "SAS/QC Graphics," for more information about producing these different kinds of graphs.

NOTE: Q-Q plots are similar to probability plots, which you can create with the PROBPLOT statement (see "PROBPLOT Statement: CAPABILITY Procedure" on page 460). Q-Q plots are preferable for graphical estimation of distribution parameters and capability indices, whereas probability plots are preferable for graphical estimation of percentiles.

Getting Started: QQPLOT Statement

The following examples illustrate the basic syntax of the QQPLOT statement. For complete details of the QQPLOT statement, see the section "Syntax: QQPLOT Statement" on page 496. Advanced examples are provided on the section "Examples: QQPLOT Statement" on page 522.

Creating a Normal Quantile-Quantile Plot

NOTE: See *Creating Normal Q-Q Plots* in the SAS/QC Sample Library.

Measurements of the distance between two holes cut into 50 steel sheets are saved as values of the variable Distance in the following data set:

```
data Sheets;
   input Distance @@;
   label Distance='Hole Distance in cm';
   datalines;
 9.80 10.20 10.27  9.70  9.76
10.11 10.24 10.20 10.24  9.63
 9.99  9.78 10.10 10.21 10.00
 9.96  9.79 10.08  9.79 10.06
10.10  9.95  9.84 10.11  9.93
10.56 10.47  9.42 10.44 10.16
10.11 10.36  9.94  9.77  9.36
 9.89  9.62 10.05  9.72  9.82
 9.99 10.16 10.58 10.70  9.54
10.31 10.07 10.33  9.98 10.15
;
```

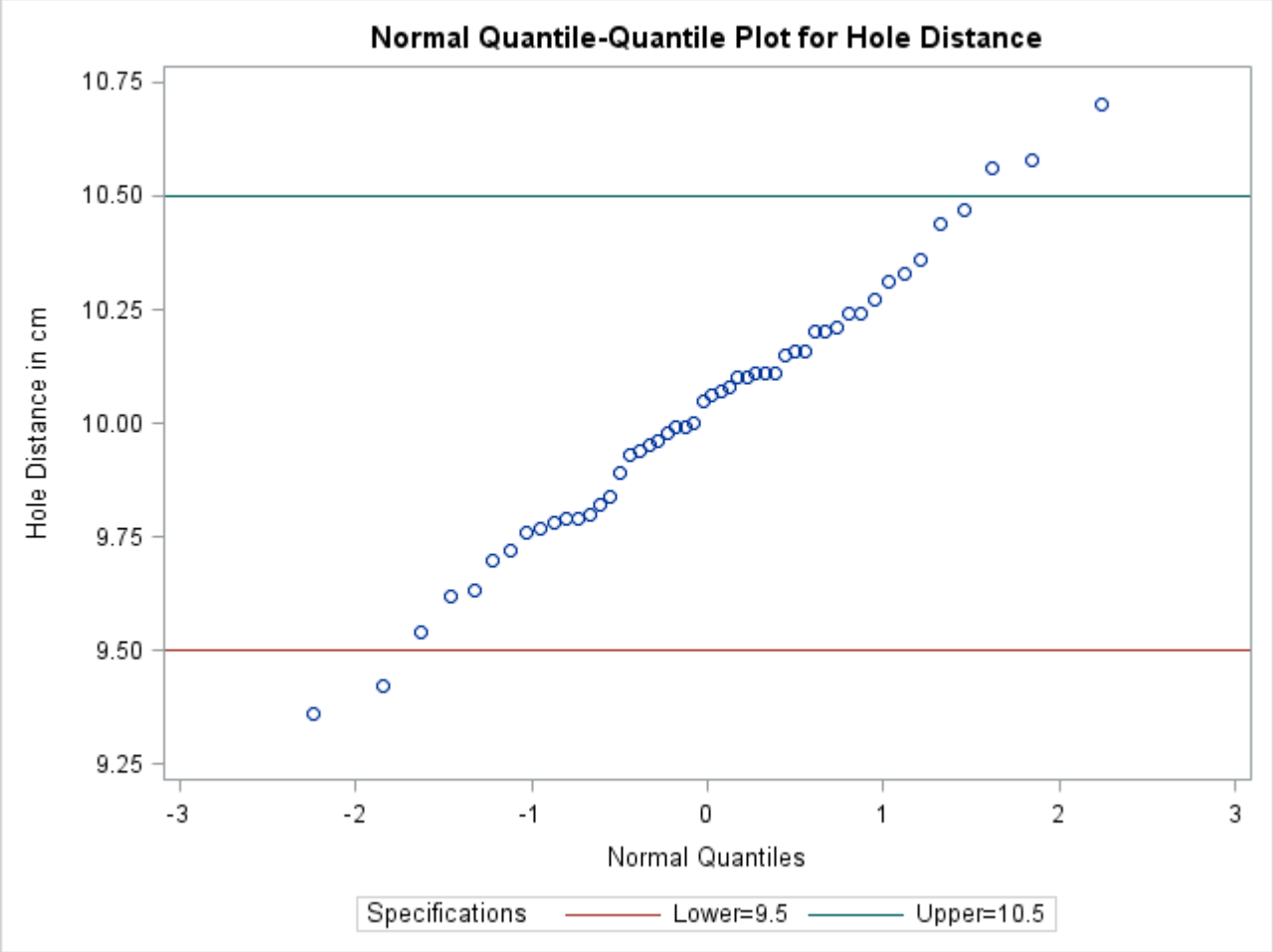
The cutting process is in control, and you decide to check whether the process distribution is normal. The following statements create a Q-Q plot for Distance, shown in Figure 6.39, with lower and upper specification lines at 9.5 cm and 10.5 cm:[6]

[6] For a P-P plot using these data, see Figure 6.31. For a probability plot using these data, see Example 6.20.

```
title 'Normal Quantile-Quantile Plot for Hole Distance';
proc capability data=Sheets noprint;
   spec lsl=9.5 usl=10.5;
   qqplot Distance / odstitle=title;
run;
```

The plot compares the ordered values of Distance with quantiles of the normal distribution. The linearity of the point pattern indicates that the measurements are normally distributed. Note that a normal Q-Q plot is created by default. The specification lines are requested with the LSL= and USL= options in the SPEC statement.

Figure 6.39 Normal Quantile-Quantile Plot Created with Traditional Graphics



Adding a Distribution Reference Line

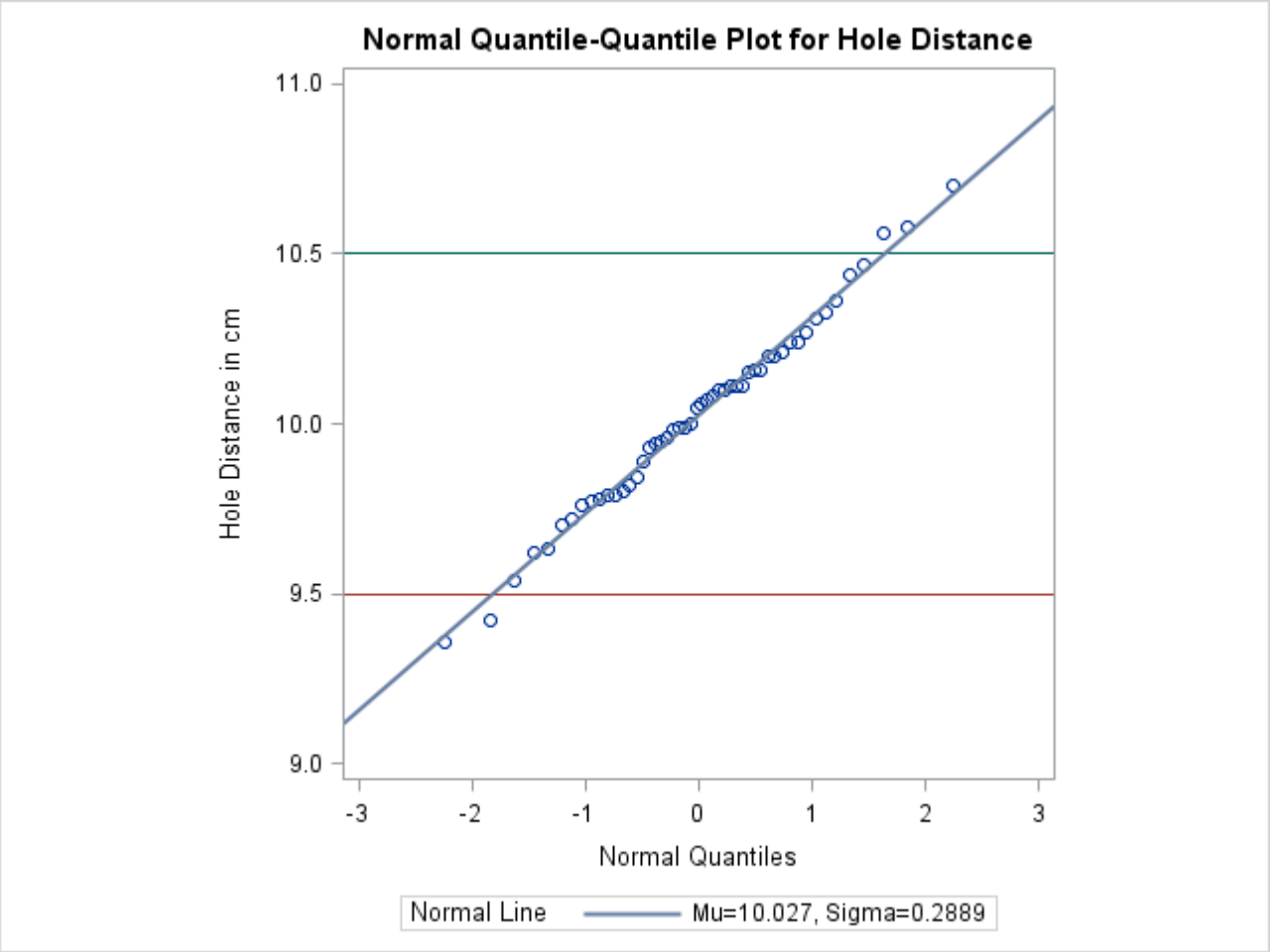
NOTE: See *Creating Normal Q-Q Plots* in the SAS/QC Sample Library.

In a normal Q-Q plot, the normal distribution with mean μ_0 and standard deviation σ_0 is represented by a reference line with intercept μ_0 and slope σ_0. The following statements reproduce the Q-Q plot in Figure 6.39, adding the line for which μ_0 and σ_0 are estimated by the sample mean and standard deviation:

```
title 'Normal Quantile-Quantile Plot for Hole Distance';
proc capability data=Sheets noprint;
   spec lsl=9.5 usl=10.5;
   qqplot Distance / normal(mu=est sigma=est)
                     square
                     nospeclegend
                     odstitle=title;
run;
```

The plot is displayed in Figure 6.40.

Figure 6.40 Adding a Distribution Reference Line to a Q-Q Plot



Specifying MU=EST and SIGMA=EST with the NORMAL option requests the reference line (alternatively, you can specify numeric values for μ_0 and σ_0 with the MU= and SIGMA= options). The COLOR= and L= options specify the color of the line and the line type. The SQUARE option displays the plot in a square format, and the NOSPECLEGEND option suppresses the legend for the specification lines.

Syntax: QQPLOT Statement

The syntax for the QQPLOT statement is as follows:

QQPLOT < *variables* > < / *options* > **;**

You can specify the keyword QQ as an alias for QQPLOT, and you can use any number of QQPLOT statements in the CAPABILITY procedure. The components of the QQPLOT statement are described as follows.

variables

are the process variables for which to create Q-Q plots. If you specify a VAR statement, the variables must also be listed in the VAR statement. Otherwise, the variables can be any numeric variables in the input data set. If you do not specify a list of variables, then by default the procedure creates a Q-Q plot for each variable listed in the VAR statement, or for each numeric variable in the DATA= data set if you do not specify a VAR statement. For example, each of the following QQPLOT statements produces two Q-Q plots, one for length and one for width:

```
proc capability data=measures;
   var length width;
   qqplot;
run;

proc capability data=measures;
   qqplot length width;
run;
```

options

specify the theoretical distribution for the plot or add features to the plot. If you specify more than one variable, the options apply equally to each variable. Specify all options after the slash (/) in the QQPLOT statement. You can specify only one option naming the distribution in each QQPLOT statement, but you can specify any number of other options. The distributions available are the beta, exponential, gamma, Gumbel, lognormal, normal, generalized Pareto, power function, Rayleigh, two-parameter Weibull, and three-parameter Weibull. By default, the procedure produces a plot for the normal distribution.

In the following example, the NORMAL option requests a normal Q-Q plot for each variable. The MU= and SIGMA= *normal-options* request a distribution reference line with intercept 10 and slope 0.3 for each plot, corresponding to a normal distribution with mean $\mu = 10$ and standard deviation $\sigma = 0.3$. The SQUARE option displays the plot in a square frame, and the CTEXT= option specifies the text color.

```
proc capability data=measures;
   qqplot length1 length2 / normal(mu=10 sigma=0.3)
                            square
                            ctext=blue;
run;
```

Summary of Options

The following tables list the QQPLOT statement options by function. For complete descriptions, see "Dictionary of Options" on page 501.

Distribution Options

Table 6.67 summarizes the options for requesting a specific theoretical distribution.

Table 6.67 Options for Specifying a Theoretical Distribution

Option	Description
BETA(*beta-options*)	specifies beta Q-Q plot for shape parameters α, β specified with mandatory ALPHA= and BETA= *beta-options*
EXPONENTIAL(*exponential-options*)	specifies exponential Q-Q plot
GAMMA(*gamma-options*)	specifies gamma Q-Q plot for shape parameter α specified with mandatory ALPHA= *gamma-option*
GUMBEL(*Gumbel-options*)	specifies Gumbel Q-Q plot
LOGNORMAL(*lognormal-options*)	specifies lognormal Q-Q plot for shape parameter σ specified with mandatory SIGMA= *lognormal-option*
NORMAL(*normal-options*)	specifies normal Q-Q plot
PARETO(*Pareto-options*)	specifies generalized Pareto Q-Q plot for shape parameter α specified with mandatory ALPHA= *Pareto-option*
POWER(*power-options*)	specifies power function Q-Q plot for shape parameter α specified with mandatory ALPHA= *power-option*
RAYLEIGH(*Rayleigh-options*)	specifies Rayleigh Q-Q plot
WEIBULL(*Weibull-options*)	specifies three-parameter Weibull Q-Q plot for shape parameter c specified with mandatory C= *Weibull-option*
WEIBULL2(*Weibull2-options*)	specifies two-parameter Weibull Q-Q plot

Table 6.68 summarizes options that specify parameter values for theoretical distributions and that control the display of a distribution reference line. Specify these options in parentheses after the distribution option. For example, the following statements use the NORMAL option to request a normal Q-Q plot with a specific distribution reference line. The MU= and SIGMA= *normal-options* display a distribution reference line with intercept 10 and slope 0.3. The COLOR= *normal-option* draws the line in red.

```
proc capability data=measures;
   qqplot length / normal(mu=10 sigma=0.3 color=red);
run;
```

Table 6.68 Distribution Options

Option	Description
Distribution Reference Line Options	
COLOR=	specifies color of distribution reference line
L=	specifies line type of distribution reference line
SYMBOL=	specifies plotting character for line printer plots
W=	specifies width of distribution reference line
Beta-Options	
ALPHA=	specifies mandatory shape parameter α
BETA=	specifies mandatory shape parameter β
SIGMA=	specifies reference line slope σ
THETA=	specifies reference line intercept θ
Exponential-Options	
SIGMA=	specifies reference line slope σ
THETA=	specifies reference line intercept θ
Gamma-Options	
ALPHA=	specifies mandatory shape parameter α
SIGMA=	specifies reference line slope σ
THETA=	specifies reference line intercept θ
Gumbel-Options	
MU=	specifies reference line intercept μ
SIGMA=	specifies reference line slope σ
Lognormal-Options	
SIGMA=	specifies mandatory shape parameter σ
SLOPE=	specifies reference line slope
THETA=	specifies reference line intercept θ
ZETA=	specifies reference line slope $\exp(\zeta_0)$
Normal-Options	
CPKREF	specifies vertical reference lines at intersection of specification limits with distribution reference line
CPKSCALE	rescales horizontal axis in C_{pk} units
MU=	specifies reference line intercept μ
SIGMA=	specifies reference line slope σ
Pareto-Options	
ALPHA=	specifies mandatory shape parameter α
SIGMA=	specifies reference line slope σ
THETA=	specifies reference line intercept θ
Power-Options	
ALPHA=	specifies mandatory shape parameter α
SIGMA=	specifies reference line slope σ
THETA=	specifies reference line intercept θ
Rayleigh-Options	
SIGMA=	specifies reference line slope σ
THETA=	specifies reference line intercept θ

Table 6.68 (*continued*)

Option	Description
Weibull-Options	
C=	specifies mandatory shape parameter c
SIGMA=	specifies reference line slope σ
THETA=	specifies reference line intercept θ
Weibull2-Options	
C=	specifies c_0 for reference line (slope is $\frac{1}{c_0}$)
SIGMA=	specifies σ_0 for reference line (intercept is $\log(\sigma_0)$)
SLOPE=	specifies reference line slope
THETA=	specifies known lower threshold θ_0

General Options

Table 6.69 lists options that control the appearance of the plots.

Table 6.69 General QQPLOT Statement Options

Option	Description
General Plot Layout Options	
CONTENTS=	specifies table of contents entry for Q-Q plot grouping
HREF=	specifies reference lines perpendicular to the horizontal axis
HREFLABELS=	specifies labels for HREF= lines
LEGEND=	specifies LEGEND statement
NADJ=	adjusts sample size (N) when computing quantiles
NOFRAME	suppresses frame around plotting area
NOLEGEND	suppresses legend
NOLINELEGEND	suppresses distribution reference line information in legend
NOSPECLEGEND	suppresses specifications information in legend
PCTLAXIS	adds a nonlinear percentile axis
PCTLMINOR	adds minor tick marks to percentile axis
PCTLSCALE	replaces theoretical quantiles with percentiles
RANKADJ=	adjusts ranks when computing quantiles
ROTATE	switches horizontal and vertical axes
SQUARE	displays Q-Q plot in square format
VREF=	specifies reference lines perpendicular to the vertical axis
VREFLABELS=	specifies labels for VREF= lines
Graphics Options	
ANNOTATE=	specifies annotate data set
CAXIS=	specifies color for axis
CFRAME=	specifies color for frame
CGRID=	specifies color for grid lines
CHREF=	specifies colors for HREF= lines
CSTATREF=	specifies colors for STATREF= lines
CTEXT=	specifies color for text

Table 6.69 (*continued*)

Option	Description
CVREF=	specifies colors for VREF= lines
DESCRIPTION=	specifies description for plot in graphics catalog
FONT=	specifies software font for text
GRID	draws grid lines perpendicular to the quantile axis
HEIGHT=	specifies height of text used outside framed areas
HMINOR=	specifies number of horizontal minor tick marks
HREFLABPOS=	specifies vertical position of labels for HREF= lines
INFONT=	specifies software font for text inside framed areas
INHEIGHT=	specifies height of text inside framed areas
LGRID=	specifies a line type for grid lines
LHREF=	specifies line styles for HREF= lines
LSTATREF=	specifies line styles for STATREF= lines
LVREF=	specifies line styles for VREF= lines
NAME=	specifies name for plot in graphics catalog
NOHLABEL	suppresses label for horizontal axis
NOVLABEL	suppresses label for vertical axis
NOVTICK	suppresses tick marks and tick mark labels for vertical axis
STATREF=	specifies reference lines at values of summary statistics
STATREFLABELS=	specifies labels for STATREF= lines
STATREFSUBCHAR=	specifies substitution character for displaying statistic values in STATREFLABELS= labels
VAXIS=	specifies AXIS statement for vertical axis
VAXISLABEL=	specifies label for vertical axis
VMINOR=	specifies number of vertical minor tick marks
VREFLABPOS=	specifies horizontal position of labels for VREF= lines
WAXIS=	specifies line thickness for axes and frame
WGRID=	specifies thickness for grid lines
Options for ODS Graphics Output	
ODSFOOTNOTE=	specifies footnote displayed on Q-Q plot
ODSFOOTNOTE2=	specifies secondary footnote displayed on Q-Q plot
ODSTITLE=	specifies title displayed on Q-Q plot
ODSTITLE2=	specifies secondary title displayed on Q-Q plot
Options for Comparative Plots	
ANNOKEY	applies annotation requested in ANNOTATE= data set to key cell only
CFRAMESIDE=	specifies color for filling frame for row labels
CFRAMETOP=	specifies color for filling frame for column labels
CPROP=	specifies color for proportion of frequency bar
CTEXTSIDE=	specifies color for row labels
CTEXTTOP=	specifies color for column labels
INTERTILE=	specifies distance between tiles
NCOLS=	specifies number of columns in comparative Q-Q plot
NROWS=	specifies number of rows in comparative Q-Q plot
OVERLAY	overlays plots for different class levels (ODS Graphics only)

Table 6.69 (*continued*)

Option	Description
Options to Enhance Line Printer Plots	
HREFCHAR=	specifies line character for HREF= lines
NOOBSLEGEND	suppresses legend for hidden points
QQSYMBOL=	specifies character for plotted points
VREFCHAR=	specifies character for VREF= lines

Dictionary of Options

The following sections provide detailed descriptions of options specific to the QQPLOT statement. See "Dictionary of Common Options: CAPABILITY Procedure" on page 533 for detailed descriptions of options common to all the plot statements.

General Options

You can specify the following options whether you are producing ODS Graphics output or traditional graphics:

ALPHA=*value-list*|**EST**

specifies values for a mandatory shape parameter α ($\alpha > 0$) for Q-Q plots requested with the BETA, GAMMA, PARETO, and POWER options. A plot is created for each value specified. For examples, see the entries for the distribution options. If you specify ALPHA=EST, a maximum likelihood estimate is computed for α.

BETA(ALPHA=*value-list*|**EST BETA=***value-list*|**EST <** *beta-options* **>)**

creates a beta Q-Q plot for each combination of the shape parameters α and β given by the mandatory ALPHA= and BETA= options. If you specify ALPHA=EST and BETA=EST, a plot is created based on maximum likelihood estimates for α and β. In the following example, the first QQPLOT statement produces one plot, the second statement produces four plots, the third statement produces six plots, and the fourth statement produces one plot:

```
proc capability data=measures;
   qqplot width / beta(alpha=2 beta=2);
   qqplot width / beta(alpha=2 3 beta=1 2);
   qqplot width / beta(alpha=2 to 3 beta=1 to 2 by 0.5);
   qqplot width / beta(alpha=est beta=est);
run;
```

To create the plot, the observations are ordered from smallest to largest, and the ith ordered observation is plotted against the quantile $B^{-1}_{\alpha\beta}\left(\frac{i-0.375}{n+0.25}\right)$, where $B^{-1}_{\alpha\beta}(\cdot)$ is the inverse normalized incomplete beta function, n is the number of nonmissing observations, and α and β are the shape parameters of the beta distribution.

The point pattern on the plot for ALPHA=α and BETA=β tends to be linear with intercept θ and slope σ if the data are beta distributed with the specific density function

$$p(x) = \begin{cases} \frac{(x-\theta)^{\alpha-1}(\theta+\sigma-x)^{\beta-1}}{B(\alpha,\beta)\sigma^{(\alpha+\beta-1)}} & \text{for } \theta < x < \theta+\sigma \\ 0 & \text{for } x \le \theta \text{ or } x \ge \theta+\sigma \end{cases}$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, and
θ = lower threshold parameter
σ = scale parameter ($\sigma > 0$)
α = first shape parameter ($\alpha > 0$)
β = second shape parameter ($\beta > 0$)

To obtain graphical estimates of α and β, specify lists of values for the ALPHA= and BETA= options, and select the combination of α and β that most nearly linearizes the point pattern. To assess the point pattern, you can add a diagonal distribution reference line with intercept θ_0 and slope σ_0 with the *beta-options* THETA=θ_0 and SIGMA=σ_0. Alternatively, you can add a line corresponding to estimated values of θ_0 and slope σ_0 with the *beta-options* THETA=EST and SIGMA=EST. Specify these options in parentheses, as in the following example:

```
proc capability data=measures;
   qqplot width / beta(alpha=2 beta=3 theta=4 sigma=5);
run;
```

Agreement between the reference line and the point pattern indicates that the beta distribution with parameters α, β, θ_0, and σ_0 is a good fit. You can specify the SCALE= option as an alias for the SIGMA= option and the THRESHOLD= option as an alias for the THETA= option.

BETA=*value-list*|**EST**
specifies values for the shape parameter β ($\beta > 0$) for Q-Q plots requested with the BETA distribution option. A plot is created for each value specified with the BETA= option. If you specify BETA=EST, a maximum likelihood estimate is computed for β. For examples, see the preceding entry for the BETA distribution option.

C=*value(-list)*|**EST**
specifies the shape parameter c ($c > 0$) for Q-Q plots requested with the WEIBULL and WEIBULL2 options. You must specify C= as a *Weibull-option* with the WEIBULL option; in this situation it accepts a list of values, or if you specify C=EST, a maximum likelihood estimate is computed for c. You can optionally specify C=*value* or C=EST as a *Weibull2-option* with the WEIBULL2 option to request a distribution reference line; in this situation, you must also specify SIGMA=*value* or SIGMA=EST. For an example, see Output 6.23.1.

CPKSCALE
rescales the quantile axis in C_{pk} units for plots requested with the NORMAL option. Specify CPKSCALE in parentheses after the NORMAL option. You can use the CPKSCALE option with the CPKREF option for graphical estimation of the capability indices *CPU*, *CPL*, and C_{pk}, as illustrated in Output 6.24.1.

EXPONENTIAL(< (*exponential-options*) **>**
EXP< (*exponential-options*) **>)**
creates an exponential Q-Q plot. To create the plot, the observations are ordered from smallest to largest, and the *i*th ordered observation is plotted against the quantile $-\log\left(1 - \frac{i-0.375}{n+0.25}\right)$, where n is the number of nonmissing observations.

The pattern on the plot tends to be linear with intercept θ and slope σ if the data are exponentially distributed with the specific density function

$$p(x) = \begin{cases} \frac{1}{\sigma} \exp\left(-\frac{x-\theta}{\sigma}\right) & \text{for } x \geq \theta \\ 0 & \text{for } x < \theta \end{cases}$$

where θ is the threshold parameter, and σ is the scale parameter ($\sigma > 0$).

To assess the point pattern, you can add a diagonal distribution reference line with intercept θ_0 and slope σ_0 with the *exponential-options* THETA=θ_0 and SIGMA=σ_0. Alternatively, you can add a line corresponding to estimated values of θ_0 and slope σ_0 with the *exponential-options* THETA=EST and SIGMA=EST. Specify these options in parentheses, as in the following example: as in the following example:

```
proc capability data=measures;
   qqplot width / exponential(theta=4 sigma=5);
run;
```

Agreement between the reference line and the point pattern indicates that the exponential distribution with parameters θ_0 and σ_0 is a good fit. You can specify the SCALE= option as an alias for the SIGMA= option and the THRESHOLD= option as an alias for the THETA= option.

GAMMA(ALPHA=*value-list***|EST <** *gamma-options* **>)**

creates a gamma Q-Q plot for each value of the shape parameter α given by the mandatory ALPHA= option or its alias, the SHAPE= option. The following example produces three probability plots:

```
proc capability data=measures;
   qqplot width / gamma(alpha=0.4 to 0.6 by 0.1);
run;
```

To create the plot, the observations are ordered from smallest to largest, and the *i*th ordered observation is plotted against the quantile $G_\alpha^{-1}\left(\frac{i-0.375}{n+0.25}\right)$, where $G_\alpha^{-1}(\cdot)$ is the inverse normalized incomplete gamma function, n is the number of nonmissing observations, and α is the shape parameter of the gamma distribution.

The pattern on the plot for ALPHA=α tends to be linear with intercept θ and slope σ if the data are gamma distributed with the specific density function

$$p(x) = \begin{cases} \frac{1}{\sigma\Gamma(\alpha)} \left(\frac{x-\theta}{\sigma}\right)^{\alpha-1} \exp\left(-\frac{x-\theta}{\sigma}\right) & \text{for } x > \theta \\ 0 & \text{for } x \leq \theta \end{cases}$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)
α = shape parameter ($\alpha > 0$)

To obtain a graphical estimate of α, specify a list of values for the ALPHA= option, and select the value that most nearly linearizes the point pattern.

To assess the point pattern, you can add a diagonal distribution reference line with intercept θ_0 and slope σ_0 with the *gamma-options* THETA=θ_0 and SIGMA=σ_0. Alternatively, you can add a line corresponding to estimated values of θ_0 and σ_0 with the *gamma-options* THETA=EST and SIGMA=EST. Specify these options in parentheses, as in the following example:

```
proc capability data=measures;
   qqplot width / gamma(alpha=2 theta=3 sigma=4);
run;
```

Agreement between the reference line and the point pattern indicates that the gamma distribution with parameters α, θ_0, and σ_0 is a good fit. You can specify the SCALE= option as an alias for the SIGMA= option and the THRESHOLD= option as an alias for the THETA= option.

GUMBEL(< *Gumbel-options* **>)**

creates a Gumbel Q-Q plot. To create the plot, the observations are ordered from smallest to largest, and the ith ordered observation is plotted against the quantile $-\log\left(-\log\left(\frac{i-0.375}{n+0.25}\right)\right)$, where n is the number of nonmissing observations.

The point pattern on the plot tends to be linear with intercept μ and slope σ if the data are Gumbel distributed with the specific density function

$$p(x) = \frac{e^{-(x-\mu)/\sigma}}{\sigma} \exp\left(-e^{-(x-\mu)/\sigma}\right)$$

where μ is a location parameter and σ is a positive scale parameter.

To assess the point pattern, you can add a diagonal distribution reference line corresponding to μ_0 and σ_0 with the *Gumbel-options* MU=μ_0 and SIGMA=σ_0. Alternatively, you can add a line corresponding to estimated values of μ_0 and σ_0 with the *Gumbel-options* MU=EST and SIGMA=EST. Specify these options in parentheses following the GUMBEL option.

Agreement between the reference line and the point pattern indicates that the Gumbel distribution with parameters μ_0 and σ_0 is a good fit.

GRID

draws reference lines perpendicular to the quantile axis at major tick marks.

LEGEND=*name* **| NONE**

specifies the name of a LEGEND statement describing the legend for specification limit reference lines and fitted curves. Specifying LEGEND=NONE is equivalent to specifying the NOLEGEND option.

LOGNORMAL(SIGMA=*value-list***|EST <** *lognormal-options* **>)**

LNORM(SIGMA=*value-list***|EST <** *lognormal-options* **>)**

creates a lognormal Q-Q plot for each value of the shape parameter σ given by the mandatory SIGMA= option or its alias, the SHAPE= option. For example,

```
proc capability data=measures;
   qqplot width/ lognormal(shape=1.5 2.5);
run;
```

To create the plot, the observations are ordered from smallest to largest, and the *i*th ordered observation is plotted against the quantile $\exp\left(\sigma\Phi^{-1}\left(\frac{i-0.375}{n+0.25}\right)\right)$, where $\Phi^{-1}(\cdot)$ is the inverse cumulative standard normal distribution, n is the number of nonmissing observations, and σ is the shape parameter of the lognormal distribution.

The pattern on the plot for SIGMA=σ tends to be linear with intercept θ and slope $\exp(\zeta)$ if the data are lognormally distributed with the specific density function

$$p(x) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}(x-\theta)} \exp\left(-\frac{(\log(x-\theta)-\zeta)^2}{2\sigma^2}\right) & \text{for } x > \theta \\ 0 & \text{for } x \leq \theta \end{cases}$$

where
θ = threshold parameter
ζ = scale parameter
σ = shape parameter ($\sigma > 0$)

To obtain a graphical estimate of σ, specify a list of values for the SIGMA= option, and select the value that most nearly linearizes the point pattern. For an illustration, see Example 6.22.

To assess the point pattern, you can add a diagonal distribution reference line corresponding to the threshold parameter θ_0 and the scale parameter ζ_0 with the *lognormal-options* THETA=θ_0 and ZETA=ζ_0. Alternatively, you can add a line corresponding to estimated values of θ_0 and ζ_0 with the *lognormal-options* THETA=EST and ZETA=EST. This line has intercept θ_0 and slope $\exp(\zeta_0)$. Agreement between the line and the point pattern indicates that the lognormal distribution with parameters σ, θ_0, and ζ_0 is a good fit. See Output 6.22.4 for an example. You can specify the THRESHOLD= option as an alias for the THETA= option and the SCALE= option as an alias for the ZETA= option.

You can also display the reference line by specifying THETA=θ_0, and you can specify the slope with the SLOPE= option. For example, the following two QQPLOT statements produce charts with identical reference lines:

```
proc capability data=measures;
   qqplot width / lognormal(sigma=2 theta=3 zeta=1);
   qqplot width / lognormal(sigma=2 theta=3 slope=2.718);
run;
```

MU=*value*|**EST**

specifies a value for the mean μ for a Q-Q plot requested with the GUMBEL and NORMAL options. For the normal distribution, you can specify MU=EST to request a distribution reference line with intercept equal to the sample mean, as illustrated in Figure 6.40. If you specify MU=EST for the Gumbel distribution, a maximum likelihood estimate is calculated.

NADJ=*value*

specifies the adjustment value added to the sample size in the calculation of theoretical quantiles. The default is $\frac{1}{4}$, as described by Blom (1958). Also refer to Chambers et al. (1983) for additional information.

NOLEGEND

LEGEND=NONE

suppresses legends for specification limits, fitted curves, distribution lines, and hidden observations. For an example, see Output 6.24.1.

NOLINELEGEND

NOLINEL

suppresses the legend for the optional distribution reference line.

NORMAL<(*normal-options*)>

NORM<(*normal-options*)>

creates a normal Q-Q plot. This is the default if you do not specify a distribution option. To create the plot, the observations are ordered from smallest to largest, and the ith ordered observation is plotted against the quantile $\Phi^{-1}\left(\frac{i-0.375}{n+0.25}\right)$, where $\Phi^{-1}(\cdot)$ is the inverse cumulative standard normal distribution, and n is the number of nonmissing observations.

The pattern on the plot tends to be linear with intercept μ and slope σ if the data are normally distributed with the specific density function

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad \text{for all } x$$

where μ is the mean, and σ is the standard deviation ($\sigma > 0$).

To assess the point pattern, you can add a diagonal distribution reference line with intercept μ_0 and slope σ_0 with the *normal-options* MU=μ_0 and SIGMA=σ_0. Alternatively, you can add a line corresponding to estimated values of μ_0 and σ_0 with the *normal-options* MU=EST and SIGMA=EST; the estimates of μ_0 and σ_0 are the sample mean and sample standard deviation. Specify these options in parentheses, as in the following example:

```
proc capability data=measures;
   qqplot length / normal(mu=10 sigma=0.3);
run;
```

For an example, see "Adding a Distribution Reference Line" on page 494. Agreement between the reference line and the point pattern indicates that the normal distribution with parameters μ_0 and σ_0 is a good fit. You can specify MU=EST and SIGMA=EST to request a distribution reference line with the sample mean and sample standard deviation as the intercept and slope.

Other *normal-options* include CPKREF and CPKSCALE. The CPKREF option draws reference lines extending from the intersections of specification limits with the distribution reference line to the theoretical quantile axis. The CPKSCALE option rescales the theoretical quantile axis in C_{pk} units. You can use the CPKREF option with the CPKSCALE option for graphical estimation of the capability indices *CPU*, *CPL*, and C_{pk}, as illustrated in Output 6.24.1.

NOSPECLEGEND
NOSPECL

suppresses the legend for specification limit reference lines. For an example, see Figure 6.40.

PARETO(< *Pareto-options* **>)**

creates a generalized Pareto Q-Q plot for each value of the shape parameter α given by the mandatory ALPHA= option. If you specify ALPHA=EST, a plot is created based on a maximum likelihood estimate for α.

To create the plot, the observations are ordered from smallest to largest, and the ith ordered observation is plotted against the quantile $(1-(1-\frac{i-0.375}{n+0.25})^{\alpha})/\alpha$ $(\alpha \neq 0)$ or $-\log(1-\frac{i-0.375}{n+0.25})$ $(\alpha = 0)$, where n is the number of nonmissing observations and α is the shape parameter of the generalized Pareto distribution.

The point pattern on the plot for ALPHA=α tends to be linear with intercept θ and slope σ if the data are generalized Pareto distributed with the specific density function

$$p(x) = \begin{cases} \frac{1}{\sigma}(1-\alpha(x-\theta)/\sigma)^{1/\alpha-1} & \text{if } \alpha \neq 0 \\ \frac{1}{\sigma}\exp(-(x-\theta)/\sigma) & \text{if } \alpha = 0 \end{cases}$$

where

θ = threshold parameter
σ = scale parameter $(\sigma > 0)$
α = shape parameter $(\alpha > 0)$

To obtain a graphical estimate of α, specify a list of values for the ALPHA= option, and select the value that most nearly linearizes the point pattern.

To assess the point pattern, you can add a diagonal distribution reference line corresponding to θ_0 and σ_0 with the *Pareto-options* THETA=θ_0 and SIGMA=σ_0. Alternatively, you can add a line corresponding to estimated values of θ_0 and σ_0 with the *Pareto-options* THETA=EST and SIGMA=EST. Specify these options in parentheses following the PARETO option.

Agreement between the reference line and the point pattern indicates that the generalized Pareto distribution with parameters α, θ_0, and σ_0 is a good fit.

PCTLAXIS(*axis-options***)**

adds a nonlinear percentile axis along the frame of the Q-Q plot opposite the theoretical quantile axis. The added axis is identical to the axis for probability plots produced with the PROBPLOT statement. When using the PCTLAXIS option, you must specify HREF= values in quantile units, and you cannot use the NOFRAME option. You can specify the following *axis-options*:

CGRID=*color*

specifies the color used for grid lines.

GRID

draws grid lines perpendicular to the percentile axis at major tick marks.

GRIDCHAR=‘*character***’**

specifies the character used to draw grid lines associated with the percentile axis on line printer plots.

LABEL='*string*'
specifies the label for the percentile axis.

LGRID=*linetype*
specifies the line type used for grid lines associated with the percentile axis.

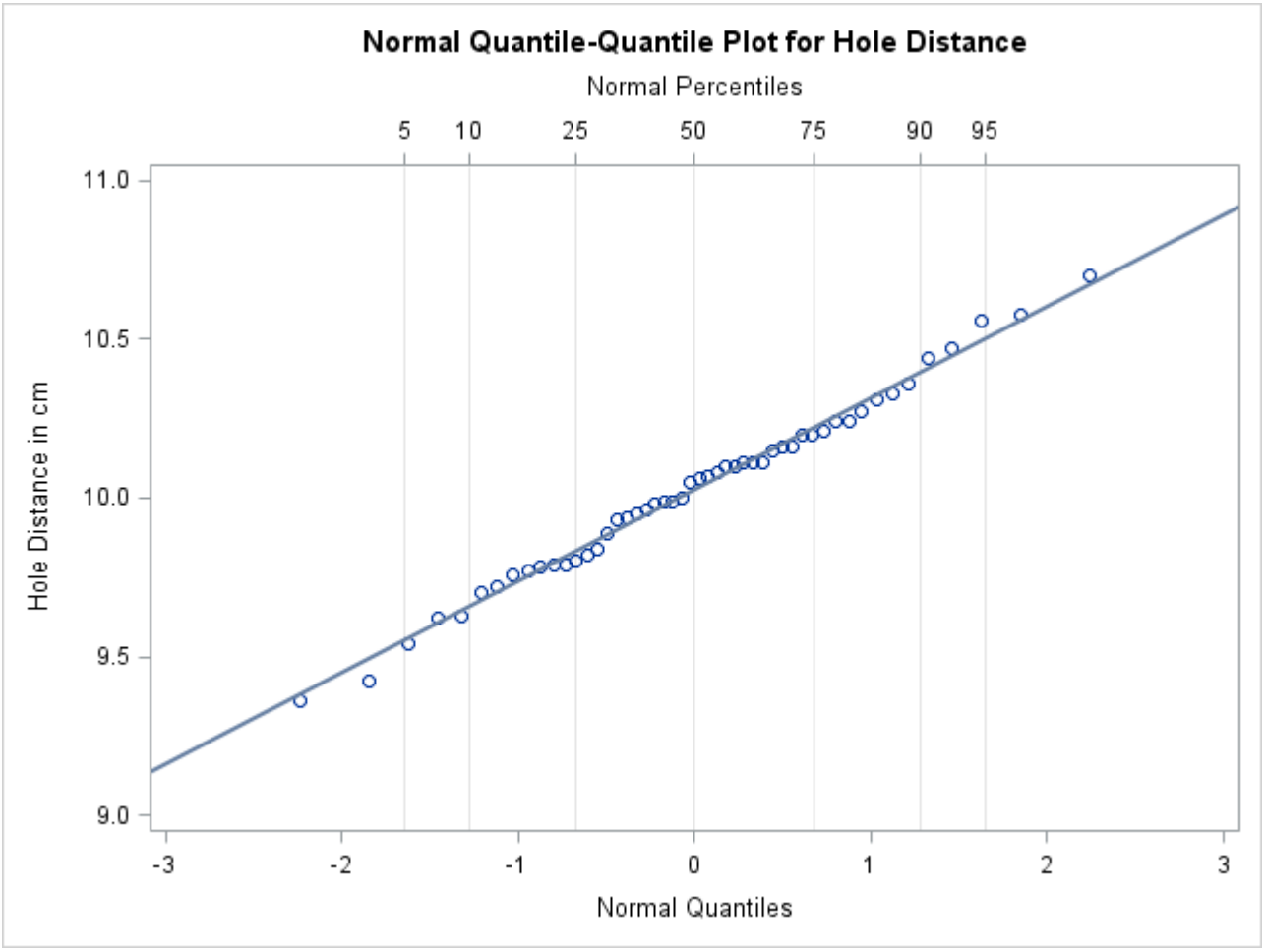
WGRID=*value*
specifies the thickness for grid lines associated with the percentile axis.

NOTE: See *Creating Normal Q-Q Plots* in the SAS/QC Sample Library.

For example, the following statements display the plot in Figure 6.41:

```
title 'Normal Quantile-Quantile Plot for Hole Distance';
proc capability data=Sheets noprint;
   qqplot Distance / normal(mu=est sigma=est)
                     nolegend
                     pctlaxis(grid label='Normal Percentiles')
                     odstitle=title;
run;
```

Figure 6.41 Normal Q-Q Plot with Percentile Axis



PCTLSCALE

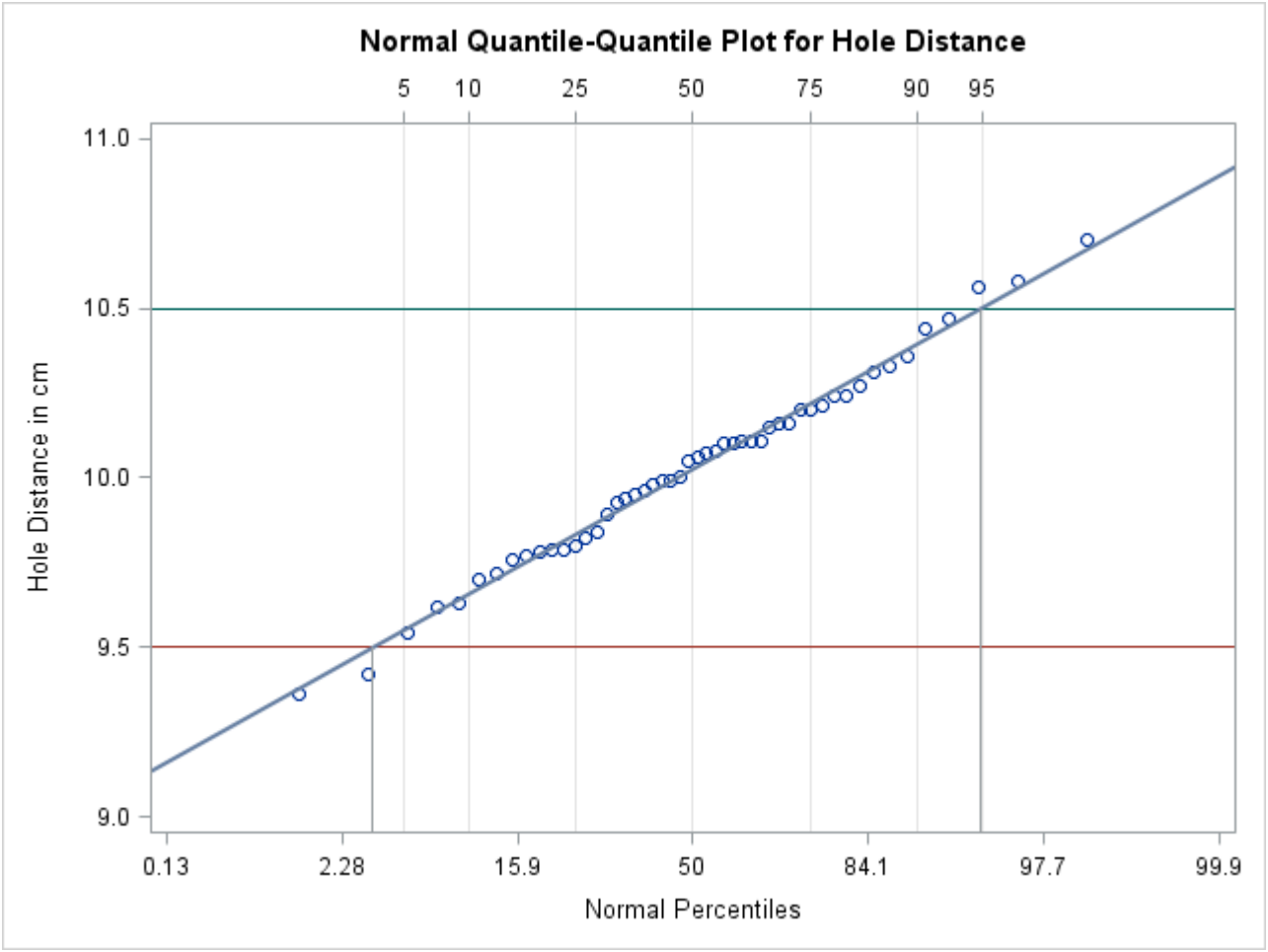
requests scale labels for the theoretical quantile axis in percentile units, resulting in a nonlinear axis scale. Tick marks are drawn uniformly across the axis based on the quantile scale. In all other respects, the plot remains the same, and you must specify HREF= values in quantile units. For a true nonlinear axis, use the PCTLAXIS option or use the PROBPLOT statement.

NOTE: See *Creating Normal Q-Q Plots* in the SAS/QC Sample Library.

For example, the following statements display the plot in Figure 6.42:

```
title 'Normal Quantile-Quantile Plot for Hole Distance';
proc capability data=Sheets noprint;
   spec lsl=9.5 usl=10.5;
   qqplot Distance / normal(mu=est sigma=est cpkref)
                     pctlaxis(grid lgrid=35)
                     nolegend pctlscale
                     odstitle=title;
run;
```

Figure 6.42 Normal Q-Q Plot for Reading Percentiles of Specification Limits



POWER(< *power-options* >)

creates a power function Q-Q plot for each value of the shape parameter α given by the mandatory ALPHA= option. If you specify ALPHA=EST, a plot is created based on a maximum likelihood estimate for α.

To create the plot, the observations are ordered from smallest to largest, and the *i*th ordered observation is plotted against the quantile $B^{-1}_{\alpha(1)}\left(\frac{i-0.375}{n+0.25}\right)$, where $B^{-1}_{\alpha(1)}(\cdot)$ is the inverse normalized incomplete beta function, n is the number of nonmissing observations, α is one shape parameter of the beta distribution, and the second shape parameter, $\beta = 1$.

The point pattern on the plot for ALPHA=α tends to be linear with intercept θ and slope σ if the data are power function distributed with the specific density function

$$p(x) = \begin{cases} \frac{\alpha}{\sigma}\left(\frac{x-\theta}{\sigma}\right)^{\alpha-1} & \text{for } \theta < x < \theta + \sigma \\ 0 & \text{for } x \le \theta \text{ or } x \ge \theta + \sigma \end{cases}$$

where

θ = threshold parameter
σ = scale parameter ($\sigma > 0$)
α = shape parameter ($\alpha > 0$)

To obtain a graphical estimate of α, specify a list of values for the ALPHA= option, and select the value that most nearly linearizes the point pattern.

To assess the point pattern, you can add a diagonal distribution reference line corresponding to θ_0 and σ_0 with the *power-options* THETA=θ_0 and SIGMA=σ_0. Alternatively, you can add a line corresponding to estimated values of θ_0 and σ_0 with the *power-options* THETA=EST and SIGMA=EST. Specify these options in parentheses following the POWER option.

Agreement between the reference line and the point pattern indicates that the power function distribution with parameters α, θ_0, and σ_0 is a good fit.

RANKADJ=*value*

specifies the adjustment value added to the ranks in the calculation of theoretical quantiles. The default is $-\frac{3}{8}$, as described by Blom (1958). Also refer to Chambers et al. (1983) for additional information.

RAYLEIGH(< *Rayleigh-options* >)

creates a Rayleigh Q-Q plot. To create the plot, the observations are ordered from smallest to largest, and the *i*th ordered observation is plotted against the quantile $\sqrt{-2\log\left(1-\frac{i-0.375}{n+0.25}\right)}$, where n is the number of nonmissing observations.

The point pattern on the plot tends to be linear with intercept θ and slope σ if the data are Rayleigh distributed with the specific density function

$$p(x) = \begin{cases} \frac{x-\theta}{\sigma^2}\exp(-(x-\theta)^2/(2\sigma^2)) & \text{for } x \ge \theta \\ 0 & \text{for } x < \theta \end{cases}$$

where θ is a threshold parameter, and σ is a positive scale parameter.

To assess the point pattern, you can add a diagonal distribution reference line corresponding to θ_0 and σ_0 with the *Rayleigh-options* THETA=θ_0 and SIGMA=σ_0. Alternatively, you can add a line corresponding to estimated values of θ_0 and σ_0 with the *Rayleigh-options* THETA=EST and SIGMA=EST. Specify these options in parentheses after the RAYLEIGH option.

Agreement between the reference line and the point pattern indicates that the Rayleigh distribution with parameters θ_0 and σ_0 is a good fit.

ROTATE

switches the horizontal and vertical axes so that the theoretical percentiles are plotted vertically while the data are plotted horizontally. Regardless of whether the plot has been rotated, horizontal axis options (such as HAXIS=) refer to the horizontal axis, and vertical axis options (such as VAXIS=) refer to the vertical axis. All other options that depend on axis placement adjust to the rotated axes.

SIGMA=*value-list*|**EST**

specifies the value of the distribution parameter σ, where $\sigma > 0$. Alternatively, you can specify SIGMA=EST to request a maximum likelihood estimate for σ_0. The use of the SIGMA= option depends on the distribution option specified, as indicated by the following table:

Distribution Option	Use of the SIGMA= Option
BETA EXPONENTIAL GAMMA PARETO POWER RAYLEIGH WEIBULL	THETA=θ_0 and SIGMA=σ_0 request a distribution reference line with intercept θ_0 and slope σ_0.
GUMBEL	MU=μ_0 and SIGMA=σ_0 request a distribution reference line corresponding to μ_0 and σ_0.
LOGNORMAL	SIGMA=$\sigma_1 \ldots \sigma_n$ requests n Q-Q plots with shape parameters $\sigma_1 \ldots \sigma_n$. The SIGMA= option is mandatory.
NORMAL	MU=μ_0 and SIGMA=σ_0 request a distribution reference line with intercept μ_0 and slope σ_0. SIGMA=EST requests a slope equal to the sample standard deviation.
WEIBULL2	SIGMA=σ_0 and C=c_0 request a distribution reference line with intercept $\log(\sigma_0)$ and slope $\frac{1}{c_0}$.

For an example using SIGMA=EST, see Output 6.24.1. For an example of lognormal plots using the SIGMA= option, see Example 6.22.

SLOPE=*value*|**EST**

specifies the slope for a distribution reference line requested with the LOGNORMAL and WEIBULL2 options.

When you use the SLOPE= option with the LOGNORMAL option, you must also specify a threshold parameter value θ_0 with the THETA= option. Specifying the SLOPE= option is an alternative to specifying ZETA=ζ_0, which requests a slope of $\exp(\zeta_0)$. See Output 6.22.4 for an example.

When you use the SLOPE= option with the WEIBULL2 option, you must also specify a scale parameter value σ_0 with the SIGMA= option. Specifying the SLOPE= option is an alternative to specifying C=c_0, which requests a slope of $\frac{1}{c_0}$.

For example, the first and second QQPLOT statements that follow produce plots identical to those produced by the third and fourth QQPLOT statements:

```
proc capability data=measures;
   qqplot width  / lognormal(sigma=2 theta=0 zeta=0);
   qqplot width  / weibull2(sigma=2 theta=0 c=0.25);
   qqplot width  / lognormal(sigma=2 theta=0 slope=1);
   qqplot width  / weibull2(sigma=2 theta=0 slope=4);
run;
```

For more information, see "Graphical Estimation" on page 517.

SQUARE

displays the Q-Q plot in a square frame. Compare Figure 6.39 with Figure 6.40. The default is a rectangular frame.

THETA=*value*|**EST**

THRESHOLD=*value*|**EST**

specifies the lower threshold parameter θ for Q-Q plots requested with the BETA, EXPONENTIAL, GAMMA, LOGNORMAL, PARETO, POWER, RAYLEIGH, WEIBULL, and WEIBULL2 options.

When used with the WEIBULL2 option, the THETA= option specifies the known lower threshold θ_0, for which the default is 0. See Output 6.23.2 for an example.

When used with the other distribution options, the THETA= option specifies θ_0 for a distribution reference line; alternatively in this situation, you can specify THETA=EST to request a maximum likelihood estimate for θ_0. To request the line, you must also specify a scale parameter. See Output 6.22.4 for an example of the THETA= option with a lognormal Q-Q plot.

WEIBULL(C=*value-list*|**EST <** *Weibull-options* **>)**

WEIB(C=*value-list* **<** *Weibull-options* **>)**

creates a three-parameter Weibull Q-Q plot for each value of the shape parameter c given by the mandatory C= option or its alias, the SHAPE= option. For example,

```
proc capability data=measures;
   qqplot width / weibull(c=1.8 to 2.4 by 0.2);
run;
```

To create the plot, the observations are ordered from smallest to largest, and the ith ordered observation is plotted against the quantile $\left(-\log\left(1-\frac{i-0.375}{n+0.25}\right)\right)^{\frac{1}{c}}$, where n is the number of nonmissing observations, and c is the Weibull distribution shape parameter.

The pattern on the plot for C=c tends to be linear with intercept θ and slope σ if the data are Weibull distributed with the specific density function

$$p(x) = \begin{cases} \frac{c}{\sigma}\left(\frac{x-\theta}{\sigma}\right)^{c-1}\exp\left(-\left(\frac{x-\theta}{\sigma}\right)^{c}\right) & \text{for } x > \theta \\ 0 & \text{for } x \le \theta \end{cases}$$

where θ is the threshold parameter, σ is the scale parameter ($\sigma > 0$), and c is the shape parameter ($c > 0$).

To obtain a graphical estimate of c, specify a list of values for the C= option, and select the value that most nearly linearizes the point pattern. For an illustration, see Example 6.23. To assess the point pattern, you can add a diagonal distribution reference line with intercept θ_0 and slope σ_0 with the *Weibull-options* THETA=θ_0 and SIGMA=σ_0. Alternatively, you can add a line corresponding to estimated values of θ_0 and σ_0 with the *Weibull-options* THETA=EST and SIGMA=EST. Specify these options in parentheses, as in the following example:

```
proc capability data=measures;
   qqplot width / weibull(c=2 theta=3 sigma=4);
run;
```

Agreement between the reference line and the point pattern indicates that the Weibull distribution with parameters c, θ_0, and σ_0 is a good fit. You can specify the SCALE= option as an alias for the SIGMA= option and the THRESHOLD= option as an alias for the THETA= option.

WEIBULL2<(*Weibull2-options***)>**

W2<(*Weibull2-options***)>**

creates a two-parameter Weibull Q-Q plot. You should use the WEIBULL2 option when your data have a *known* lower threshold θ_0. You can specify the threshold value θ_0 with the THETA= option or its alias, the THRESHOLD= option. If you are uncertain of the lower threshold value, you can estimate θ_0 graphically by specifying a list of values for the THETA= option. Select the value that most linearizes the point pattern. The default is $\theta_0 = 0$.

To create the plot, the observations are ordered from smallest to largest, and the log of the shifted ith ordered observation $x_{(i)}$, $\log(x_{(i)} - \theta_0)$, is plotted against the quantile $\log\left(-\log\left(1 - \frac{i-0.375}{n+0.25}\right)\right)$, where n is the number of nonmissing observations. Unlike the three-parameter Weibull quantile, the preceding expression is free of distribution parameters. This is why the C= shape parameter option is not mandatory with the WEIBULL2 option.

The pattern on the plot for THETA=θ_0 tends to be linear with intercept $\log(\sigma)$ and slope $\frac{1}{c}$ if the data are Weibull distributed with the specific density function

$$p(x) = \begin{cases} \frac{c}{\sigma}\left(\frac{x-\theta_0}{\sigma}\right)^{c-1} \exp\left(-\left(\frac{x-\theta_0}{\sigma}\right)^{c}\right) & \text{for } x > \theta_0 \\ 0 & \text{for } x \le \theta_0 \end{cases}$$

where θ_0 is a known lower threshold parameter, σ is a scale parameter ($\sigma > 0$), and c is a shape parameter ($c > 0$).

The advantage of a two-parameter Weibull plot over a three-parameter Weibull plot is that you can visually estimate the shape parameter c and the scale parameter σ from the slope and intercept of the point pattern; see Example 6.23 for an illustration of this method. The disadvantage is that the two-parameter Weibull distribution applies only in situations where the threshold parameter is known. See "Graphical Estimation" on page 517 for more information.

To assess the point pattern, you can add a diagonal distribution reference line corresponding to the scale parameter σ_0 and shape parameter c_0 with the *Weibull2-options* SIGMA=σ_0 and C=c_0. Alternatively, you can add a distribution reference line corresponding to estimated values of σ_0 and c_0 with the

Weibull2-options SIGMA=EST and C=EST. This line has intercept $\log(\sigma_0)$ and slope $\frac{1}{c_0}$. Agreement between the line and the point pattern indicates that the Weibull distribution with parameters c_0, θ_0, and σ_0 is a good fit. You can specify the SCALE= option as an alias for the SIGMA= option and the SHAPE= option as an alias for the C= option.

You can also display the reference line by specifying SIGMA=σ_0, and you can specify the slope with the SLOPE= option. For example, the following QQPLOT statements produce identical plots:

```
proc capability data=measures;
   qqplot width / weibull2(theta=3 sigma=4 c=2);
   qqplot width / weibull2(theta=3 sigma=4 slope=0.5);
run;
```

ZETA=*value*|**EST**
specifies a value for the scale parameter ζ for lognormal Q-Q plots requested with the LOGNORMAL option. Specify THETA=θ_0 and ZETA=ζ_0 to request a distribution reference line with intercept θ_0 and slope $\exp(\zeta_0)$.

Options for Traditional Graphics

You can specify the following options if you are producing traditional graphics:

CGRID=*color*
specifies the color for the grid lines associated with the quantile axis, requested by the GRID option.

LGRID=*linetype*
specifies the line type for the grid lines associated with the quantile axis, requested by the GRID option.

CPKREF
draws reference lines extending from the intersections of the specification limits with the distribution reference line to the quantile axis in plots requested with the NORMAL option. Specify CPKREF in parentheses after the NORMAL option. You can use the CPKREF option with the CPKSCALE option for graphical estimation of the capability indices *CPU*, *CPL*, and C_{pk}, as illustrated in Output 6.24.1.

PCTLMINOR
requests minor tick marks for the percentile axis displayed when you use the PCTLAXIS option. See the entry for the PCTLAXIS option for an example.

WGRID=*n*
specifies the width of the grid lines associated with the quantile axis, requested with the GRID option. If you use the WGRID= option, you do not need to specify the GRID option.

Options for Legacy Line Printer Plots

You can specify the following options if you are producing legacy line printer plots:

NOOBSLEGEND

NOOBSL
suppresses the legend that indicates the number of hidden observations.

QQSYMBOL=‘*character*’
specifies the character used to plot the Q-Q points in line printer plots. The default is the plus sign (+).

SYMBOL=‘*character*’
specifies the character used for a distribution reference line in a line printer plot. The default character is the first letter of the distribution option keyword.

Details: QQPLOT Statement

This section provides details on the following topics:

- construction of Q-Q plots
- interpretation of Q-Q plots
- distributions supported by the QQPLOT statement
- graphical estimation of shape parameters, location and scale parameters, theoretical percentiles, and capability indices
- SYMBOL statement options

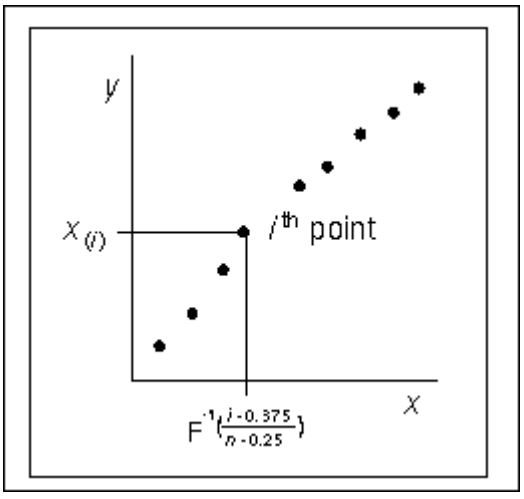
Construction of Quantile-Quantile and Probability Plots

Figure 6.43 illustrates how a Q-Q plot is constructed. First, the n nonmissing values of the variable are ordered from smallest to largest:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

Then the ith ordered value $x_{(i)}$ is represented on the plot by a point whose y-coordinate is $x_{(i)}$ and whose x-coordinate is $F^{-1}\left(\frac{i-0.375}{n+0.25}\right)$, where $F(\cdot)$ is the theoretical distribution with zero location parameter and unit scale parameter.

Figure 6.43 Construction of a Q-Q Plot



You can modify the adjustment constants –0.375 and 0.25 with the RANKADJ= and NADJ= options. This default combination is recommended by Blom (1958). For additional information, refer to Chambers et al. (1983). Because $x_{(i)}$ is a quantile of the empirical cumulative distribution function (ecdf), a Q-Q plot compares quantiles of the ecdf with quantiles of a theoretical distribution. Probability plots (see "PROBPLOT Statement: CAPABILITY Procedure" on page 460) are constructed the same way, except that the x-axis is scaled nonlinearly in percentiles.

Interpretation of Quantile-Quantile and Probability Plots

The following properties of Q-Q plots and probability plots make them useful diagnostics of how well a specified theoretical distribution fits a set of measurements:

- If the quantiles of the theoretical and data distributions agree, the plotted points fall on or near the line $y = x$.
- If the theoretical and data distributions differ only in their location or scale, the points on the plot fall on or near the line $y = ax + b$. The slope a and intercept b are visual estimates of the scale and location parameters of the theoretical distribution.

Q-Q plots are more convenient than probability plots for graphical estimation of the location and scale parameters because the x-axis of a Q-Q plot is scaled linearly. On the other hand, probability plots are more convenient for estimating percentiles or probabilities.

There are many reasons why the point pattern in a Q-Q plot may not be linear. Chambers et al. (1983) and Fowlkes (1987) discuss the interpretations of commonly encountered departures from linearity, and these are summarized in the following table.

Table 6.70 Quantile-Quantile Plot Diagnostics

Description of Point Pattern	Possible Interpretation
All but a few points fall on a line	Outliers in the data
Left end of pattern is below the line; right end of pattern is above the line	Long tails at both ends of the data distribution
Left end of pattern is above the line; right end of pattern is below the line	Short tails at both ends of the data distribution
Curved pattern with slope increasing from left to right	Data distribution is skewed to the right
Curved pattern with slope decreasing from left to right	Data distribution is skewed to the left
Staircase pattern (plateaus and gaps)	Data have been rounded or are discrete

In some applications, a nonlinear pattern may be more revealing than a linear pattern. However, Chambers et al. (1983) note that departures from linearity can also be due to chance variation.

Summary of Theoretical Distributions

You can use the QQPLOT statement to request Q-Q plots based on the theoretical distributions summarized in Table 6.71.

Table 6.71 QQPLOT Statement Distribution Options

Distribution	Density Function $p(x)$	Range	Parameters		
			Location	Scale	Shape
Beta	$\frac{(x-\theta)^{\alpha-1}(\theta+\sigma-x)^{\beta-1}}{B(\alpha,\beta)\sigma^{(\alpha+\beta-1)}}$	$\theta < x < \theta+\sigma$	θ	σ	α, β
Exponential	$\frac{1}{\sigma}\exp\left(-\frac{x-\theta}{\sigma}\right)$	$x \geq \theta$	θ	σ	
Gamma	$\frac{1}{\sigma\Gamma(\alpha)}\left(\frac{x-\theta}{\sigma}\right)^{\alpha-1}\exp\left(-\frac{x-\theta}{\sigma}\right)$	$x > \theta$	θ	σ	α
Gumbel	$\frac{e^{-(x-\mu)/\sigma}}{\sigma}\exp\left(-e^{-(x-\mu)/\sigma}\right)$	all x	μ	σ	
Lognormal (3-parameter)	$\frac{1}{\sigma\sqrt{2\pi}(x-\theta)}\exp\left(-\frac{(\log(x-\theta)-\zeta)^2}{2\sigma^2}\right)$	$x > \theta$	θ	ζ	σ
Normal	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	all x	μ	σ	
Generalized Pareto	$\alpha \neq 0$ $\frac{1}{\sigma}(1-\alpha(x-\theta)/\sigma)^{1/\alpha-1}$ $\alpha = 0$ $\frac{1}{\sigma}\exp(-(x-\theta)/\sigma)$	$x > \theta$	θ	σ	α
Power Function	$\frac{\alpha}{\sigma}\left(\frac{x-\theta}{\sigma}\right)^{\alpha-1}$	$x > \theta$	θ	σ	α
Rayleigh	$\frac{x-\theta}{\sigma^2}\exp(-(x-\theta)^2/(2\sigma^2))$	$x \geq \theta$	θ	σ	
Weibull (3-parameter)	$\frac{c}{\sigma}\left(\frac{x-\theta}{\sigma}\right)^{c-1}\exp\left(-\left(\frac{x-\theta}{\sigma}\right)^{c}\right)$	$x > \theta$	θ	σ	c
Weibull (2-parameter)	$\frac{c}{\sigma}\left(\frac{x-\theta_0}{\sigma}\right)^{c-1}\exp\left(-\left(\frac{x-\theta_0}{\sigma}\right)^{c}\right)$	$x > \theta_0$ (known)	θ_0	σ	c

You can request these distributions with the BETA, EXPONENTIAL, GAMMA, LOGNORMAL, NORMAL, WEIBULL, and WEIBULL2 options, respectively. If you do not specify a distribution option, a normal Q-Q plot is created.

Graphical Estimation

You can use Q-Q plots to estimate shape, location, and scale parameters and to estimate percentiles. If you are working with a normal Q-Q plot, you can also estimate certain capability indices.

Shape Parameters

Some distribution options in the QQPLOT statement require that you specify one or two shape parameters in parentheses after the distribution keyword. These are summarized in Table 6.72.

You can visually estimate a shape parameter by specifying a list of values for the shape parameter option. A separate plot is displayed for each value, and you can then select the value that linearizes the point pattern. Alternatively, you can request that the plot be created using an estimated shape parameter. See the entries for the distribution options in the section "Dictionary of Options" on page 501. for details on specification of shape parameters. Example 6.22 and Example 6.23 illustrate shape parameter estimation with lognormal and Weibull Q-Q plots.

Note that for Q-Q plots requested with the WEIBULL2 option, you can estimate the shape parameter c from a linear pattern using the fact that the slope of the pattern is $\frac{1}{c}$. For an illustration, see Example 6.23.

Table 6.72 Shape Parameter Options for the QQPLOT Statement

Distribution Keyword	Mandatory Shape Parameter Option	Range
BETA	ALPHA=α, BETA=β	$\alpha > 0, \beta > 0$
EXPONENTIAL	None	
GAMMA	ALPHA=α	$\alpha > 0$
GUMBEL	None	
LOGNORMAL	SIGMA=σ	$\sigma > 0$
NORMAL	None	
PARETO	ALPHA=α	$\alpha > 0$
POWER	ALPHA=α	$\alpha > 0$
RAYLEIGH	None	
WEIBULL	C=c	$a > 0$
WEIBULL2	None	

Location and Scale Parameters

When the point pattern on a Q-Q plot is linear, its intercept and slope provide estimates of the location and scale parameters. (An exception to this rule is the two-parameter Weibull distribution, for which the intercept and slope are related to the scale and shape parameters.) Table 6.73 shows how the intercept and slope are related to the parameters for each distribution supported by the QQPLOT statement.

Table 6.73 Intercept and Slope of Linear Q-Q Plots

Distribution	Parameters			Linear Pattern	
	Location	Scale	Shape	Intercept	Slope
Beta	θ	σ	α , β	θ	σ
Exponential	θ	σ		θ	σ
Gamma	θ	σ	α	θ	σ
Gumbel	μ	σ		μ	σ
Lognormal	θ	ζ	σ	θ	$\exp(\zeta)$
Normal	μ	σ		μ	σ
Generalized Pareto	θ	σ	α	θ	σ
Power Function	θ	σ	α	θ	σ
Rayleigh	θ	σ		θ	σ
Weibull (3-parameter)	θ	σ	c	θ	σ
Weibull (2-parameter)	θ_0 (known)	σ	c	$\log(\sigma)$	$\frac{1}{c}$

You can enhance a Q-Q plot with a diagonal *distribution reference line* by specifying the parameters that determine the slope and intercept of the line; alternatively, you can request estimates for these parameters. This line is an aid to checking the linearity of the point pattern, and it facilitates parameter estimation. For instance, specifying MU=3 and SIGMA=2 with the NORMAL option requests a line with intercept 3 and slope 2. Specifying SIGMA=1 and C=2 with the WEIBULL2 option requests a line with intercept $\log(1) = 0$ and slope $\frac{1}{2}$.

With the LOGNORMAL and WEIBULL2 options, you can specify the slope directly with the SLOPE= option. That is, for the LOGNORMAL option, specifying THETA=θ_0 and SLOPE=$\exp(\zeta_0)$ gives the same reference line as specifying THETA=θ_0 and ZETA=ζ_0. For the WEIBULL2 option, specifying SIGMA=σ_0 and SLOPE=$\frac{1}{c_0}$ gives the same reference line as specifying SIGMA=σ_0 and C=c_0.

For an example of parameter estimation using a normal Q-Q plot, see "Adding a Distribution Reference Line" on page 494. Example 6.22 illustrates parameter estimation using a lognormal plot, and Example 6.23 illustrates estimation using two-parameter and three-parameter Weibull plots.

Theoretical Percentiles

There are two ways to estimate percentiles from a Q-Q plot:

- Specify the PCTLAXIS option, which adds a percentile axis opposite the theoretical quantile axis. The scale for the percentile axis ranges between 0 and 100 with tick marks at percentile values such as 1, 5, 10, 25, 50, 75, 90, 95, and 99. See Figure 6.41 for an example.
- Specify the PCTLSCALE option, which relabels the horizontal axis tick marks with their percentile equivalents but does not alter their spacing. For example, on a normal Q-Q plot, the tick mark labeled "0" is relabeled as "50" because the 50th percentile corresponds to the zero quantile. See Figure 6.42 for an example.

You can also estimate percentiles using probability plots created with the PROBPLOT statement. See Output 6.20.1 for an example.

Capability Indices

When the point pattern on a normal Q-Q plot is linear, you can estimate the capability indices CPU, CPL, and C_{pk} from the plot, as explained by Rodriguez (1992). This method exploits the fact that the horizontal axis of a Q-Q plot indicates the distance in standard deviation units (multiple of σ) between a measurement or specification limit and the process average.

In particular, one-third the standardized distance between an upper specification limit and the mean is the one-sided capability index *CPU*.

$$CPU = \frac{USL - \mu}{3\sigma}$$

Likewise, one-third the standardized distance between a lower specification limit and the mean is the one-sided capability index *CPL*.

$$CPL = \frac{\mu - LSL}{3\sigma}$$

Consequently, if you *rescale* the quantile axis of a normal Q-Q plot by a factor of three, you can read *CPU* and *CPL* from the horizontal coordinates of the points at which the upper and lower specification lines intersect the point pattern. Because C_{pk} is defined as the minimum of *CPU* and *CPL*, this method also provides a graphical estimate of C_{pk}. For an illustration, see Example 6.24.

SYMBOL Statement Options

In earlier releases of SAS/QC software, graphical features of lower and upper specification lines and diagonal distribution reference lines were controlled with options in the SYMBOL2, SYMBOL3, and SYMBOL4 statements, respectively. These options are still supported, although they have been superseded by options in the QQPLOT and SPEC statements. Table 6.74 summarizes the two sets of options. **NOTE:** These statements have no effect on ODS Graphics output.

Table 6.74 SYMBOL Statement Options

Feature	Statement and Options	Alternative Statement and Options
Symbol markers	**SYMBOL1 Statement**	
character	VALUE=*special-symbol*	
color	COLOR=*color*	
font	FONT=*font*	
height	HEIGHT=*value*	
Lower specification line	**SPEC Statement**	**SYMBOL2 Statement**
position	LSL=*value*	
color	CLSL=*color*	COLOR=*color*
line type	LLSL=*linetype*	LINE=*linetype*
width	WLSL=*value*	WIDTH=*value*
Upper specification line	**SPEC Statement**	**SYMBOL3 Statement**
position	USL=*value*	
color	CUSL=*color*	COLOR=*color*
line type	LUSL=*linetype*	LINE=*linetype*
width	WUSL=*value*	WIDTH=*value*
Target reference line	**SPEC Statement**	
position	TARGET=*value*	
color	CTARGET=*color*	
line type	LTARGET=*linetype*	
width	WTARGET=*value*	
Distribution reference line	**QQPLOT Statement**	**SYMBOL4 Statement**
color	COLOR=*color*	COLOR=*color*
line type	LINE=*linetype*	LINE=*linetype*
width	WIDTH=*value*	WIDTH=*value*

ODS Graphics

Before you create ODS Graphics output, ODS Graphics must be enabled (for example, by using the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" (Chapter 21, *SAS/STAT User's Guide*).

The appearance of a graph produced with ODS Graphics is determined by the style associated with the ODS destination where the graph is produced. QQPLOT options used to control the appearance of traditional graphics are ignored for ODS Graphics output.

When ODS Graphics is in effect, the QQPLOT statement assigns a name to the graph it creates. You can use this name to reference the graph when using ODS. The name is listed in Table 6.75.

Table 6.75 ODS Graphics Produced by the QQPLOT Statement

ODS Graph Name	Plot Description
QQPlot	Q-Q plot

See Chapter 4, "SAS/QC Graphics," for more information about ODS Graphics and other methods for producing charts.

Examples: QQPLOT Statement

This section provides advanced examples of the QQPLOT statement.

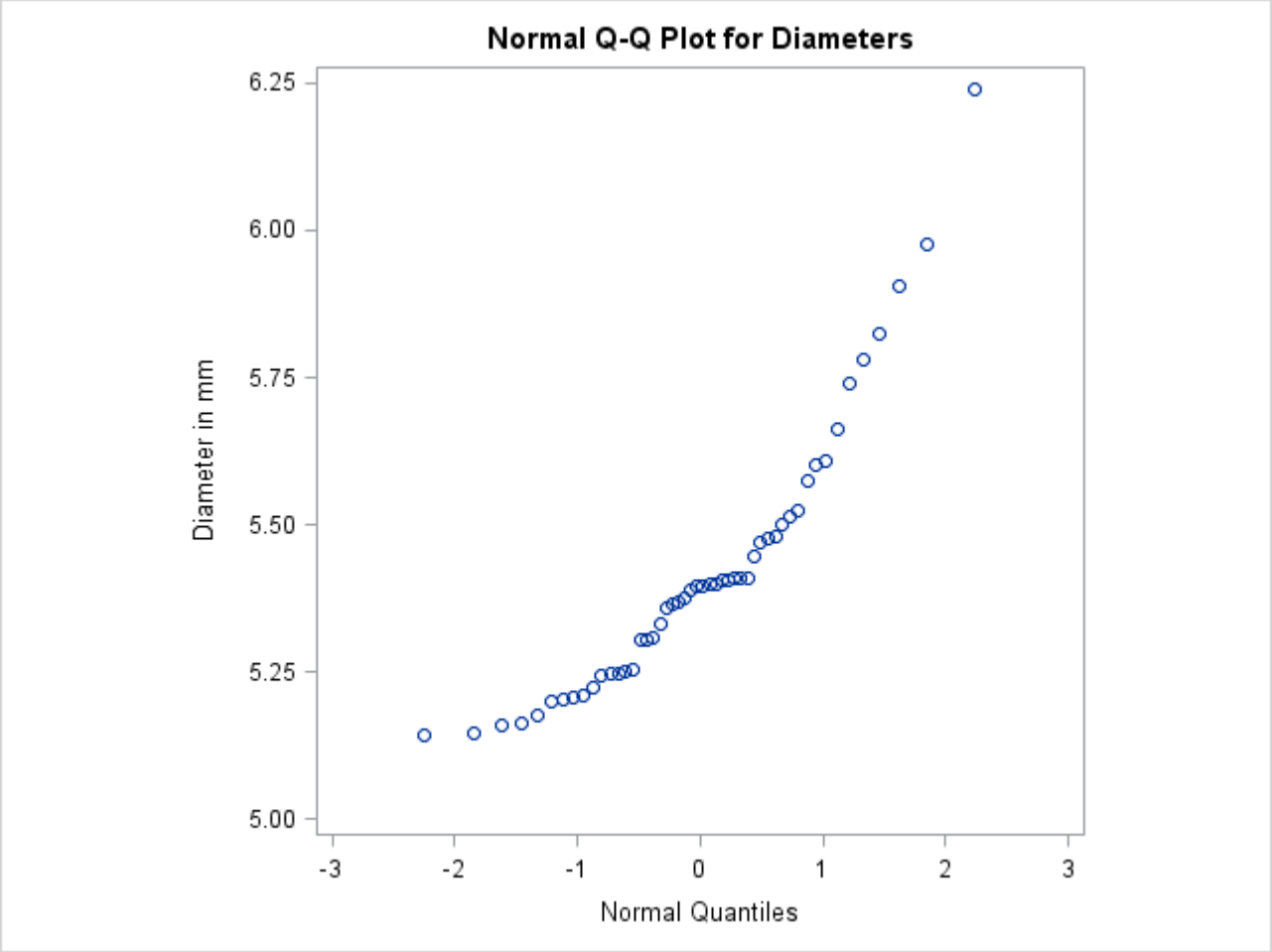
Example 6.21: Interpreting a Normal Q-Q Plot of Nonnormal Data

NOTE: See *Creating Lognormal Q-Q Plots* in the SAS/QC Sample Library.

The following statements produce the normal Q-Q plot in Output 6.21.1:

```
data Measures;
   input Diameter @@;
   label Diameter='Diameter in mm';
   datalines;
5.501  5.251  5.404  5.366  5.445  5.576  5.607
5.200  5.977  5.177  5.332  5.399  5.661  5.512
5.252  5.404  5.739  5.525  5.160  5.410  5.823
5.376  5.202  5.470  5.410  5.394  5.146  5.244
5.309  5.480  5.388  5.399  5.360  5.368  5.394
5.248  5.409  5.304  6.239  5.781  5.247  5.907
5.208  5.143  5.304  5.603  5.164  5.209  5.475
5.223
 ;

title 'Normal Q-Q Plot for Diameters';
proc capability data=Measures noprint;
   qqplot Diameter / normal square odstitle=title;
run;
```

Output 6.21.1 Normal Quantile-Quantile Plot of Nonnormal Data



The nonlinearity of the points in Output 6.21.1 indicates a departure from normality. Because the point pattern is curved with slope increasing from left to right, a theoretical distribution that is skewed to the right, such as a lognormal distribution, should provide a better fit than the normal distribution. The mild curvature suggests that you should examine the data with a series of lognormal Q-Q plots for small values of the shape parameter, as illustrated in the next example.

Example 6.22: Estimating Parameters from Lognormal Plots

This example, which is a continuation of Example 6.21, demonstrates techniques for estimating the shape parameter, location and scale parameters, and theoretical percentiles for a lognormal distribution.

Three-Parameter Lognormal Plots

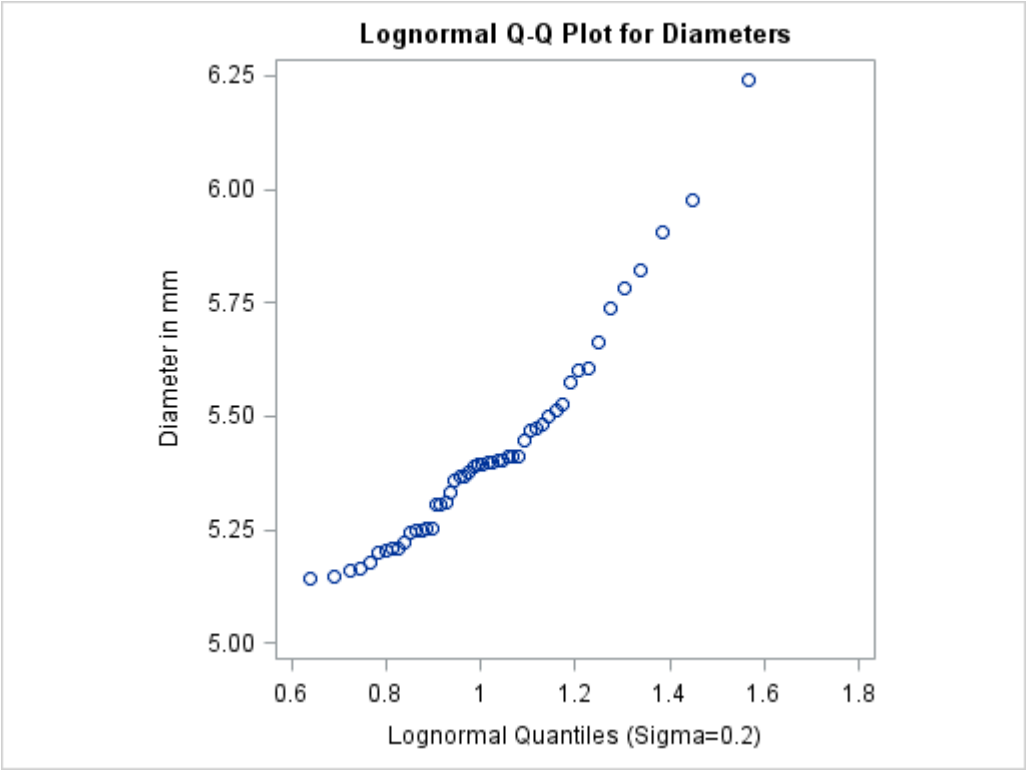
NOTE: See *Creating Lognormal Q-Q Plots* in the SAS/QC Sample Library.

The three-parameter lognormal distribution depends on a threshold parameter θ, a scale parameter ζ, and a shape parameter σ. You can estimate σ from a series of lognormal Q-Q plots with different values of σ. The estimate is the value of σ that linearizes the point pattern. You can then estimate the threshold and scale

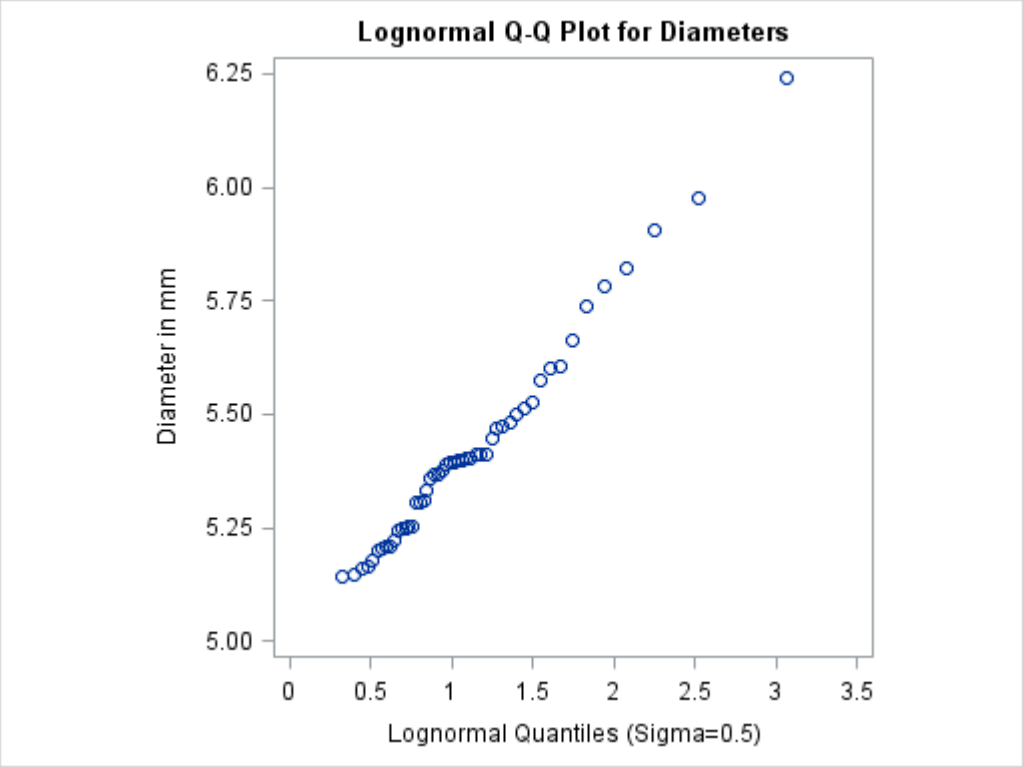
parameters from the intercept and slope of the point pattern. The following statements create the series of plots in Output 6.22.1 through Output 6.22.3 for σ values of 0.2, 0.5, and 0.8:

```
title 'Lognormal Q-Q Plot for Diameters';
proc capability data=Measures noprint;
   qqplot Diameter / lognormal(sigma=0.2 0.5 0.8)
                     square
                     odstitle=title;
run;
```

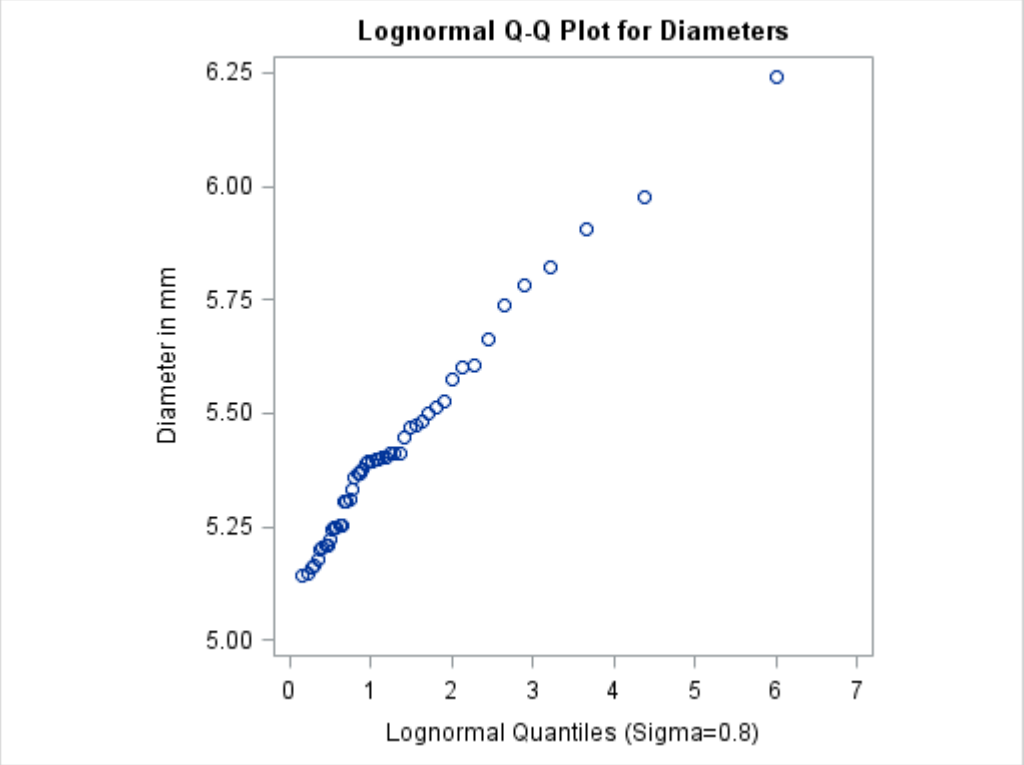
Output 6.22.1 Lognormal Quantile-Quantile Plot (σ =0.2)



Output 6.22.2 Lognormal Quantile-Quantile Plot (σ =0.5)



Output 6.22.3 Lognormal Quantile-Quantile Plot (σ =0.8)



NOTE: You must specify a value for the shape parameter σ for a lognormal Q-Q plot with the SIGMA= option or its alias, the SHAPE= option.

The plot in Output 6.22.2 displays the most linear point pattern, indicating that the lognormal distribution with $\sigma = 0.5$ provides a reasonable fit for the data distribution.

Data with this particular lognormal distribution have the density function

$$p(x) = \begin{cases} \frac{\sqrt{2}}{\sqrt{\pi}(x-\theta)} \exp\left(-2(\log(x-\theta)-\zeta)^2\right) & \text{for } x > \theta \\ 0 & \text{for } x \leq \theta \end{cases}$$

The points in the plot fall on or near the line with intercept θ and slope $\exp(\zeta)$. Based on Output 6.22.2, $\theta \approx 5$ and $\exp(\zeta) \approx \frac{1.2}{3} = 0.4$, giving $\zeta \approx \log(0.4) \approx -0.92$.

Estimating Percentiles

NOTE: See *Creating Lognormal Q-Q Plots* in the SAS/QC Sample Library.

You can use a Q-Q plot to estimate percentiles such as the 95th percentile of the lognormal distribution.[7]

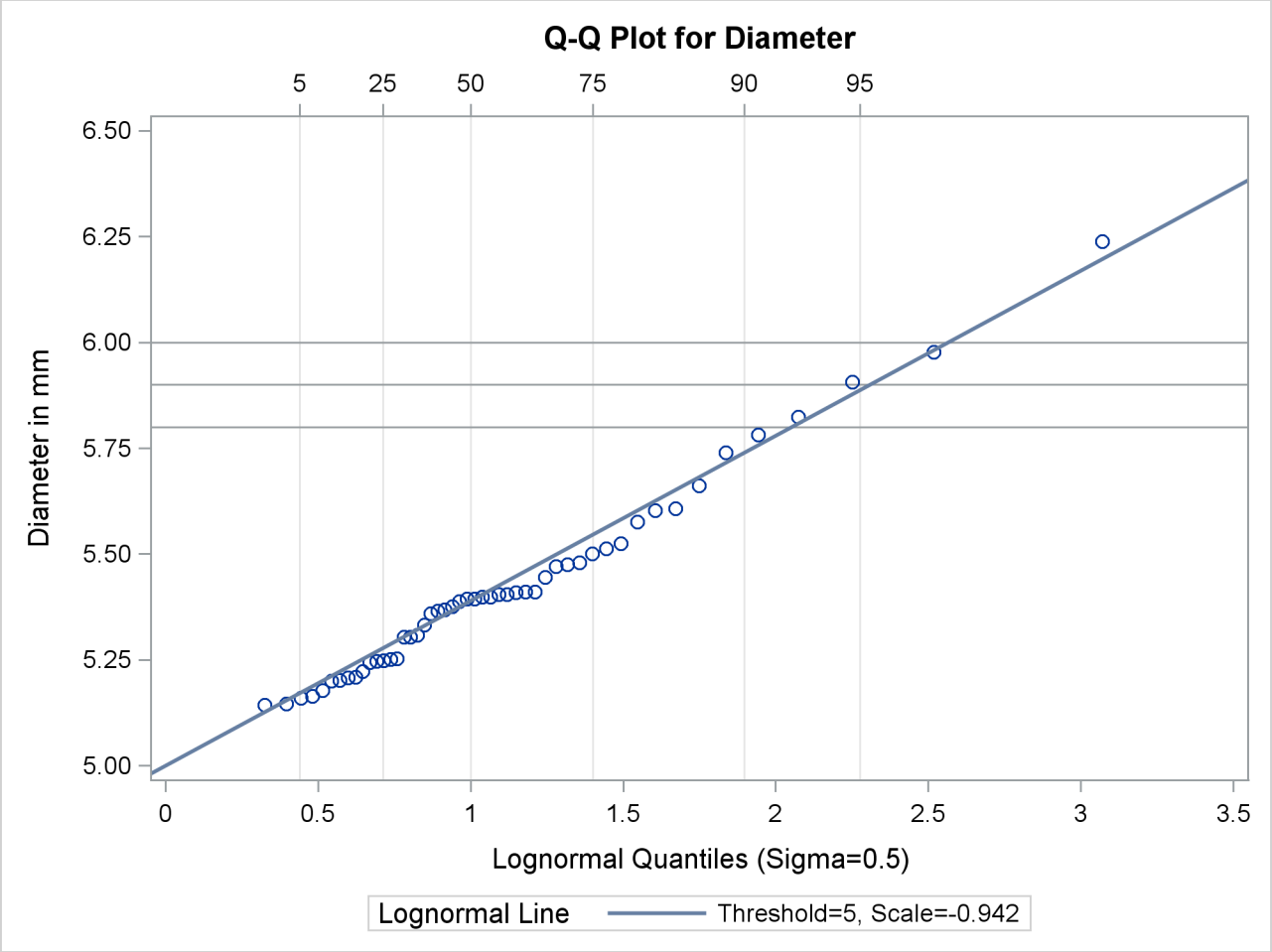
The point pattern in Output 6.22.2 has a slope of approximately 0.39 and an intercept of 5. The following statements reproduce this plot, adding a lognormal reference line with this slope and intercept.

```
title 'Lognormal Q-Q Plot for Diameters';
proc capability data=Measures noprint;
   qqplot Diameter / lognormal(sigma=0.5 theta=5 slope=0.39)
                     pctlaxis(grid)
                     vref  = 5.8 5.9 6.0;
run;
```

The ODS GRAPHICS ON statement specified before the PROC CAPABILITY statement enables ODS Graphics, so the Q-Q plot is created using ODS Graphics instead of traditional graphics. The result is shown in Output 6.22.4.

[7] You can also use a probability plot for this purpose. See Output 6.20.1.

Output 6.22.4 Lognormal Q-Q Plot Identifying Percentiles



The PCTLAXIS option labels the major percentiles, and the GRID option draws percentile axis reference lines. The 95th percentile is 5.9, because the intersection of the distribution reference line and the 95th reference line occurs at this value on the vertical axis.

Alternatively, you can compute this percentile from the estimated lognormal parameters. The 100αth percentile of the lognormal distribution is

$$P_\alpha = \exp(\sigma\Phi^{-1}(\alpha) + \zeta) + \theta$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative standard normal distribution. Consequently,

$$P_{0.95} \approx \exp\left(\frac{1}{2}\Phi^{-1}(0.95) + \log(0.39)\right) + 5 \approx \exp\left(\frac{1}{2} \times 1.645 - 0.94\right) + 5 \approx 5.89$$

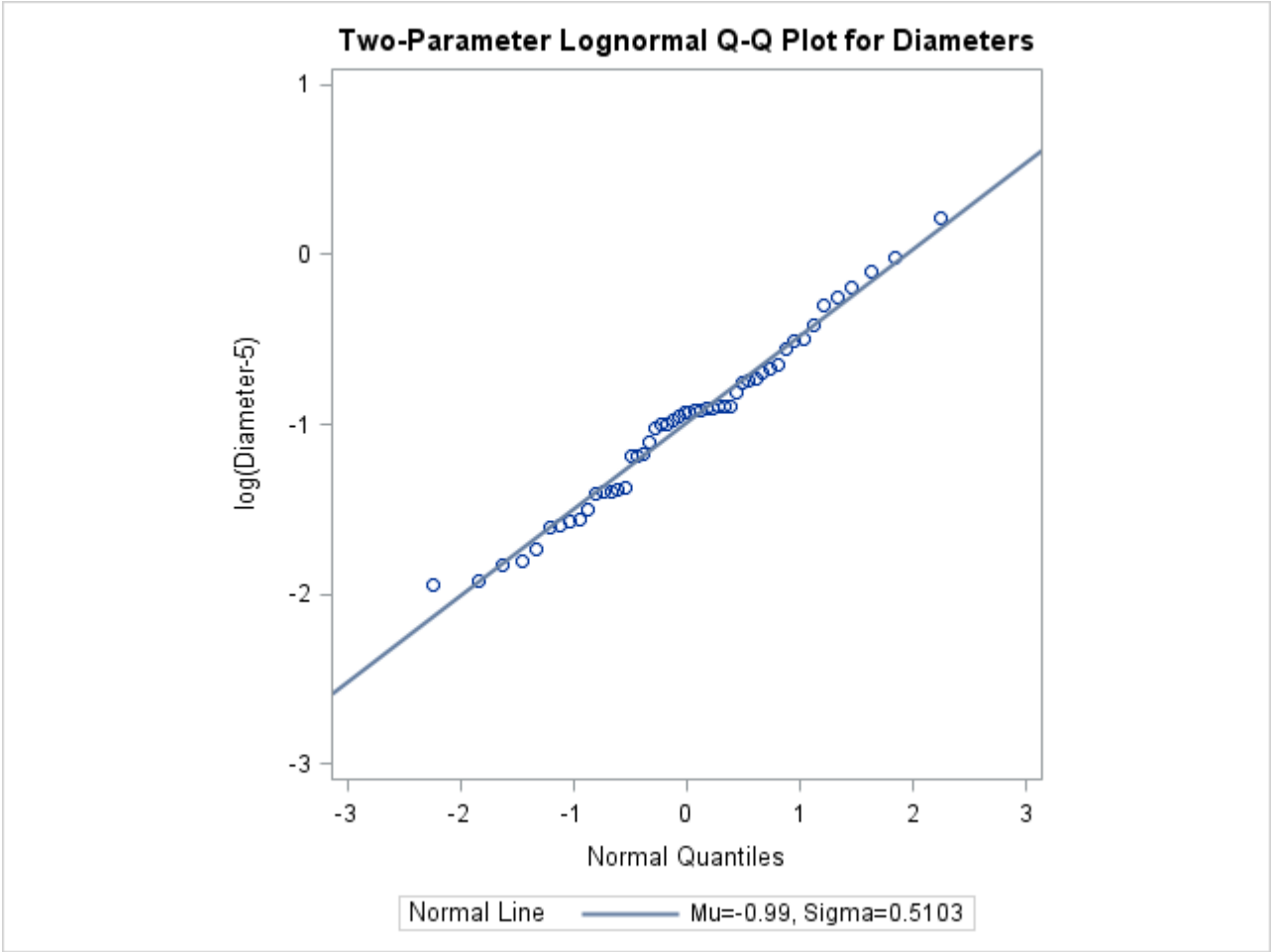
Two-Parameter Lognormal Plots

NOTE: See *Creating Lognormal Q-Q Plots* in the SAS/QC Sample Library.

If a known threshold parameter is available, you can construct a two-parameter lognormal Q-Q plot by subtracting the threshold from the data and requesting a normal Q-Q plot. The following statements create this plot for Diameter, assuming a known threshold of five:

```
data Measures;
   set Measures;
   Logdiam=log(Diameter-5);
   label Logdiam='log(Diameter-5)';
run;

title 'Two-Parameter Lognormal Q-Q Plot for Diameters';
proc capability data=Measures noprint;
   qqplot Logdiam / normal(mu=est sigma=est)
                    square
                    odstitle=title;
run;
```

Output 6.22.5 Two-Parameter Lognormal Q-Q Plot for Diameters



Because the point pattern in Output 6.22.5 is linear, you can estimate the lognormal parameters ζ and σ as the normal plot estimates of μ and σ, which are –0.99 and 0.51. These values correspond to the previous estimates of –0.92 for ζ and 0.5 for σ.

Example 6.23: Comparing Weibull Q-Q Plots

NOTE: See *Creating Weibull Q-Q Plots* in the SAS/QC Sample Library.

This example compares the use of three-parameter and two-parameter Weibull Q-Q plots for the failure times in months for 48 integrated circuits. The times are assumed to follow a Weibull distribution.

```
data Failures;
   input Time @@;
   label Time='Time in Months';
   datalines;
29.42 32.14 30.58 27.50 26.08 29.06 25.10 31.34
29.14 33.96 30.64 27.32 29.86 26.28 29.68 33.76
29.32 30.82 27.26 27.92 30.92 24.64 32.90 35.46
30.28 28.36 25.86 31.36 25.26 36.32 28.58 28.88
26.72 27.42 29.02 27.54 31.60 33.46 26.78 27.82
29.18 27.94 27.66 26.42 31.00 26.64 31.44 32.52
 ;
```

Three-Parameter Weibull Plots

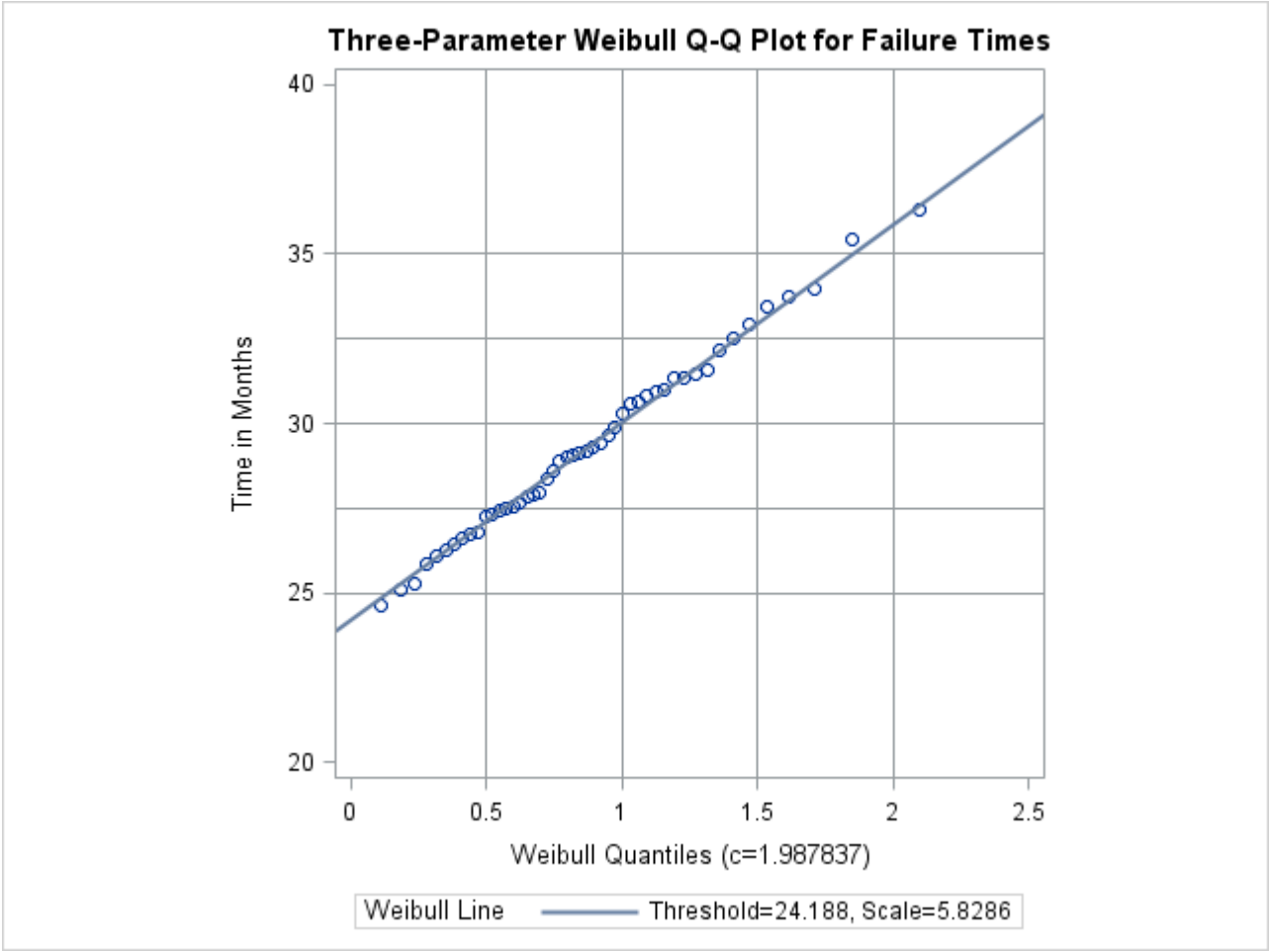
If no assumption is made about the parameters of this distribution, you can use the WEIBULL option to request a three-parameter Weibull plot. As in the previous example, you can visually estimate the shape parameter c by requesting plots for different values of c and choosing the value of c that linearizes the point pattern. Alternatively, you can request a maximum likelihood estimate for c, as illustrated in the following statements produce Weibull plots for c = 1, 2 and 3:

```
title 'Three-Parameter Weibull Q-Q Plot for Failure Times';
proc capability data=Failures noprint;
   qqplot Time / weibull(c=est theta=est sigma=est)
                 square
                 href=0.5 1 1.5 2
                 vref=25 27.5 30 32.5 35
                 odstitle=title;
run;
```

NOTE: When using the WEIBULL option, you must either specify a list of values for the Weibull shape parameter c with the C= option, or you must specify C=EST.

Output 6.23.1 displays the plot for the estimated value c = 1.99. The reference line corresponds to the estimated values for the threshold and scale parameters of ($\hat{\theta}_0$=24.19 and $\hat{\sigma}_0$=5.83, respectively.

Output 6.23.1 Three-Parameter Weibull Q-Q Plot for c = 2



Two-Parameter Weibull Plots

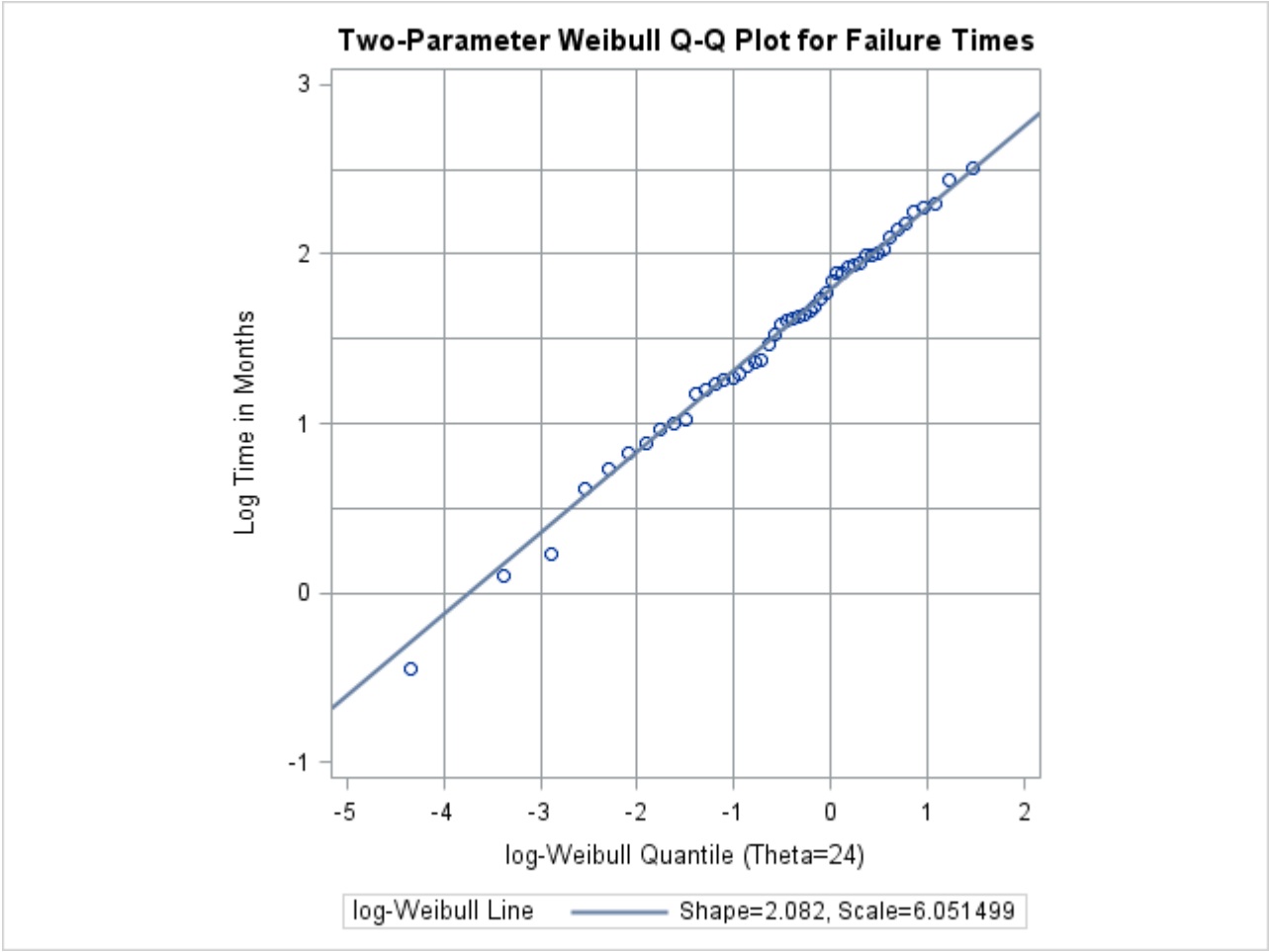
NOTE: See *Creating Weibull Q-Q Plots* in the SAS/QC Sample Library.

Now, suppose it is known that the circuit lifetime is at least 24 months. The following statements use the threshold value $\theta_0 = 24$ to produce the two-parameter Weibull Q-Q plot shown in Output 6.23.2:

```
title 'Two-Parameter Weibull Q-Q Plot for Failure Times';
proc capability data=Failures noprint;
   qqplot Time / weibull2(theta=24 c=est sigma=est) square
                 href= -4 to 1
                 vref= 0 to 2.5 by 0.5
                 odstitle=title;
run;
```

The reference line is based on maximum likelihood estimates $\hat{c}$=2.08 and $\hat{\sigma}$=6.05. These estimates agree with those of the previous example.

Output 6.23.2 Two-Parameter Weibull Q-Q Plot for $\theta_0 = 24$



Example 6.24: Estimating Cpk from a Normal Q-Q Plot

NOTE: See *Creating Normal Q-Q Plots* in the SAS/QC Sample Library.

This example illustrates how you can use a normal Q-Q plot to estimate the capability index C_{pk}. The data used here are the distance measurements provided in the section "Creating a Normal Quantile-Quantile Plot" on page 493.

The linearity of the point pattern in Figure 6.40 indicates that the measurements are normally distributed (recall that normality should be checked when process capability indices are reported). Furthermore, Figure 6.40 shows that the upper specification limit is about 1.7 standard deviation units above the mean, and the lower specification limit is about 1.8 standard deviation units below the mean. Because *CPU* is defined as

$$CPU = \frac{USL - \mu}{3\sigma}$$

and *CPL* is defined as

$$CPL = \frac{\mu - LSL}{3\sigma}$$

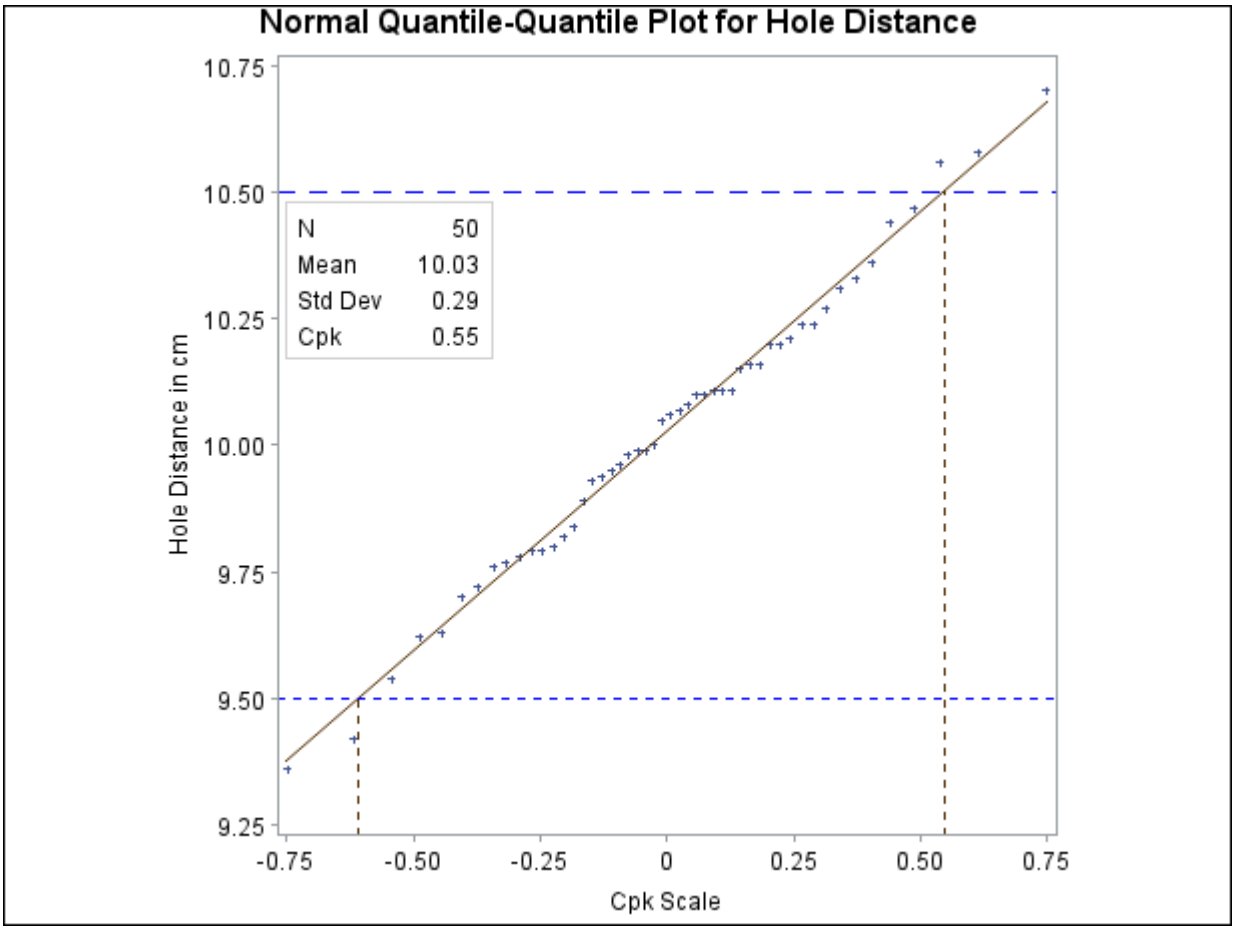
it follows that an estimate of *CPU* is $1.7/3 = 0.57$, and an estimate of *CPL* is $1.8/3 = 0.6$. Thus, except for a factor of three, you can estimate *CPU* and *CPL* from the points of intersection between the specification lines and the point pattern.

The following statements facilitate this type of estimation by creating a Q-Q plot, displayed in Output 6.24.1, in which the horizontal axis is rescaled by a factor of three:

```
ods graphics off;
symbol v=plus;
title "Normal Quantile-Quantile Plot for Hole Distance";
proc capability data=Sheets noprint;
   spec  lsl=9.5   llsl=2   clsl=blue
         usl=10.5  lusl=20  cusl=blue;
   qqplot Distance / normal(mu=est sigma=est cpkscale cpkref)
                     nolegend
                     square;
   inset n mean (5.2) std="Std Dev" (4.2) cpk (4.2) /
         pos=(-0.75,10.48) data refpoint=tl;
run;
```

The CPKSCALE option rescales the horizontal axis, and the CPKREF option adds reference lines indicating the intersections of the distribution reference line and the specification limits.

Output 6.24.1 Normal Q-Q Plot With C_{pk} Scaling



Using this display, you can estimate *CPU* and *CPL* directly from the horizontal axis as 0.55 and 0.60, respectively (the negative sign for –0.60 is ignored). The minimum of these values (0.55) is an estimate of C_{pk}. Note that this estimate agrees with the numerically obtained estimate for C_{pk} that is displayed on the plot with the INSET statement.

See Rodriguez (1992) for further discussion concerning the use of Q-Q plots in process capability analysis.

Dictionary of Common Options: CAPABILITY Procedure

This chapter provides detailed descriptions of options that you can specify in the following chart statements:

- CDFPLOT
- COMPHISTOGRAM
- HISTOGRAM
- PPPLOT
- PROBPLOT
- QQPLOT

As noted, some options are applicable only to comparative plots produced by the COMPHISTOGRAM statement or by another plot statement in conjunction with a CLASS statement.

General Options

ALPHADELTA=*value*

specifies the change in successive estimates of $\hat{\alpha}$ at which iteration terminates in the Newton-Raphson approximation of the maximum likelihood estimate of α for gamma distributions requested with the GAMMA option. Enclose the ALPHADELTA= option in parentheses after the GAMMA keyword. Iteration continues until the change in α is less than the value specified or the number of iterations exceeds the value of the MAXITER= option. The default value is 0.00001.

ALPHAINITIAL=*value*

specifies the initial value for $\hat{\alpha}$ in the Newton-Raphson approximation of the maximum likelihood estimate of α for gamma distributions requested with the GAMMA option. Enclose the ALPHAINITIAL= option in parentheses after the GAMMA keyword. The default value is Thom's approximation of the estimate of α. See Johnson, Kotz, and Balakrishnan (1995).

CDELTA=*value*

specifies the change in successive estimates of c at which iterations terminate in the Newton-Raphson approximation of the maximum likelihood estimate of c for Weibull distributions requested by the WEIBULL option. Enclose the CDELTA= option in parentheses after the WEIBULL keyword. Iteration continues until the change in c between consecutive steps is less than the *value* specified or until the number of iterations exceeds the value of the MAXITER= option. The default value is 0.00001.

CINITIAL=*value*

specifies the initial value for $\hat{c}$ in the Newton-Raphson approximation of the maximum likelihood estimate of c for Weibull distributions requested with the WEIBULL or WEIBULL2 option. The default value is 1.8. See Johnson, Kotz, and Balakrishnan (1995).

CONTENTS='*string*'

specifies the table of contents grouping entry for output produced by the plot statement. You can specify CONTENTS='' to suppress the grouping entry.

CPROP

CPROP=*color* | **EMPTY**

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 in a comparative plot. By default, no proportion bars are displayed. You can specify the keyword EMPTY to display empty bars.

For traditional graphics with the GSTYLE system option in effect, you can specify CPROP with no argument to produce proportion bars using an appropriate color from the ODS style. The CPROP option is not available with ODS Graphics.

HAXIS=*value*

specifies the name of an AXIS statement describing the horizontal axis.

HREF=*values*

draws reference lines that are perpendicular to the horizontal axis at the values that you specify. Also see the CHREF= and LHREF= options.

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

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

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

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

HREFLABPOS=*n*

specifies the vertical position of HREFLABELS= labels, as described in the following table.

n	Position
1	along top of plot
2	staggered from top to bottom of plot
3	along bottom of plot
4	staggered from bottom to top of plot

By default, HREFLABPOS=1. **NOTE:** HREFLABPOS=2 and HREFLABPOS=4 are not supported for ODS Graphics output.

INTERTILE=*value*

specifies the distance in horizontal percentage screen units between the framed areas, called *tiles*, of a comparative plot. By default, INTERTILE=0.75 percentage screen units. You can specify INTERTILE=0 to create contiguous tiles.

MAXITER=*n*

specifies the maximum number of iterations in the Newton-Raphson approximation of the maximum likelihood estimate of α for gamma distributions requested with the GAMMA option and c for Weibull distributions requested with the WEIBULL and WEIBULL2 options. Enclose the MAXITER= option in parentheses after the GAMMA, WEIBULL, or WEIBULL2 keywords. The default value of n is 20.

NCOLS=*n*

NCOL=*n*

specifies the number of columns per panel in a comparative plot. By default, NCOLS=1 if you specify only one CLASS variable, and NCOLS=2 if you specify two CLASS variables. If you specify two CLASS variables, you can use the NCOLS= option with the NROWS= option.

NOHLABEL

suppresses the label for the horizontal axis. You can use this option to reduce clutter.

NOVLABEL

suppresses the label for the vertical axis. You can use this option to reduce clutter.

NOVTICK

suppresses the tick marks and tick mark labels for the vertical axis. This option also suppresses the label for the vertical axis.

NROWS=*n*

NROW=*n*

specifies the number of rows per panel in a comparative plot. By default, NROWS=2. If you specify two CLASS variables, you can use the NCOLS= option with the NROWS= option.

ODSFOOTNOTE=FOOTNOTE | FOOTNOTE1 | '*string***'**

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

\n	analysis variable name
\l	analysis variable label (or name if the analysis variable has no label)

ODSFOOTNOTE2=FOOTNOTE2 | '*string***'**

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

\n	analysis variable name
\l	analysis variable label (or name if the analysis variable has no label)

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

TITLE (or TITLE1)	uses the value of SAS TITLE statement as the graph title.
NONE	suppresses all titles from the graph.
DEFAULT	uses the default ODS Graphics title (a descriptive title consisting of the plot type and the analysis variable name.)
LABELFMT	uses the default ODS Graphics title with the variable label instead of the variable name.

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

\n	analysis variable name
\l	analysis variable label (or name if the analysis variable has no label)

ODSTITLE2=TITLE2 | '*string*'
specifies a secondary title for ODS Graphics output. If you specify the TITLE2 keyword, the value of SAS TITLE2 statement is used as the secondary graph title. If you specify a quoted string, that is used as the secondary title. The quoted string can contain the following escaped characters, which are replaced with the appropriate values from the analysis:

\n	analysis variable name
\l	analysis variable label (or name if the analysis variable has no label)

OVERLAY
specifies that plots associated with different levels of a CLASS variable be overlaid onto a single plot, rather than displayed as separate cells in a comparative plot. If you specify the OVERLAY option with one CLASS variable, the output associated with each level of the CLASS variable is overlaid on a single plot. If you specify the OVERLAY option with two CLASS variables, a comparative plot based on the first CLASS variable's levels is produced. Each cell in this comparative plot contains overlaid output associated with the levels of the second CLASS variable.

The OVERLAY option applies only to ODS Graphics output. It is not available in the COMPHISTOGRAM statement.

SCALE=*value*
is an alias for the SIGMA= option for distributions requested by the BETA, EXPONENTIAL, GAMMA, SB, SU, WEIBULL, and WEIBULL2 options and for the ZETA= option for distributions requested by the LOGNORMAL option.

SHAPE=*value*
is an alias for the ALPHA= option for distributions requested by the GAMMA option, for the SIGMA= option for distributions requested by the LOGNORMAL option, and for the C= option for distributions requested by the WEIBULL and WEIBULL2 options.

STATREF=*keyword-list*

draws reference lines at the values of the statistics requested in the *keyword-list*. These reference lines are perpendicular to the horizontal axis in a histogram or cdf plot, and perpendicular to the vertical axis in a probability or Q-Q plot (unless the ROTATE option is specified). The STATREF= option does not apply to the PPPLOT statement.

Valid keywords are listed in the following table.

Keyword	Statistic
MAX	largest value
MEAN	sample mean
MEDIAN | Q2	median (50th percentile)
MIN	smallest value
MODE	most frequent value
P *pctl*	*pctl*th percentile
Q1	lower quartile (25th percentile)
Q3	upper quartile (75th percentile)
factor STD	*factor* standard deviations from the mean

Note that the *factor* specified with the STD keyword can be positive (which puts a reference line above the mean) or negative (below the mean).

Also see the CSTATREF=, LSTATREF=, STATREFLABELS=, and STATREFSUBCHAR= options.

STATREFLABELS='*label1*' **...** '*labeln*'

STATREFLABEL='*label1*' **...** '*labeln*'

STATREFLAB='*label1*' **...** '*labeln*'

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

STATREFSUBCHAR='*keyword-list*'

specifies a substitution character (such as #) for labels specified with the STATREFLABELS= option. When the labels are displayed on a graph, the first occurrence of the specified character in each label is replaced with the value of the corresponding STATREF= statistic.

VAXIS=*name*

VAXIS=*value-list*

specifies the name of an AXIS statement describing the vertical axis. In a COMPHISTOGRAM or HISTOGRAM statement, you can alternatively specify a *value-list* for the vertical axis.

VAXISLABEL='*label*'

specifies a label for the vertical axis. Labels can have up to 40 characters.

VREF=*value-list*

draws reference lines perpendicular to the vertical axis at the values specified. Also see the CVREF= and LVREF= options.

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

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

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

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

VREFLABPOS=*n*

specifies the horizontal position of VREFLABELS= labels. If you specify VREFLABPOS=1, the labels are positioned at the left of the plot. If you specify VREFLABPOS=2, the labels are positioned at the right of the plot. By default, VREFLABPOS=1 for traditional graphics and 2 for ODS Graphics.

Options for Traditional Graphics

ANNOKEY

applies the annotation requested with the ANNOTATE= option only to the key cell of a comparative plot. By default, the procedure applies annotation to all of the cells. You can use the KEYLEVEL= option in the CLASS statement or the CLASSKEY= option in the COMPHISTOGRAM statement to specify the key cell.

ANNOTATE=*SAS-data-set*

ANNO=*SAS-data-set*

specifies an input data set that contains annotate variables, as described in *SAS/GRAPH: Help*, for annotating traditional graphics. The ANNOTATE= data set you specify in the plot 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 (see "ANNOTATE= Data Sets" on page 221).

CAXIS=*color*

CAXES=*color*

CA=*color*

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

CFRAME=*color*

specifies the color for the area that is enclosed by the axes and frame.

CFRAMESIDE=*color*

specifies the color to fill the frame area for the row labels that display along the left side of a comparative plot. This color also fills the frame area for the label of the corresponding CLASS variable, if you associate a label with the variable.

CFRAMETOP=*color*

specifies the color to fill the frame area for the column labels that display across the top of a comparative plot. This color also fills the frame area for the label of the corresponding CLASS variable, if you associate a label with the variable.

CHREF=*color* | *(color-list)*

CH=*color* | *(color-list***)**

specifies the colors for horizontal axis reference lines requested by the HREF= option. If you specify a single color, it is used for all HREF= lines. Otherwise, if there are fewer colors specified than reference lines requested, the remaining lines are displayed with the default reference line color. You can also specify the value *_default* in the color list to request the default color.

COLOR=*color*

COLOR=*color-list*

specifies the color of the curve or reference line associated with a distribution or kernel density estimate. Enclose the COLOR= option in parentheses after a distribution option or the KERNEL option. In a HISTOGRAM statement, you can specify a list of colors in parentheses for multiple density curves.

CSTATREF=*color* | *(color-list)*

specifies the colors for reference lines requested by the STATREF= option. If you specify a single color, it is used for all STATREF= lines. Otherwise, if there are fewer colors specified than reference lines requested, the remaining lines are displayed with the default reference line color. You can also specify the value *_default* in the color list to request the default color.

CTEXT=*color*

CT=*color*

specifies the color for tick mark values and axis labels.

CTEXTSIDE=*color*

specifies the color for the row labels that display along the left side of a comparative plot. If you do not specify the CTEXTSIDE= option, the color specified with the CTEXT= option is used. You can specify the CFRAMESIDE= option to change the background color for the row labels.

CTEXTTOP=*color*

specifies the color for the column labels that display along the left side of a comparative plot. If you do not specify the CTEXTTOP= option, the color specified with the CTEXT= option is used. You can use the CFRAMETOP= option to change the background color for the column labels.

CVREF=*color* | *(color-list)*

CV=*color* | *(color-list)*

specifies the colors for lines requested with the VREF= option. If you specify a single color, it is used for all VREF= lines. Otherwise, if there are fewer colors specified than reference lines requested, the remaining lines are displayed with the default reference line color. You can also specify the value *_default* in the color list to request the default color.

DESCRIPTION='*string*'

DES='*string*'

specifies a description, up to 256 characters long, for the GRSEG catalog entry for a traditional graphics chart. The default value is the analysis variable name.

FONT=*font*

specifies a font for reference line and axis labels. You can also specify fonts for axis labels in an AXIS statement. The FONT= option takes precedence over the FTEXT= font specified in the GOPTIONS statement. For a list of software fonts, see *SAS/GRAPH: Help*.

HEIGHT=*value*

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

HMINOR=*n*

HM=*n*

specifies the number of minor tick marks between each major tick mark on the horizontal axis. Minor tick marks are not labeled. By default, HMINOR=0.

INFONT=*font*

specifies a font to use for text inside the framed areas of the plot. The INFONT= option takes precedence over the FTEXT= option in the GOPTIONS statement. For a list of software fonts, see *SAS/GRAPH: Help*.

INHEIGHT=*value*

specifies the height, in percentage screen units, of text used inside the framed areas of the plot. If you do not specify the INHEIGHT= option, the height specified with the HEIGHT= option is used.

L=*linetype*

L=*linetype-list*

specifies the line type of the curve or reference line associated with a distribution or kernel density estimate. Enclose the L= option in parentheses after the distribution option or the KERNEL option. In a HISTOGRAM statement, you can specify a list of line types in parentheses for multiple density curves.

LHREF=*linetype* | *linetype-list*

LH=*linetype* | *linetype-list*

specifies the line types for the reference lines that you request with the HREF= option. If you specify a single line type, it is used for all HREF= lines. Otherwise, if there are fewer line types specified than reference lines requested, the remaining lines are displayed with the default reference line type. You can also specify line type 0 to request the default color.

LSTATREF=*linetype* | *linetype-list*

specifies the line types for the reference lines that you request with the STATREF= option. If you specify a single line type, it is used for all STATREF= lines. Otherwise, if there are fewer line types specified than reference lines requested, the remaining lines are displayed with the default reference line type. You can also specify line type 0 to request the default color.

LVREF=*linetype* | *linetype-list*

LV=*linetype* | *linetype-list*

specifies the line types for lines requested with the VREF= option. If you specify a single line type, it is used for all VREF= lines. Otherwise, if there are fewer line types specified than reference lines requested, the remaining lines are displayed with the default reference line type. You can also specify line type 0 to request the default color.

NAME='*string*'

specifies the name of the GRSEG catalog entry for a traditional graphics plot, and the name of the graphics output file if one is created. The name can be up to 256 characters long, but the GRSEG name is truncated to eight characters. The default value is 'CAPABILI'.

NOFRAME
suppresses the frame around the subplot area.

TURNVLABELS
TURNVLABEL
turns the characters in the vertical axis labels so that they display vertically.

VMINOR=*n*
VM=*n*
specifies the number of minor tick marks between each major tick mark on the vertical axis. Minor tick marks are not labeled. The default is zero.

W=*value*
W=*value-list*
specifies the width in pixels of the curve or reference line associated with a distribution or kernel density estimate. Enclose the W= option in parentheses after the distribution option or the KERNEL option. In a HISTOGRAM statement, you can specify a list of widths in parentheses for multiple density curves.

WAXIS=*n*
specifies the line thickness, in pixels, for the axes and frame.

Options for Legacy Line Printer Charts

HREFCHAR='*character***'**
specifies the character used to form the lines requested by the HREF= option for a line printer chart. The default is the vertical bar (|).

VREFCHAR='*character***'**
VREF= option for a line printer chart. specifies the character used to form the lines requested by the The default is the hyphen (-).

References

Bai, D. S., and Choi, I. S. (1997). "Process Capability Indices for Skewed Populations." Unpublished manuscript, Korean Advanced Institute of Science and Technology, Taejon, Korea.

Bissell, A. F. (1990). "How Reliable Is Your Capability Index?" *Journal of the Royal Statistical Society, Series C* 39:331–340.

Blom, G. (1958). *Statistical Estimates and Transformed Beta Variables*. New York: John Wiley & Sons.

Bowman, K. O., and Shenton, L. R. (1983). "Johnson's System of Distributions." In *Encyclopedia of Statistical Sciences*, vol. 4, edited by S. Kotz, N. L. Johnson, and C. B. Read. New York: John Wiley & Sons.

Boyles, R. A. (1991). "The Taguchi Capability Index." *Journal of Quality Technology* 23:107–126.

Boyles, R. A. (1992). *Cpm for Asymmetrical Tolerances*. Technical report, Precision Castparts Corp., Portland, OR.

Boyles, R. A. (1994). "Process Capability with Asymmetric Tolerances." *Communications in Statistics—Simulation and Computation* 23:615–643.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Belmont, CA: Wadsworth International Group.

Chen, H. F., and Kotz, S. (1996). "An Asymptotic Distribution of Wright's Process Capability Index Sensitive to Skewness." *Journal of Statistical Computation and Simulation* 55:147–158.

Chen, K. S. (1998). "Incapability Index with Asymmetric Tolerances." *Statistica Sinica* 8:253–262.

Chou, Y., Owen, D. B., and Borrego, S. A. (1990). "Lower Confidence Limits on Process Capability Indices." *Journal of Quality Technology* 22:223–229; corrigenda, 24, 251.

Cohen, A. C. (1951). "Estimating Parameters of Logarithmic-Normal Distributions by Maximum Likelihood." *Journal of the American Statistical Association* 46:206–212.

Croux, C., and Rousseeuw, P. J. (1992). "Time-Efficient Algorithms for Two Highly Robust Estimators of Scale." *Computational Statistics* 1:411–428.

D'Agostino, R. B., and Stephens, M., eds. (1986). *Goodness-of-Fit Techniques*. New York: Marcel Dekker.

Dixon, W. J., and Tukey, J. W. (1968). "Approximate Behavior of the Distribution of Winsorized *t* (Trimming/Winsorization 2)." *Technometrics* 10:83–98.

Ekvall, D. N., and Juran, J. M. (1974). "Manufacturing Planning." In *Quality Control Handbook*, 3rd ed., edited by J. M. Juran. New York: McGraw-Hill.

Elandt, R. C. (1961). "The Folded Normal Distribution: Two Methods of Estimating Parameters from Moments." *Technometrics* 3:551–562.

Fowlkes, E. B. (1987). *A Folio of Distributions: A Collection of Theoretical Quantile-Quantile Plots*. New York: Marcel Dekker.

Gnanadesikan, R. (1997). *Statistical Data Analysis of Multivariate Observations*. New York: John Wiley & Sons.

Gupta, A. K., and Kotz, S. (1997). "A New Process Capability Index." *Metrika* 45:213–224.

Hahn, G. J. (1969). "Factors for Calculating Two-Sided Prediction Intervals for Samples from a Normal Distribution." *Journal of the American Statistical Association* 64:878–898.

Hahn, G. J. (1970a). "Additional Factors for Calculating Prediction Intervals for Samples from a Normal Distribution." *Journal of the American Statistical Association* 65:1668–1676.

Hahn, G. J. (1970b). "Statistical Intervals for a Normal Population, Part 2: Formulas, Assumptions, Some Derivations." *Journal of Quality Technology* 2:195–206.

Hahn, G. J., and Meeker, W. Q. (1991). *Statistical Intervals: A Guide for Practitioners*. New York: John Wiley & Sons.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Distributions.* 2nd ed. Vol. 1. New York: John Wiley & Sons.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions.* 2nd ed. Vol. 2. New York: John Wiley & Sons.

Johnson, N. L., Kotz, S., and Pearn, W. L. (1994). "Flexible Process Capability Indices." *Pakistan Journal of Statistics* 10:23–31.

Kane, V. E. (1986). "Process Capability Indices." *Journal of Quality Technology* 1:41–52.

Kotz, S., and Johnson, N. L. (1993). *Process Capability Indices*. London: Chapman & Hall.

Kotz, S., and Lovelace, C. R. (1998). *Process Capability Indices in Theory and Practice*. London: Edward Arnold.

Krishnamoorthy, K., and Mathew, T. (2009). *Statistical Tolerance Regions: Theory, Applications, and Computation*. Hoboken, NJ: John Wiley & Sons.

Kushler, R. H., and Hurley, P. (1992). "Confidence Bounds for Capability Indices." *Journal of Quality Technology* 24:188–195.

Lehmann, E. L., and D'Abrera, H. J. M. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day.

Luceño, A. (1996). "A Process Capability Index with Reliable Confidence Intervals." *Communications in Statistics—Simulation and Computation* 25:235–245.

Marcucci, M. O., and Beazley, C. F. (1988). "Capability Indices: Process Performance Measures." *Transactions of ASQC Congress* 42:516–523.

Montgomery, D. C. (1996). *Introduction to Statistical Quality Control*. 3rd ed. New York: John Wiley & Sons.

Odeh, R. E., and Owen, D. B. (1980). *Tables for Normal Tolerance Limits, Sampling Plans, and Screening*. New York: Marcel Dekker.

Owen, D. B., and Hua, T. A. (1977). "Tables of Confidence Limits on the Tail Area of the Normal Distribution." *Communications in Statistics—Simulation and Computation* 6:285–311.

Pearn, W. L., Kotz, S., and Johnson, N. L. (1992). "Distributional and Inferential Properties of Process Capability Indices." *Journal of Quality Technology* 24:216–231.

Rodriguez, R. N. (1992). "Recent Developments in Process Capability Analysis." *Journal of Quality Technology* 24:176–187.

Rodriguez, R. N., and Bynum, R. A. (1992). "Examples of Short Run Process Control Methods with the SHEWHART Procedure in SAS/QC Software." Unpublished manuscript available from the authors.

Rousseeuw, P. J., and Croux, C. (1993). "Alternatives to the Median Absolute Deviation." *Journal of the American Statistical Association* 88:1273–1283.

Royston, J. P. (1992). "Approximating the Shapiro-Wilk W Test for Nonnormality." *Statistics and Computing* 2:117–119.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. New York: Chapman & Hall.

Slifker, J. F., and Shapiro, S. S. (1980). "The Johnson System: Selection and Parameter Estimation." *Technometrics* 22:239–246.

Terrell, G. R., and Scott, D. W. (1985). "Oversmoothed Nonparametric Density Estimates." *Journal of the American Statistical Association* 80:209–214.

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

Tukey, J. W., and McLaughlin, D. H. (1963). "Less Vulnerable Confidence and Significance Procedures for Location Based on a Single Sample: Trimming/Winsorization 1." *Sankhyā, Series A* 25:331–352.

Vännmann, K. (1995). "A Unified Approach to Capability Indices." *Statistica Sinica* 5:805–820.

Vännmann, K. (1997). "A General Class of Capability Indices in the Case of Asymmetric Tolerances." *Communications in Statistics—Theory and Methods* 26:2049–2072.

Velleman, P. F., and Hoaglin, D. C. (1981). *Applications, Basics, and Computing of Exploratory Data Analysis*. Boston: Duxbury Press.

Wadsworth, H. M., Stephens, K. S., and Godfrey, A. B. (1986). *Modern Methods for Quality Control and Improvement*. New York: John Wiley & Sons.

Wainer, H. (1974). "The Suspended Rootogram and Other Visual Displays: An Empirical Validation." *American Statistician* 28:143–145.

Wilk, M. B., and Gnanadesikan, R. (1968). "Probability Plotting Methods for the Analysis of Data." *Biometrika* 49:525–545.

Wright, P. A. (1995). "A Process Capability Index Sensitive to Skewness." *Journal of Statistical Computation and Simulation* 52:195–203.

Zhang, N. F., Stenback, G. A., and Wardrop, D. M. (1990). "Interval Estimation of Process Capability Index Cpk." *Communications in Statistics—Theory and Methods* 19:4455–4470.

Subject Index

Syntax Index