Chapter 4
HISTOGRAM Statement

Chapter Table of Contents

Chapter 4
HISTOGRAM Statement

Overview

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

- specify the midpoints for histogram intervals

- display specification limits on histograms

- display density curves for fitted theoretical distributions (beta, exponential, gamma, Johnson S_B, Johnson S_U, lognormal, normal, and Weibull) 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

Getting Started

This section introduces the HISTOGRAM statement with examples that illustrate commonly used options. Complete syntax for the HISTOGRAM statement is presented in the "Syntax" section on page 124, and advanced examples are given in the "Examples" section on page 170.

Creating a Histogram with Specification Limits

See CAPHST1
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 4.1:

```
title 'Process Capability Analysis of Plating Thickness';
proc capability data=trans noprint;
   spec lsl=3.45 llsl=2 usl=3.55 lusl=2;
   histogram thick;
run;
```

A histogram is created for each variable listed after the keyword HISTOGRAM. If you specify the LINEPRINTER option in the PROC CAPABILITY statement, the histogram is displayed in line printer output, as shown in Figure 4.2. * The SPEC statement, which is optional, provides the specification limits that are displayed on the histogram. For more information on the SPEC statement, see "Syntax for the SPEC Statement" on page 26.

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 9 for an example of this output.

---

*In Release 6.12 and previous releases of SAS/QC software, the keyword GRAPHICS was required in the PROC CAPABILITY statement to specify that the chart be created with a graphics device. In Version 7, you can specify the LINEPRINTER option to request line printer plots.
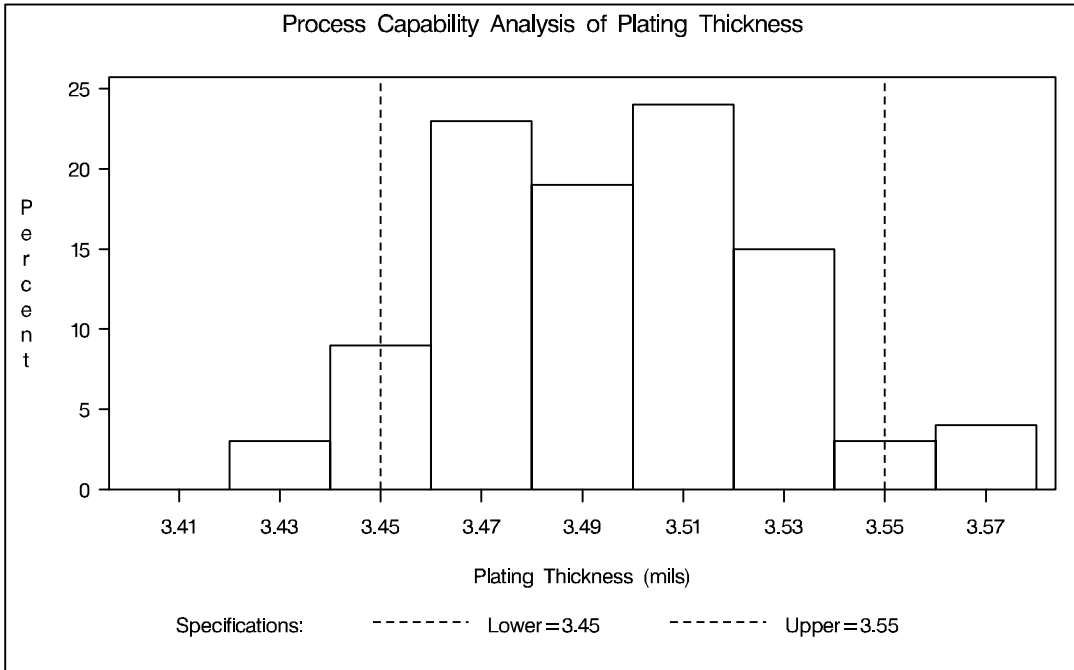
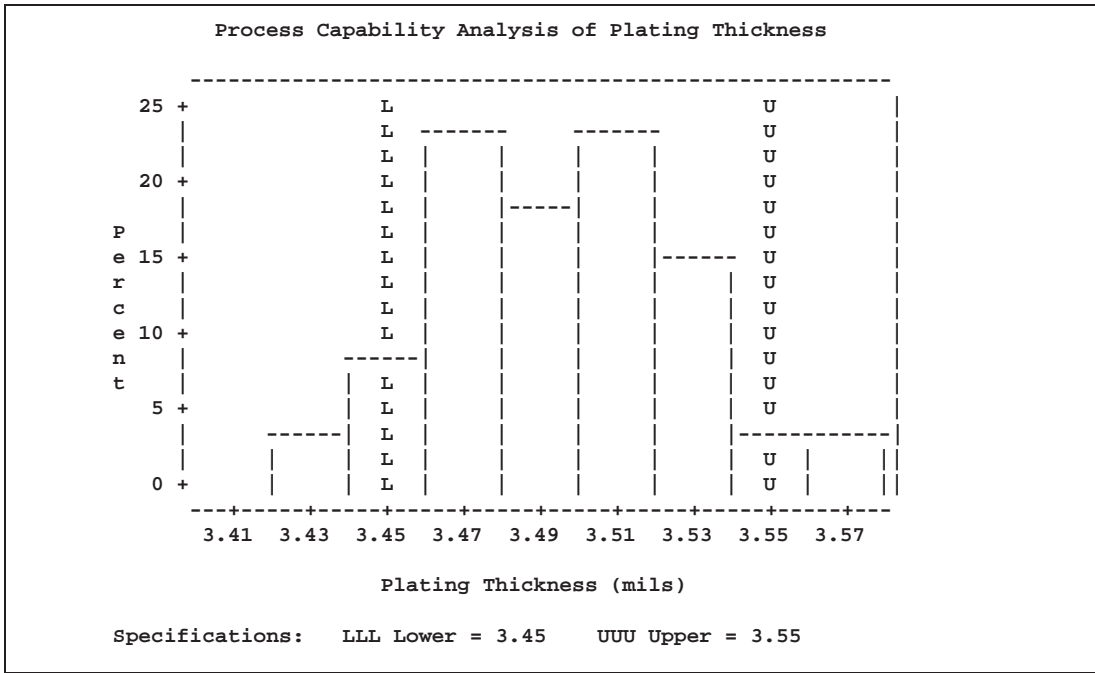**Figure 4.1.** Histogram Created with Graphics Device

Figure 4.2. Histogram Created with Line Printer

119

Adding a Normal Curve to the Histogram

This example is a continuation of the preceding example.

The following statements fit a normal distribution using the thickness measurements and superimpose the fitted density curve on the histogram:

```
title 'Process Capability Analysis of Plating Thickness';
proc capability data=trans noprint;
   spec lsl=3.45 llsl=2 usl=3.55 lusl=2;
   histogram thick / normal;
run;
```

The NORMAL option summarizes the fitted distribution in the printed output shown in Figure 4.3, and it specifies that the normal curve be displayed on the histogram shown in Figure 4.4.

```
                    The CAPABILITY Procedure
               Fitted Normal Distribution for thick

                 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


              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
```
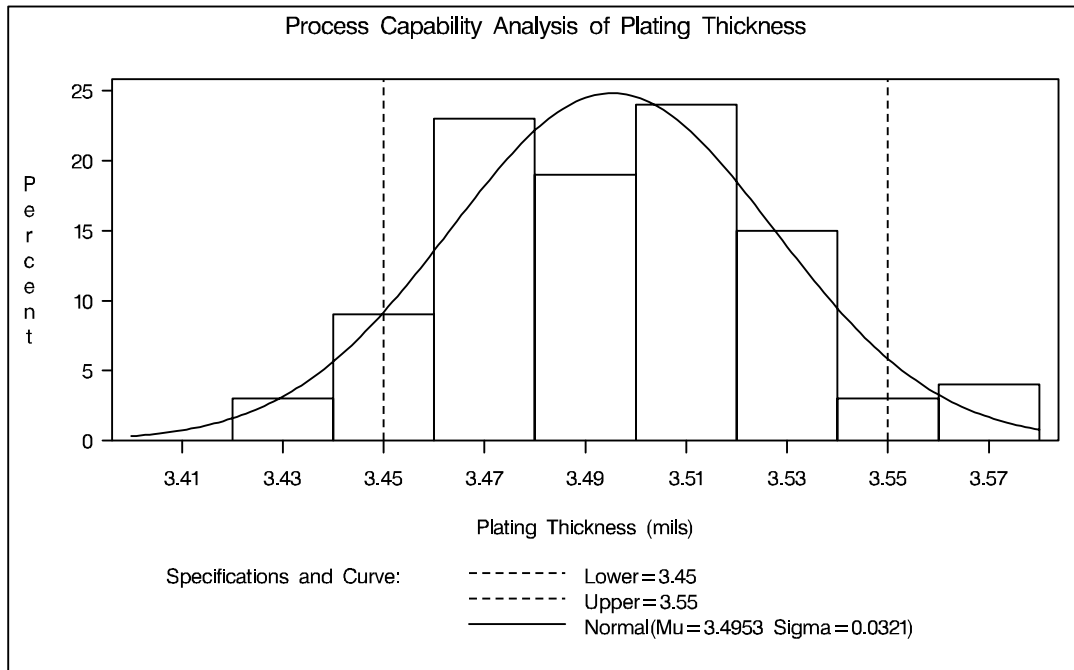
Figure 4.3. Summary for Fitted Normal Distribution

Figure 4.4. 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 ($\hat{\sigma} = 0.03211691$).

- a chi-square goodness-of-fit test. Compared to the usual cutoff values of 0.05 and 0.10, the p-value of 0.2229 for this test indicates that the thicknesses are normally distributed.

- 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 smaller 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 0.2229 for this test indicates that the thicknesses are normally distributed. In general EDF tests (when available) are preferable to chi-square tests. See the "EDF Goodness-of-Fit Tests" section on page 159 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 157. 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 4.3, specify the NOPRINT option enclosed in parentheses after the NORMAL option, as on page 122.

The NORMAL option is one of many options that you can specify in the HISTOGRAM statement. See the "Syntax" section on page 124 for a complete list of options or the "Dictionary of Options" section on page 130 for detailed descriptions of options.

Customizing a Histogram

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 llsl=2 usl=3.55 lusl=3;
   histogram thick / normal( noprint )
                     midpoints = 3.4 to 3.6 by 0.025
                     vscale    = count
                     cfill     = yellow
                     nospeclegend ;
   inset lsl usl                / cfill=blank;
   inset n mean (5.2) cpk (5.2) / cfill=blank;
run;
```

The histogram is displayed in Figure 4.5.
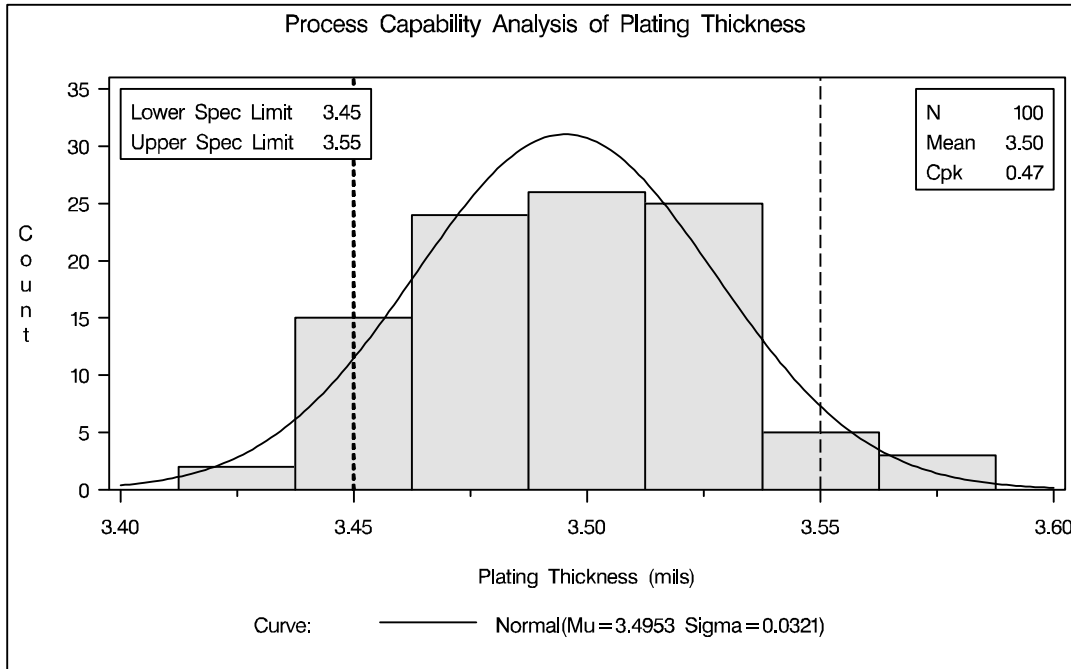
Figure 4.5. Customizing the Appearance of the Histogram

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 CFILL= option specifies a color for the histogram bars. 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 4.4.

For more information about HISTOGRAM statement options, see "Dictionary of Options" on page 130. For details on the INSET statement, see Chapter 5, "INSET Statement" on page 191.

Syntax

The syntax for the HISTOGRAM statement is as follows:

HISTOGRAM $<variables> < I\ options>$;

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;
```

Likewise, the following statements 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 130.

Parametric Density Estimation Options

Table 4.1 lists options that display a parametric density estimate on the histogram.

Table 4.1. Parametric Distribution Options

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
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
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

125

Table 4.2 through Table 4.10 list options that specify parameters for fitted parametric distributions and that control the display of fitted curves. Specify these options in parentheses after the distribution keyword. For example, the following statements fit a normal curve with the keyword NORMAL:

```
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.

Table 4.2. Options Used with All Parametric Distribution Options

COLOR=*color*	specifies color of fitted density curve
FILL	fills area under fitted density curve
INDICES	calculates capability indices based on fitted distribution
L=*linetype*	specifies line type of fitted curve
MIDPERCENTS	prints table of midpoints of histogram intervals
NOPRINT	suppresses printed output summarizing fitted curve
PERCENTS=*value-list*	lists percents for which quantiles calculated from data and quantiles estimated from fitted curve are tabulated
SYMBOL='*character*'	specifies character used to plot fitted density curve if histogram is produced on a line printer
W=*n*	specifies width of fitted density curve

Table 4.3. Beta-Options

ALPHA=*value*	specifies first shape parameter α for fitted beta curve
BETA=*value*	specifies second shape parameter β for fitted beta curve
SIGMA=*value*|EST	specifies scale parameter σ for fitted beta curve
THETA=*value*|EST	specifies lower threshold parameter θ for fitted beta curve

Table 4.4. Exponential-Options

SIGMA=*value*	specifies scale parameter σ for fitted exponential curve
THETA=*value*|EST	specifies threshold parameter θ for fitted exponential curve

Table 4.5. Gamma-Options

ALPHADELTA=*value*	specifies change in successive estimates of α at which the Newton-Raphson approximation of $\hat{\alpha}$ terminates
ALPHAINITIAL=*value*	specifies initial value for α in Newton-Raphson approximation of $\hat{\alpha}$
MAXITER=*n*	specifies maximum number of iterations in Newton-Raphson approximation of $\hat{\alpha}$
SIGMA=*value*	specifies scale parameter σ for fitted gamma curve
ALPHA=*value*	specifies shape parameter α for fitted gamma curve
THETA=*value*|EST	specifies threshold parameter θ for fitted gamma curve

Table 4.6. Lognormal-Options

ZETA=*value*	specifies scale parameter ζ for fitted lognormal curve
SIGMA=*value*	specifies shape parameter σ for fitted lognormal curve
THETA=*value*|EST	specifies threshold parameter θ for fitted lognormal curve

Table 4.7. Normal-Options

MU=*value*	specifies mean μ for fitted normal curve
SIGMA=*value*	specifies standard deviation σ for fitted normal curve

Table 4.8. S_B-Options

DELTA=*value*	specifies first shape parameter δ for fitted S_B curve
FITINTERVAL=*value*	specifies z-value for method of percentiles
FITMETHOD=MLE| PERCENTILE| MOMENTS	specifies method of parameter estimation
GAMMA=*value*	specifies second shape parameter γ for fitted S_B curve
SIGMA=*value*|EST	specifies scale parameter σ for fitted S_B curve
THETA=*value*|EST	specifies lower threshold parameter θ for fitted S_B curve
FITTOLERANCE=*value*	specifies tolerance for method of percentiles

127

Table 4.9. S_U-Options

DELTA=*value*	specifies first shape parameter δ for fitted S_U curve
FITINTERVAL=*value*	specifies z-value for method of percentiles
FITMETHOD=MLE| PERCENTILE| MOMENTS	specifies method of parameter estimation
GAMMA=*value*	specifies second shape parameter γ for fitted S_U curve
SIGMA=*value*|EST	specifies scale parameter σ for fitted S_U curve
THETA=*value*|EST	specifies lower threshold parameter θ for fitted S_U curve
FITTOLERANCE=*value*	specifies tolerance for method of percentiles

Table 4.10. Weibull-Options

C=*value*	specifies shape parameter c for fitted Weibull curve
CDELTA=*value*	specifies change in successive estimates of c at which the Newton-Raphson approximation of $\hat{c}$ terminates
CINITIAL=*value*	specifies initial value for c in Newton-Raphson approximation of $\hat{c}$
MAXITER=*n*	specifies maximum number of iterations in Newton-Raphson approximation of $\hat{c}$
SIGMA=*value*	specifies scale parameter σ for fitted Weibull curve
THETA=*value*|EST	specifies threshold parameter θ for fitted Weibull curve

Nonparametric Density Estimation Options

Table 4.11. Kernel Density Estimation Options

KERNEL(*kernel-options*)	fits kernel density estimates

Specify the options listed in Table 4.12 in parentheses after the keyword KERNEL to control features of kernel density estimates requested with the KERNEL option.

Table 4.12. Kernel-Options

C=*value* | MISE	specifies standardized bandwidth parameter c for fitted kernel density estimate
COLOR=*color*	specifies color of the fitted kernel density curve
FILL	fills area under fitted kernel density curve
K=NORMAL | QUADRATIC | TRIANGULAR	specifies type of kernel function
L=*linetype*	specifies line type used for fitted kernel density curve
SYMBOL='*character*'	specifies character used to plot fitted kernel density curve if the histogram is produced on a line printer
W=*n*	specifies line width for fitted kernel density curve

General Options

Table 4.13 through Table 4.16 summarize general options for the HISTOGRAM statement, including options for enhancing charts and producing output data sets.

Table 4.13. General Histogram Layout Options

CURVELEGEND=*name* | NONE	specifies LEGEND statement for curves
FORCEHIST	forces creation of histogram
HANGING	constructs hanging histogram
HREF=*value-list*	specifies reference lines perpendicular to the horizontal axis
HREFLABELS='*label1*' . . . '*labeln*'	specifies labels for HREF= lines
MIDPERCENTS	prints table of histogram intervals
MIDPOINTS=*value-list*	lists midpoints for histogram intervals
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
RTINCLUDE	includes right endpoint in interval
SPECLEGEND=*name* | NONE	specifies LEGEND statement for specification limits
VREF=*value-list*	specifies reference lines perpendicular to the vertical axis
VREFLABELS='*label1*' . . . '*labeln*'	specifies labels for VREF= lines
VSCALE=COUNT | PERCENT | PROPORTION	specifies scale for vertical axis

Table 4.14. Options to Create Output Data Sets

OUTFIT=*SAS-data-set*	specifies information on fitted curves
OUTHISTOGRAM=*SAS-data-set*	specifies information on histogram intervals

Table 4.15. Options to Enhance Histograms Produced on Line Printers

HREFCHAR='*character*'	specifies line character for HREF= lines
VREFCHAR='*character*'	specifies line character for VREF= lines

Table 4.16. Options to Enhance Histograms Produced on Graphics Devices

ANNOTATE=*SAS-data-set*	specifies annotate data set
CAXIS=*color*	specifies color for axis
CBARLINE=*color*	specifies color of outlines of histogram bars
CFILL=*color*	specifies color for filling under curve
CFRAME=*color*	specifies color for frame
CHREF=*color*	specifies color for HREF= lines
CTEXT=*color*	specifies color for text
CVREF=*color*	specifies color for VREF= lines
DESCRIPTION='*string*'	specifies description for plot in graphics catalog
FONT=*font*	specifies software font for text
HAXIS=*name*	specifies AXIS statement for horizontal axis
HMINOR=*n*	specifies number of horizontal minor tick marks
LEGEND=*name* | NONE	identifies LEGEND statement
LHREF=*linetype*	specifies line style for HREF= lines
LVREF=*linetype*	specifies line style for VREF= lines
MIDPTAXIS=*name*	specifies name of AXIS statement for horizontal axis
NAME='*string*'	specifies name for plot in graphics catalog
PCTAXIS=*name*|*value-list*	specifies AXIS statement or values for vertical axis
PFILL=*pattern*	specifies pattern for filling under curve
VAXIS=*name*|*value-list*	specifies AXIS statement or values for vertical axis
VMINOR=*n*	specifies number of vertical minor tick marks
WBARLINE=*n*	specifies line thickness for bar outlines

Dictionary of Options

The following entries provide detailed descriptions of options for the HISTOGRAM statement. The marginal notes *Graphics* and *Line Printer* identify options that can be used only with graphics devices and line printers, respectively.

ALPHA=*value*

specifies the shape parameter α for fitted curves requested with the BETA and GAMMA options. Enclose the ALPHA= option in parentheses after the BETA or GAMMA options. If you do not specify a value for α, the procedure calculates a maximum likelihood estimate. See Example 4.1 on page 170. 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*.

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

curves requested by the GAMMA option. Enclose the ALPHADELTA= option in parentheses after the GAMMA option. Iteration continues until the change in α is less than the value specified or until the number of iterations exceeds the value of the MAXITER= option (see page 140). 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 fitted gamma distributions requested with the GAMMA option. Enclose the ALPHAINITIAL= option in parentheses after the GAMMA option. The default value is Thom's approximation of the estimate of α. Refer to Johnson *et al.* (1994).

ANNOTATE=*SAS-data-set*
ANNO=*SAS-data-set*

specifies an input data set containing annotate variables as described in *SAS/GRAPH Software: Reference*. See Example 4.7 on page 182. The ANNOTATE= data set you specify in the HISTOGRAM statement is used for all plots created by the statement. You can also specify an ANNOTATE= data set in the PROC CAPABILITY statement to enhance all plots created by the procedure; for more information, see "ANNOTATE= Data Sets" on page 31.

| *Graphics* |

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

displays a fitted beta density curve on the histogram. The curve equation is

$$p(x) = \begin{cases} \frac{(x-\theta)^{\alpha-1}(\sigma+\theta-x)^{\beta-1}}{B(\alpha,\beta)\sigma^{(\alpha+\beta-1)}}h \times 100\% & \text{for } \theta < x < \theta + \sigma \\ 0 & \text{for } x \leq \theta \text{ or } x \geq \theta + \sigma \end{cases}$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and

$\theta = $ lower threshold parameter (lower endpoint parameter)
$\sigma = $ scale parameter $(\sigma > 0)$
$\alpha = $ shape parameter $(\alpha > 0)$
$\beta = $ shape parameter $(\beta > 0)$
$h = $ width of histogram interval

The beta distribution is bounded below by the parameter θ and above by the value $\theta + \sigma$. You can specify θ and σ using the THETA= and SIGMA= *beta-options*. The following statements fit a beta distribution bounded between 50 and 75, 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 4.2 (page 126) and Table 4.3 (page 126) list options you can specify with the BETA option. See Example 4.1 on page 170. Also see "Formulas for Fitted Curves" on page 149.

BETA=*value*
B=*value*

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 4.1 on page 170.

C=*value*

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 4.2 on page 172. 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 4.5 on page 179.

If you do not specify a value for c, the bandwidth that minimizes the approximate MISE is used for all the estimates.

CAXIS=_color_
CAXES=_color_

> specifies the color used for the axes and tick marks. This option overrides any COLOR= specifications in an AXIS statement. The default is the first color in the device color list.

Graphics

CBARLINE=_color_

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

Graphics

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 fitted Weibull curves requested by the WEIBULL option. Enclose the CDELTA= option in parentheses after the WEIBULL option. 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 (see page 140). The default value is 0.00001. For examples, see the entry for the WEIBULL option.

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 4.5 on page 122 and Output 4.1.1 on page 171. Refer to *SAS/GRAPH Software: Reference* for a list of colors. By default, bars and curve areas are not filled.

Graphics

CFRAME=_color_
CFR=_color_

> specifies the color for the area enclosed by the axes and frame. The area is not filled by default.

Graphics

CHREF=_color_
CH=_color_

> specifies the color for horizontal axis reference lines requested by the HREF= option. The default is the first color in the device color list.

Graphics

CINITIAL=_value_

> specifies the initial value for $\hat{c}$ in the Newton-Raphson approximation of the maximum likelihood estimate of c for Weibull curves requested with the WEIBULL option. Enclose the CINITIAL= option in parentheses after the WEIBULL option. The default value is 1.8 (refer to Johnson *et al.* 1994).

COLOR=_color_

> specifies the color of the density curve. Enclose the COLOR= option in parentheses after the distribution option or the KERNEL option. See Example 4.1 on page 170. If you use the COLOR= option with the KERNEL option, you can specify a list of up to five colors in parentheses for multiple kernel density estimates. If there are more estimates than colors, the last color specified is used for the remaining estimates.

Graphics

CTEXT=*color*

specifies the color for tick mark values and axis labels. The default is the color specified for the CTEXT= option in the GOPTIONS statement. In the absence of a GOPTIONS statement, the default color is the first color in the device color list.

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.

CVREF=*color*

CV=*color*

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

DELTA=*value*

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.

DESCRIPTION='*string***'**

DES='*string***'**

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

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

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

displays a fitted exponential density curve on the histogram. The curve equation is

$$p(x) = \begin{cases} \frac{h \times 100\%}{\sigma} \exp\left(-\left(\frac{x-\theta}{\sigma}\right)\right) & \text{for } x \geq \theta \\ 0 & \text{for } x < \theta \end{cases}$$

where

θ = threshold parameter

σ = scale parameter $(\sigma > 0)$

h = width of histogram interval

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 4.2 (page 126) and Table 4.4 (page 126) list options you can specify with the EXPONENTIAL option. See "Formulas for Fitted Curves" on page 149.

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;
```

Graphics

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=*color*
limits	HISTOGRAM	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, 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 4.1.1 on page 171. Refer to *SAS/GRAPH Software: Reference* 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 *value* 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.

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.

FONT=*font*

specifies a software font for reference line and axis labels. You can also specify fonts for axis labels in an AXIS statement. The FONT= font takes precedence over the FTEXT= font specified in the GOPTIONS statement. Hardware characters are used by default.

Graphics

135

FORCEHIST

forces the creation of a histogram if there is only one unique observation. By default, a histogram is not created if the standard deviation of the data is zero.

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

displays a fitted gamma density curve on the histogram. The curve equation is

$$p(x) = \begin{cases} \frac{h \times 100\%}{\Gamma(\alpha)\sigma} \left(\frac{x-\theta}{\sigma}\right)^{\alpha-1} \exp\left(-\left(\frac{x-\theta}{\sigma}\right)\right) & \text{for } x > \theta \\ 0 & \text{for } x \leq \theta \end{cases}$$

where

θ = threshold parameter
σ = scale parameter $(\sigma > 0)$
α = shape parameter $(\alpha > 0)$
h = width of histogram interval

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 4.2 (page 126) and Table 4.5 (page 127) list the options you can specify with the GAMMA option. See Example 4.2 on page 172 and "Formulas for Fitted Curves" on page 149.

GAMMA=*value*

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.

HANGING
HANG

requests a hanging histogram, as illustrated in Figure 4.6.
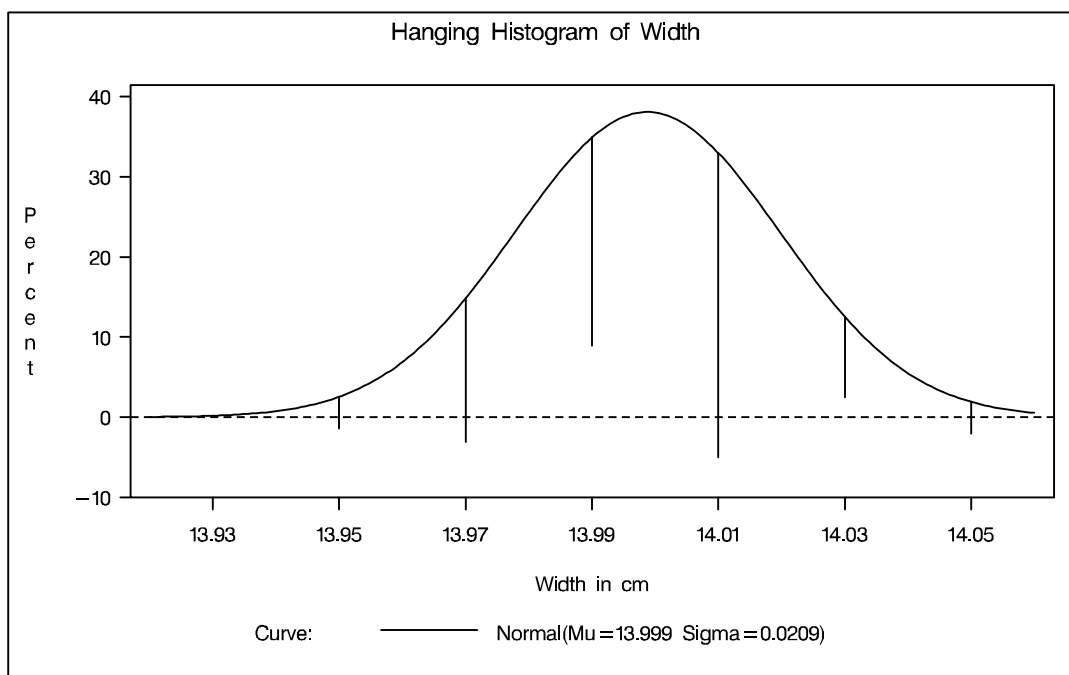
Figure 4.6. Hanging Histogram

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).

HAXIS=*name*

specifies the name of an AXIS statement describing the horizontal axis. You can specify the MIDPTAXIS= option as an alias for the HAXIS= option. See the entry for the MIDPOINTS= option for a syntax example.   *Graphics*

HMINOR=*n*
HM=*n*

specifies the number of minor tick marks between each major tick mark on the horizontal axis. Minor tick marks are not labeled. The default is 0.   *Graphics*

HREF=*value-list*

draws reference lines perpendicular to the horizontal axis at the values specified. See Output 4.1.1 on page 171. Also see the CHREF=, HREFCHAR=, and LHREF= options.

HREFCHAR=*'character'*

specifies the character used to form the lines requested by the HREF= option. The default is the vertical bar (|).   *Line Printer*

137

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. See Output 4.1.1 on page 171.

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 162 for computational details and see Output 4.4.2 on page 179.

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:

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 function
C=	specifies the smoothing parameter
SYMBOL=	specifies the character used to plot the kernel density curve if the histogram is produced on a line printer

> 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 3.1.1 on page 111 and "Kernel Density Estimates" on page 156. By default, kernel density estimates are computed using the AMISE method.

L=*linetype*

specifies the line type used for fitted density curves. If used with the KERNEL option, you can specify a list of up to five line types for multiple kernel density estimates. See the entries for the C= and K= options for details on specifying multiple kernel density estimates. The default is 1, which produces a solid line.

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.

$\boxed{\textit{Graphics}}$

LHREF=*linetype*
LH=*linetype*

specifies the line type for lines requested with the HREF= option. See Output 4.1.1 on page 171. The default is 2, which produces a dashed line.

$\boxed{\textit{Graphics}}$

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

displays a fitted lognormal density curve on the histogram. The curve equation is

$$p(x) = \begin{cases} \frac{h \times 100\%}{\sigma\sqrt{2\pi}(x-\theta)} \exp\left(-\frac{(\log(x-\theta)-\zeta)^2}{2\sigma^2}\right) & \text{for } x > \theta \\ 0 & \text{for } x \leq \theta \end{cases}$$

where

$\theta = \text{threshold parameter}$
$\zeta = \text{scale parameter}$
$\sigma = \text{shape parameter } (\sigma > 0)$
$h = \text{width of histogram interval}$

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 4.2 on page 126 and Table 4.6 on page 127 list options that you can specify with

the LOGNORMAL option. See Example 4.2 on page 172 and "Formulas for Fitted Curves" on page 149.

LVREF=*linetype*
LV=*linetype*

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

MAXITER=*n*

specifies the maximum number of iterations in the Newton-Raphson approximation of the maximum likelihood estimate of α for fitted gamma curves requested with the GAMMA option and c for fitted Weibull curves requested with the WEIBULL option. Enclose the MAXITER= option in parentheses after the GAMMA or WEIBULL option. The default is 20.

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 4.7:

```
proc capability;
   histogram length / midpercents;
run;
```

```
        Midpoint of          Percent of
      Histogram Interval     Observations

          10.02000             12.000
          10.08000             32.000
          10.14000             28.000
          10.20000             18.000
          10.26000              6.000
          10.32000              4.000
```

Figure 4.7. Table of Midpoints and Observed Percentages

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 4.8:

```
proc capability;
   histogram length / gamma(theta=3 midpercents)
run;
```

```
        Bin    -------Percent------
      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
```

Figure 4.8. Table of Observed and Expected Percentages

MIDPOINTS=_value-list_

    lists midpoints for the histogram intervals. The midpoints 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. The same _value-list_ is used for all variables. See Output 4.2.1 on page 173.

    If you specify the MIDPOINTS= option, the range of the 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 must fall between 1.75 and 10.25 (otherwise, a default list of midpoints is used).

    By default, the number of midpoints is determined 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 display the histogram with a graphic device 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 described earlier in this section.   `Graphics`

MU=_value_

    specifies the parameter μ for 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.

NAME=_'string'_

    specifies a name for the plot, up to eight characters, that appears in the PROC GRE-PLAY master menu. The default is 'CAPABILI'.   `Graphics`

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 5.1.1 on page 211.

NOFRAME

suppresses the frame around the subplot area.

NOLEGEND

suppresses legends for specification limits, fitted curves, distribution lines, and hidden observations. See Example 4.6 on page 181. 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 4.4 on page 178.

NOPRINT

suppresses printed output summarizing the fitted curve. Enclose the NOPRINT option in parentheses following the distribution option. See "Customizing a Histogram" on page 122 for an example.

NORMAL<**(***normal-options***)**>

displays a fitted normal density curve on the histogram. The curve equation is

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

where

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

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 4.2 (page 126) and Table 4.7 (page 127) list options that you can specify with the NORMAL option. See Figure 4.4 on page 121 and "Formulas for Fitted Curves" on page 149.

NOSPECLEGEND
NOSPECL

suppresses the portion of the legend for specification limit reference lines. See Figure 4.5 on page 122.

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 164.

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 164.

PCTAXIS=*name|value-list*

is an alias for the VAXIS= option.

<div style="float:right; border:1px solid; padding:2px;">*Graphics*</div>

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 4.9:

```
proc capability;
   histogram length / lognormal(percents=1 3 5 95 97 99);
run;
```

| | ------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

Figure 4.9. Estimated and Observed Quantiles for the Lognormal Curve

PFILL=*pattern*

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

RTINCLUDE

includes the right endpoint of each histogram interval in that interval. By default, the left endpoint is included in the histogram interval.

143

SB$<$**(**S_B*-options* **)**$>$

displays a fitted Johnson S_B density curve on the histogram. The curve equation is

$$
p(x) = \begin{cases} \frac{\delta h \times 100\%}{\sigma\sqrt{2\pi}} \left[\left(\frac{x-\theta}{\sigma}\right)\left(1 - \frac{x-\theta}{\sigma}\right)\right]^{-1} \times \\ \exp\left[-\frac{1}{2}\left(\gamma + \delta \log(\frac{x-\theta}{\theta+\sigma-x})\right)^2\right] & \text{for } \theta < x < \theta + \sigma \\ 0 & \text{for } x \le \theta \text{ or } x \ge \theta + \sigma \end{cases}
$$

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

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 4.2 (page 126) and Table 4.8 (page 127) list options you can specify with the SB option.

144

SCALE=_value_

is an alias for the SIGMA= option for curves requested by the BETA, EXPONENTIAL, GAMMA, SB, SU, and WEIBULL options and an alias for the ZETA= option for curves requested by the LOGNORMAL option. See Example 4.1 on page 170.

SHAPE=_value_

is an alias for the ALPHA= option for curves requested with the GAMMA option, an alias for the SIGMA= option for curves requested with the LOGNORMAL option, and an alias for the C= option for curves requested with the WEIBULL option.

SIGMA=_value_**|EST**

specifies the parameter σ for curves requested with the BETA, EXPONENTIAL, GAMMA, LOGNORMAL, NORMAL, SB, SU, and WEIBULL options. Enclose the SIGMA= option in parentheses after the distribution option. 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=
LOGNORMAL	shape parameter σ	maximum likelihood estimate	SHAPE=
NORMAL	scale parameter σ	standard deviation	
SB	scale parameter σ	1	SCALE=
SU	scale parameter σ	percentile-based estimate	
WEIBULL	scale parameter σ	maximum likelihood estimate	SCALE=

With the BETA distribution option, you can specify SIGMA=EST to request a maximum likelihood estimate for σ. For syntax examples, see the entries for the BETA and NORMAL options.

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.

SU$<$**(**S_U*-options* **)**$>$

displays a fitted Johnson S_U density curve on the histogram. The curve equation is

$$p(x) = \begin{cases} \frac{\delta h \times 100\%}{\sigma\sqrt{2\pi}} \frac{1}{\sqrt{1+((x-\theta)/\sigma)^2}} \times \\ \exp\left[-\frac{1}{2}\left(\gamma + \delta \sinh^{-1}\left(\frac{x-\theta}{\sigma}\right)\right)^2\right] & \text{for } x > \theta \\ 0 & \text{for } x \leq \theta \end{cases}$$

where

$\theta =$ location parameter $(-\infty < \theta < \infty)$
$\sigma =$ scale parameter $(\sigma > 0)$
$\delta =$ shape parameter $(\delta > 0)$
$\gamma =$ shape parameter $(-\infty < \gamma < \infty)$
$h =$ width of histogram interval

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 4.2 (page 126) and Table 4.9 (page 128) list options you can specify with the SU option.

SYMBOL='*character*'

specifies the *character* used to plot the density curve or kernel density curve if the histogram is produced on a line printer. 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 denisty estimates. If there are more estimates than characters, the last character specified is used for the remaining estimates.

THETA=value**|EST**

specifies the lower threshold parameter θ for curves requested with the BETA, EXPO-NENTIAL, GAMMA, LOGNORMAL, 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 4.1 on page 170. The default *value* is zero. If you specify THETA=EST, an estimate is computed for θ.

THRESHOLD=value

is an alias for the THETA= option. See the preceding entry for the THETA= option.

VAXIS=name|value-list

specifies the name of an AXIS statement describing the vertical axis. Alternatively, you can specify a *value-list* for the vertical axis. The PCTAXIS= option is an alias for the VAXIS= option. See Example 4.1 (page 170).

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.

VREF=*value-list*

draws reference lines perpendicular to the vertical axis at the values specified. Also see the CVREF=, LVREF=, and VREFCHAR= options.

VREFCHAR='*character***'**

specifies the character used to form the lines requested by the VREF= option for a line printer. The default is a hyphen (-).   `Line Printer`

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.

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 4.5 on page 122 for an illustration of VSCALE=COUNT. The default is PERCENT.

W=*n*

specifies the width in pixels of the fitted curve or the kernel density estimate curve. Enclose the W= option in parentheses after the distribution option or the KERNEL option (with the KERNEL option, you can specify a list of up to five W= values). For example, the following statements display a normal curve with a width of 3:   `Graphics`

```
proc capability;
   histogram length / normal(w=3);
run;
```

The default is 1.

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

displays a fitted Weibull density curve on the histogram. The curve equation is

$$p(x) = \begin{cases} \frac{ch \times 100\%}{\sigma} \left(\frac{x-\theta}{\sigma}\right)^{c-1} \exp\left(-\left(\frac{x-\theta}{\sigma}\right)^c\right) & \text{for } x > \theta \\ 0 & \text{for } x \leq \theta \end{cases}$$

where

θ = threshold parameter
σ = scale parameter $(\sigma > 0)$
c = shape parameter $(c > 0)$
h = width of histogram interval

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

147

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 4.2 (page 126) and Table 4.10 (page 128) list the options that you can specify with the WEIBULL option. See Example 4.2 on page 172 and "Formulas for Fitted Curves" on page 149.

ZETA=*value*

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.

Details

This section provides details on the following topics:

- formulas for fitted distributions
- formulas for kernel density estimates
- printed output
- OUTFIT= and OUTHISTOGRAM= data sets
- graphical enhancements to histograms

Formulas for Fitted Curves

The following sections provide information on the families of parametric distributions that you can fit with the HISTOGRAM statement. Properties of these distributions are discussed by Johnson *et al.* (1994, 1995).

Beta Distribution

The fitted density function is

$$
p(x) = \begin{cases} \frac{(x-\theta)^{\alpha-1}(\sigma+\theta-x)^{\beta-1}}{B(\alpha,\beta)\sigma^{(\alpha+\beta-1)}} h \times 100\% & \text{for } \theta < x < \theta + \sigma \\ 0 & \text{for } x \leq \theta \text{ or } x \geq \theta + \sigma \end{cases}
$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and

$\theta =$ lower threshold parameter (lower endpoint parameter)
$\sigma =$ scale parameter $(\sigma > 0)$
$\alpha =$ shape parameter $(\alpha > 0)$
$\beta =$ shape parameter $(\beta > 0)$
$h =$ width of histogram interval

Note: This notation is consistent with that of other distributions that you can fit with the HISTOGRAM statement. However, many texts, including Johnson *et al.* (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:

$\sigma = b - a$
$\theta = a$
$\alpha = p$
$\beta = q$

149

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 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

$$
p(x) = \begin{cases} \frac{h \times 100\%}{\sigma} \exp\left(-\left(\frac{x-\theta}{\sigma}\right)\right) & \text{for } x \geq \theta \\ 0 & \text{for } x < \theta \end{cases}
$$

where

$\theta = $ threshold parameter
$\sigma = $ scale parameter $(\sigma > 0)$
$h = $ width of histogram interval

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

$$p(x) = \begin{cases} \frac{h \times 100\%}{\Gamma(\alpha)\sigma} \left(\frac{x-\theta}{\sigma}\right)^{\alpha-1} \exp\left(-\left(\frac{x-\theta}{\sigma}\right)\right) & \text{for } x > \theta \\ 0 & \text{for } x \leq \theta \end{cases}$$

where

θ = threshold parameter
σ = scale parameter $(\sigma > 0)$
α = shape parameter $(\alpha > 0)$
h = width of histogram interval

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

$$p(x) = \begin{cases} \frac{1}{2\Gamma(\frac{\nu}{2})} \left(\frac{x}{2}\right)^{\frac{\nu}{2}-1} \exp(-\frac{x}{2}) & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

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

$$p(x) = \begin{cases} \left[2^{\frac{\nu}{2}-1}\Gamma(\frac{\nu}{2})\right]^{-1} x^{\nu-1} \exp(-\frac{x^2}{2}) & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

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.

Johnson S_B Distribution

The fitted density function is

$$p(x) = \begin{cases} \frac{\delta h \times 100\%}{\sigma\sqrt{2\pi}} \left[\left(\frac{x-\theta}{\sigma}\right)\left(1 - \frac{x-\theta}{\sigma}\right)\right]^{-1} \times \\ \exp\left[-\frac{1}{2}\left(\gamma + \delta \log(\frac{x-\theta}{\theta+\sigma-x})\right)^2\right] & \text{for } \theta < x < \theta + \sigma \\ 0 & \text{for } x \leq \theta \text{ or } x \geq \theta + \sigma \end{cases}$$

151

where

$$\theta = \text{threshold parameter } (-\infty < \theta < \infty)$$
$$\sigma = \text{scale parameter } (\sigma > 0)$$
$$\delta = \text{shape parameter } (\delta > 0)$$
$$\gamma = \text{shape parameter } (-\infty < \gamma < \infty)$$
$$h = \text{width of histogram interval}$$

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:

$$m = x_{3z} - x_z$$
$$n = x_{-z} - x_{-3z}$$
$$p = x_z - x_{-z}$$

It was demonstrated by Slifker and Shapiro (1980) that

$$\frac{mn}{p^2} > 1 \quad \text{for any } S_U \text{ distribution}$$
$$\frac{mn}{p^2} < 1 \quad \text{for any } S_B \text{ distribution}$$
$$\frac{mn}{p^2} = 1 \quad \text{for any } S_L \text{ (lognormal) distribution}$$

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 pareqntheses 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

$$
p(x) = \begin{cases} \dfrac{\delta h \times 100\%}{\sigma\sqrt{2\pi}} \dfrac{1}{\sqrt{1+((x-\theta)/\sigma)^2}} \times \\ \exp\left[-\frac{1}{2}\left(\gamma + \delta \sinh^{-1}\left(\frac{x-\theta}{\sigma}\right)\right)^2\right] & \text{for } x > \theta \\ 0 & \text{for } x \leq \theta \end{cases}
$$

where

$\theta = \text{location parameter } (-\infty < \theta < \infty)$
$\sigma = \text{scale parameter } (\sigma > 0)$
$\delta = \text{shape parameter } (\delta > 0)$
$\gamma = \text{shape parameter } (-\infty < \gamma < \infty)$
$h = \text{width of histogram interval}$

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 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. You can specify the value of z with the FITINTERVAL= option (specified in parentheses after the SU option).

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

$$
\begin{aligned}
\frac{mn}{p^2} &> 1 \quad \text{for any } S_U \text{ distribution} \\
\frac{mn}{p^2} &< 1 \quad \text{for any } S_B \text{ distribution} \\
\frac{mn}{p^2} &= 1 \quad \text{for any } S_L \text{ (lognormal) distribution}
\end{aligned}
$$

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.

Lognormal Distribution

The fitted density function is

$$
p(x) = \begin{cases} \frac{h \times 100\%}{\sigma \sqrt{2\pi}(x-\theta)} \exp\left(-\frac{(\log(x-\theta)-\zeta)^2}{2\sigma^2}\right) & \text{for } x > \theta \\ 0 & \text{for } x \leq \theta \end{cases}
$$

where

$$
\begin{aligned}
\theta &= \text{threshold parameter} \\
\zeta &= \text{scale parameter } (-\infty < \zeta < \infty) \\
\sigma &= \text{shape parameter } (\sigma > 0) \\
h &= \text{width of histogram interval}
\end{aligned}
$$

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.

154

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, normal, 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

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

where

$\mu = \text{mean}$
$\sigma = \text{standard deviation } (\sigma > 0)$
$h = \text{width of histogram interval}$

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.

Weibull Distribution

The fitted density function is

$$p(x) = \begin{cases} \frac{ch \times 100\%}{\sigma}\left(\frac{x-\theta}{\sigma}\right)^{c-1} \exp\left(-\left(\frac{x-\theta}{\sigma}\right)^c\right) & \text{for } x > \theta \\ 0 & \text{for } x \leq \theta \end{cases}$$

where

$\theta = \text{threshold parameter}$
$\sigma = \text{scale parameter } (\sigma > 0)$
$c = \text{shape parameter } (c > 0)$
$h = \text{width of histogram interval}$

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 4.5 on page 179.

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 $\quad K_0(t) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}t^2) \quad$ for $-\infty < t < \infty$

Quadratic $\quad K_0(t) = \frac{3}{4}(1 - t^2) \qquad$ for $|t| \le 1$

Triangular $\quad K_0(t) = 1 - |t| \qquad$ for $|t| \le 1$

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

$$\text{AMISE}(\lambda) = \frac{1}{4}\lambda^4 \left(\int_t t^2 K(t) dt\right)^2 \int_x \left(f''(x)\right)^2 dx + \frac{1}{n\lambda} \int_t K(t)^2 dt$$

156

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.

Printed Output

If you request a fitted parametric distribution, printed output summarizing the fit is produced in addition to the graphical display. Figure 4.10 shows the printed output for a fitted lognormal distribution requested by the following statements:

```
proc capability;
   spec target=14 lsl=13.95 usl=14.05;
   histogram / lognormal(indices midpercents);
run;
```

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 149.

157

```
                    Fitted Lognormal Distribution for width

                 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


                 Capability Indices Based on Lognormal Distribution

                          Index        Value

                          Cp         0.795463
                          CPL        0.776822
                          CPU        0.814021
                          Cpk        0.776822
                          Cpm        0.792237


                 Histogram Bin Percents for Lognormal Distribution

                        Bin    -------Percent------
                     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
```
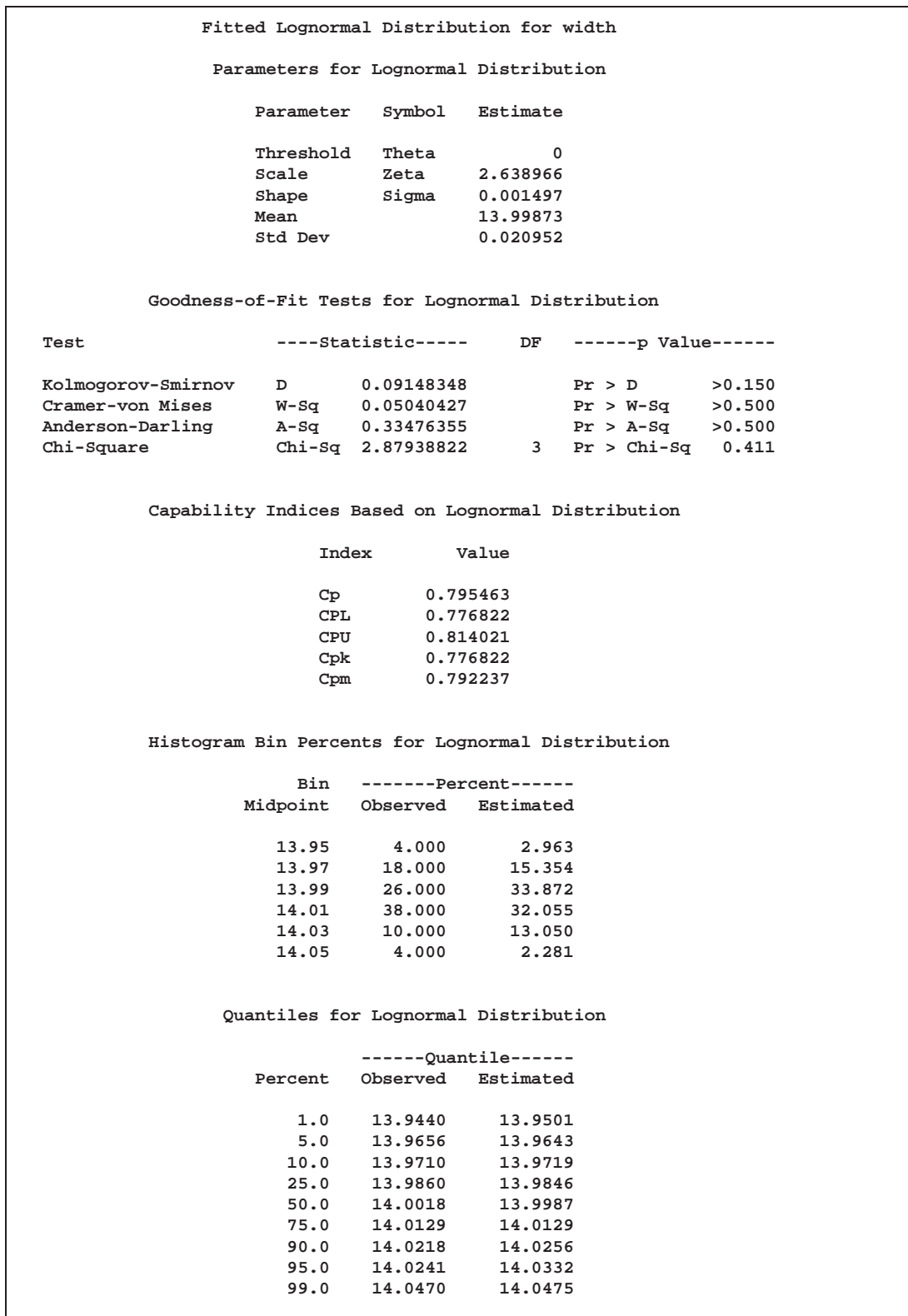
Figure 4.10. Sample Summary of Fitted Distribution

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 percentage in i^{th} histogram interval
E_i = expected percentage in i^{th} 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 percentages in the OUTFIT= data set discussed in "Output Data Sets" on page 164.

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

$$
\begin{aligned}
F_n(x) &= 0, & x &< X_{(1)} \\
F_n(x) &= \tfrac{i}{n}, & X_{(i)} &\le x < X_{(i+1)} \quad i = 1, \ldots, n-1 \\
F_n(x) &= 1, & X_{(n)} &\le x
\end{aligned}
$$

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

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

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

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 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. 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 4.17 summarizes different combinations of estimated parameters for which EDF tests are available.

Note: The threshold (THETA=) parameter for the beta, exponential, gamma, log-normal, and Weibull distributions is assumed to be known. If you do not specify its value, it is assumed to be zero and known. Likewise, the SIGMA= parameter, which determines the upper threshold (SIGMA) for the beta distribution, is assumed to be known; if you do not specify its value, it is assumed to be one. These parameters are not listed in Table 4.17 because they are assumed to be known in all cases, and they do not affect which EDF statistics are computed.

Table 4.17. Availability of EDF Tests

Distribution	Parameters	EDF Tests Available
Beta	α and β unknown	none
	α known, β unknown	none
	α unknown, β known	none
	α and β known	all
Exponential	σ unknown	all
	σ known	all
Gamma	α and σ unknown	none
	α known, σ unknown	none
	α unknown, σ known	none
	α and σ known	all
Lognormal	ζ and σ unknown	all
	ζ known, σ unknown	A^2 and W^2
	ζ unknown, σ known	A^2 and W^2
	ζ and σ known	all
Normal	μ and σ unknown	all
	μ known, σ unknown	A^2 and W^2
	μ unknown, σ known	A^2 and W^2
	μ and σ known	all
Weibull	c and σ unknown	A^2 and W^2
	c known, σ unknown	A^2 and W^2
	c unknown, σ known	A^2 and W^2
	c and σ known	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 on page 157 that produce Figure 4.10. 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 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 $100\alpha^{\text{th}}$ percentile, the generalized capability indices are as follows:

$$C_{pl} = \frac{P_{0.5} - LSL}{P_{0.5} - P_{0.00135}}$$

$$C_{pu} = \frac{USL - P_{0.5}}{P_{0.99865} - P_{0.5}}$$

$$C_p = \frac{USL - LSL}{P_{0.99865} - P_{0.00135}}$$

$$C_{pk} = \min \left(\frac{P_{0.5} - \text{LSL}}{P_{0.5} - P_{0.00135}}, \frac{\text{USL} - P_{0.5}}{P_{0.99865} - P_{0.5}} \right)$$

$$K = 2 \times \frac{\left| \frac{1}{2}(\text{USL} + \text{LSL}) - P_{0.5} \right|}{\text{USL} - \text{LSL}}$$

$$C_{pm} = \frac{\min \left(\frac{T - LSL}{P_{0.5} - P_{0.00135}}, \frac{USL - T}{P_{0.99865} - P_{0.5}} \right)}{\sqrt{1 + \left(\frac{\mu - T}{\sigma} \right)^2}}$$

If the data are normally distributed, these formulas reduce to the formulas for the standard capability indices, which are given on page 46.

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 (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 on page 157 that produce Figure 4.10. 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 140.

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 4.10 is the default list. See the entry for the PERCENTS= option on page 143.

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 Sets

The OUTFIT= data set contains the parameters of fitted density curves, information on chi-square and EDF goodness-of-fit tests, specification limit information, and capability indices based on the fitted distribution. Since 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 4.18 on page 165.

Table 4.18. 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 normal distribution, this is either the value of μ specified with the MU= option or the sample mean. For all other distributions, this is either the value specified with the THRESHOLD= option or zero.
LSL	lower specification limit
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.
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 4.18. *(continued)*

Variable	Description
SCALE	value of scale parameter for fitted distribution. For the normal distribution, this is either the value of σ specified with the SIGMA= option or the sample standard deviation. For all other distributions, this is either the value specified with the SCALE= option or the value estimated by the procedure.
SHAPE1	value of shape parameter for fitted distribution. For distributions without a shape parameter (normal and exponential distributions), _SHAPE1_ is set to missing. For the gamma, lognormal, and Weibull distributions, the value of _SHAPE1_ is either the value specified with the SHAPE= option or the value estimated by the procedure. For the beta distribution, _SHAPE1_ is either the value of α specified with the ALPHA= option or the value estimated by the procedure.
SHAPE2	value of shape parameter for fitted distribution. For the beta distribution, _SHAPE2_ is either the value of β specified with the BETA= option or the value 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 Sets

The OUTHISTOGRAM= data set contains information about histogram intervals. Since 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 since 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 4.19.

Table 4.19. Variables in the OUTHISTOGRAM= Data Set

Variable	Description
CURVE	name of fitted distribution (if requested in HISTOGRAM statement)
EXPPCT	estimated percent of population in histogram interval determined from optional fitted distribution
MIDPT	midpoint of histogram interval
OBSPCT	percent of variable values in histogram interval
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 4.20. ODS Tables Produced with the HISTOGRAM Statement

Table Name	Description	Option
Bins	histogram bins	MIDPERCENTS sub-option with any distribution option, such as NORMAL(MIDPERCENTS)
FitIndices	capability indices computed from fitted distribution	INDICES sub-option 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

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. 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.

Table 4.21. 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 4.22. 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 4.23. 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*
S_B Curve	S_B-options	SYMBOL10 Statement
color	COLOR=*color*	C=*color*
line type	L=*linetype*	L=*linetype*
width	W=*value*	W=*value*
S_U Curve	S_U-options	SYMBOL11 Statement
color	COLOR=*color*	C=*color*
line type	L=*linetype*	L=*linetype*
width	W=*value*	W=*value*

Examples

This section provides advanced examples of the HISTOGRAM statement.

Example 4.1. Fitting a Beta Curve

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:

```
title 'Fitted Beta Distribution of Offsets';
proc capability data=measures noprint;
   specs usl=10.25 lusl=20 cusl=black cright=orange;
   histogram length /
      beta(theta=10 scale=0.5 color=red fill)
      cfill    = yellow
      href     = 10
      hreflabel = 'Lower Bound'
      lhref    = 2
      vaxis    = axis1;
   axis1 label=(a=90 r=0);
   inset n = 'Sample Size'
         beta( pchisq='P-Value' ) / pos=ne cfill=blank;
run;
```

The histogram is shown in Output 4.1.1. 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 4.1.1. Superimposing a Histogram with a Fitted Beta Curve
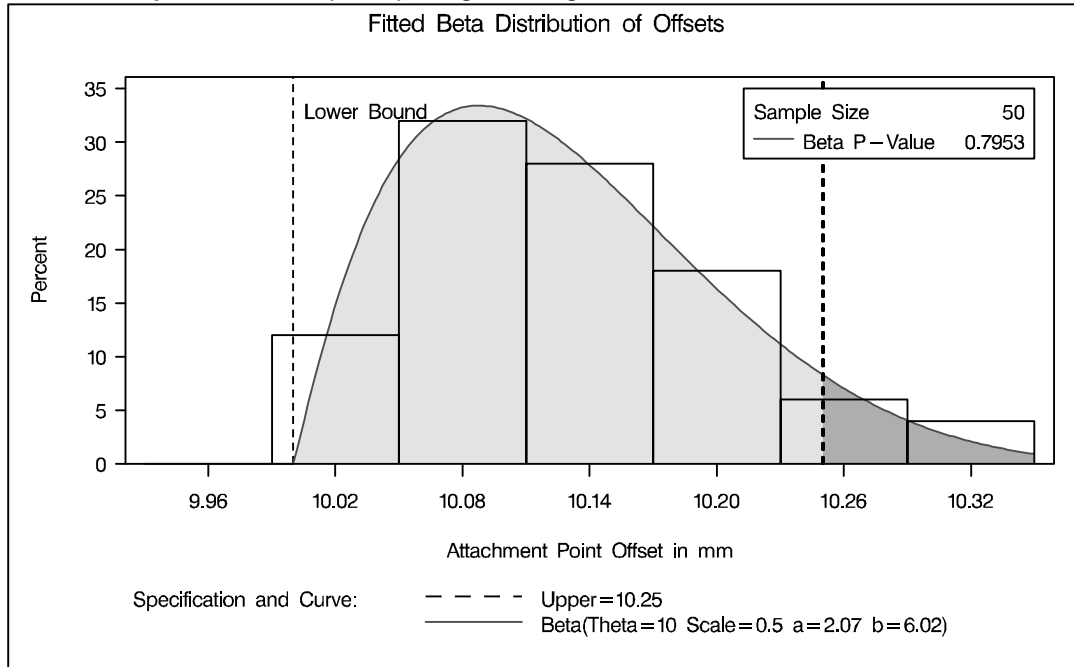
The FILL *beta-option* specifies that the area under the curve is to be filled with the CFILL= color. (If FILL were omitted, the CFILL= color would be used to fill the histogram bars instead.) The CRIGHT= option in the SPEC statement specifies the color under the curve to the right of the upper specification 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 4.1.2. The output tabulates the parameters for the curve, the chi-square goodness-of-fit test whose *p*-value is shown in Output 4.1.1, 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 "Formulas for Fitted Curves" on page 149.

171

Output 4.1.2. Summary of Fitted Beta Distribution

```
                Fitted Beta Distribution of Offsets

                Fitted Beta Distribution for length

                 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

                 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 4.2. Fitting Lognormal, Weibull, and Gamma Curves

See CAPCURV
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 HIS-
TOGRAM 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:

```
title1 'Distribution of Plate Gaps';
proc capability data=plates noprint;
   specs  lsl = 0.3  usl = 0.8
          llsl = 3   lusl = 20;
   histogram gap /
      midpoints = 0.2 to 1.8 by 0.2
      lognormal (l=1)
      weibull   (l=2)
      gamma     (l=8)
      nospeclegend
      vaxis     = axis1;
   inset n mean (5.3) std='Std Dev' (5.3) skewness (5.3) /
      header = 'Summary Statistics'
      pos    = ne;
   axis1  label=(a=90 r=0);
run;
```

The LOGNORMAL, WEIBULL, and GAMMA options superimpose fitted curves on the histogram in Output 4.2.1. The L= options specify distinct line types for the curves. 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 4.2.1. Superimposing a Histogram with Fitted Curves
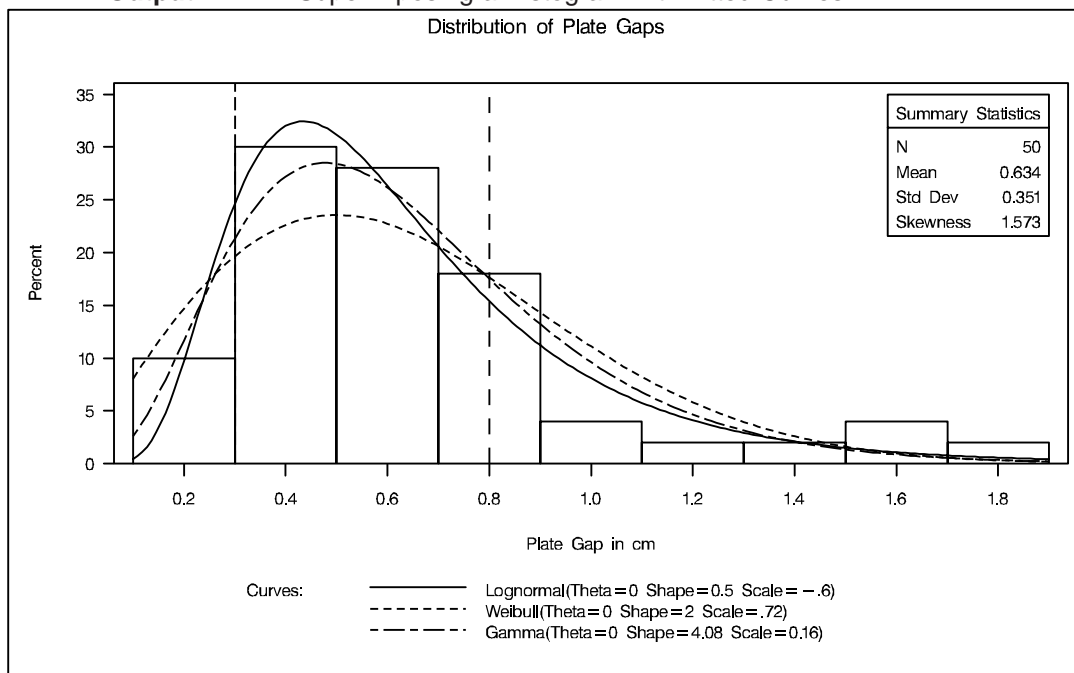
The LOGNORMAL, WEIBULL, and GAMMA options also produce the summaries for the fitted distributions shown in Output 4.2.2, Output 4.2.3, and Output 4.2.4.

Output 4.2.2.   Summary of Fitted Lognormal Distribution

```
                    Distribution of Plate Gaps

            Fitted Lognormal Distribution for gap

            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 7.51762213    6     Pr > Chi-Sq    0.276


            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 4.2.2 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 159 and "EDF Goodness-of-Fit Tests" on page 159 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 4.2.3. Summary of Fitted Weibull Distribution

```
                  Distribution of Plate Gaps

             Fitted Weibull Distribution for gap

             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 15.0252996       6    Pr > Chi-Sq    0.020


               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 4.2.3 provides two EDF goodness-of-fit tests for the Weibull distribution: the Anderson-Darling and the Cramer-von Mises tests. (See Table 4.17 on page 162 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 4.2.4. Summary of Fitted Gamma Distribution

```
                    Distribution of Plate Gaps

                  Fitted Gamma Distribution for gap

                  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------

       Chi-Square           Chi-Sq  12.3075959      6    Pr > Chi-Sq    0.055


                  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 4.2.4 provides a chi-square goodness-of-fit test for the gamma distribution. (None of the EDF tests are currently supported when the scale and shape parameter of the gamma distribution are estimated; see Table 4.17 on page 162.) 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 4.2.2.

Example 4.3. Comparing Goodness-of-Fit Tests

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.

See CAPGOF in the SAS/QC Sample Library

In Example 4.2, 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 using default midpoints (0.3 to 1.8 by 0.3).

```
title1 'Distribution of Plate Gaps';
proc capability data=plates noprint;
   specs  lsl = 0.3  usl = 0.8
          llsl = 2   lusl = 20;
   histogram gap /
      lognormal (l=1)
      nospeclegend
      vaxis=axis1;
   inset n mean (5.3) std='Std Dev' (5.3) skewness (5.3) /
      header = 'Summary Statistics'
      pos    = ne;
   axis1 label=(a=90 r=0);
run;
```

The histogram is shown in Output 4.3.1.

Output 4.3.1. Lognormal Curve Fit with Default Midpoints
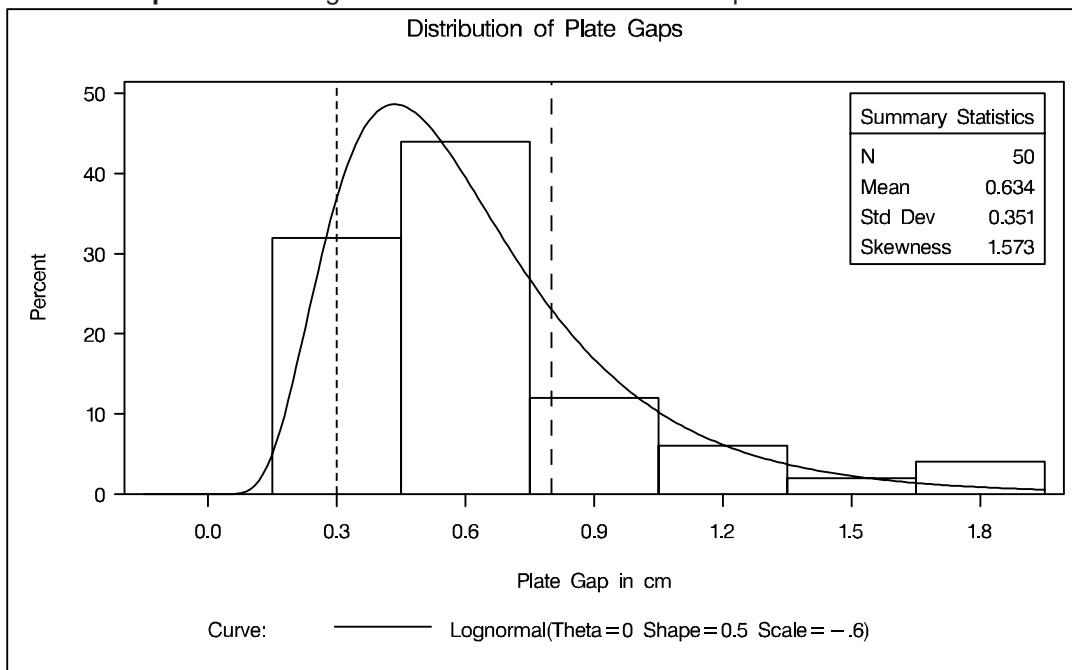
A summary of the lognormal fit is shown in Output 4.3.2. The *p*-value for the chi-square goodness-of-fit test is 0.0822. Since this value is less than 0.10 (a typical cutoff level), the conclusion is that the lognormal distribution is not an appropriate model for the data. This is the *opposite* conclusion drawn from the chi-square test

in Example 4.2, which is based on a different set of midpoints and has a *p*-value of 0.2756 (see Output 4.2.2). Moreover, the results of the EDF goodness-of-fit tests are the same since 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 4.3.2. Printed Output for the Lognormal Curve

```
                    Distribution of Plate Gaps

              Fitted Lognormal Distribution for gap

              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 4.4. Computing Capability Indices for Nonnormal Distributions

See CAPIND
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 4.2 and Example 4.3, 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 4.4.1 and Output 4.4.2:

```
proc capability data=plates;
   specs lsl=0.3 usl=0.8 alpha=0.05;
   histogram gap / lognormal(indices) noplot;
run;
```

The PROC CAPABILITY statement computes the standard capability indices that are shown in Output 4.4.1.

Output 4.4.1. Standard Capability Indices for Variable GAP

```
                    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 ALPHA= option in the SPECS statement requests a Kolmogorov-Smirnov 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 4.2 concluded that the fitted lognormal distribution summarized in Output 4.2.2 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 4.4.2. Formulas and recommendations for these indices are given in "Indices Using Fitted Curves" on page 162.

Output 4.4.2. Fitted Lognormal Distribution Information

```
      Capability Indices Based on Lognormal Distribution

                 Index         Value

                 Cp          0.210804
                 CPL         0.595156
                 CPU         0.124927
                 Cpk         0.124927
```

Example 4.5. Computing Kernel Density Estimates

This example illustrates the use of kernel density estimates to visualize a nonnormal data distribution.

See CAPKERN1 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, which is partially listed in Output 4.5.1.

Output 4.5.1. Partial Listing of the Data Set CHANNEL

```
            Obs     lot      length

              1    Lot 1    0.90979
              2    Lot 1    1.01131
              3    Lot 1    0.95001
              4    Lot 1    1.12591
              5    Lot 1    1.11707
              .      .        .
              .      .        .
              .      .        .
           1224    Lot 3    1.74088
           1225    Lot 3    1.91107
```

179

When you use kernel density estimates to explore a data distribution, you should try several choices for the bandwidth parameter c since 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
                            l = 1 20 2 34
                            color=red);
run;
```

The L= option specifies distinct line types for the curves (the L= values are paired with the C= values in the order listed). The display, shown in Output 4.5.2, 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 4.5.2. Multiple Kernel Density Estimates
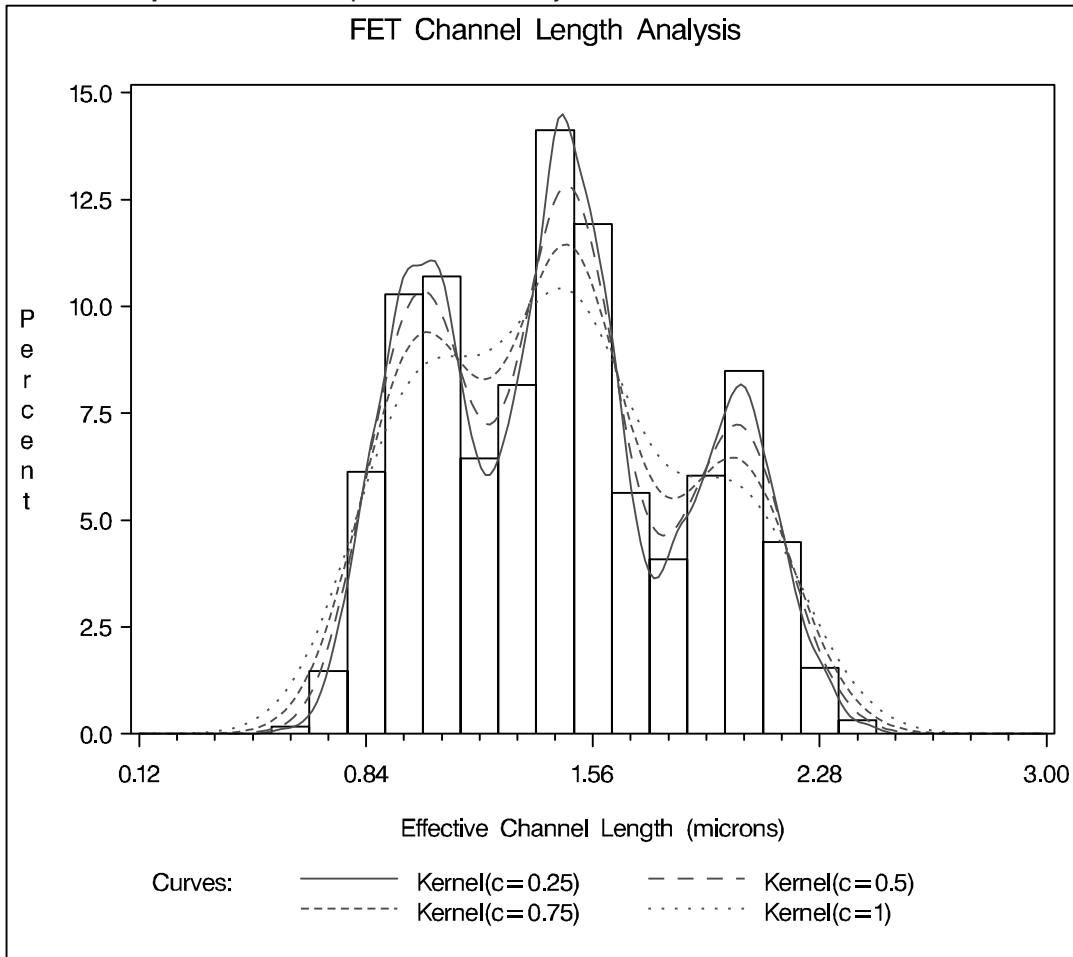
Output 4.5.2 reveals strong trimodality in the data, which are explored further in "Creating a One-Way Comparative Histogram" on page 88.

180

Example 4.6. Fitting a Three-Parameter Lognormal Curve

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 on page 154, 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 4.6.1:

```
title 'Three-Parameter Lognormal Fit';
proc capability data=plastic noprint;
   spec lsl=25 cleft=green;
   histogram strength / lognormal(fill theta=est)
                        cfill=white
                        nolegend;
   inset lsl='LSL' lslpct / cfill=blank pos=nw;
   inset lognormal        / format=6.2 pos=ne;
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 that you can specify THETA=EST as a *Weibull-option* to fit a three-parameter Weibull distribution.

Output 4.6.1. Three-Parameter Lognormal Fit
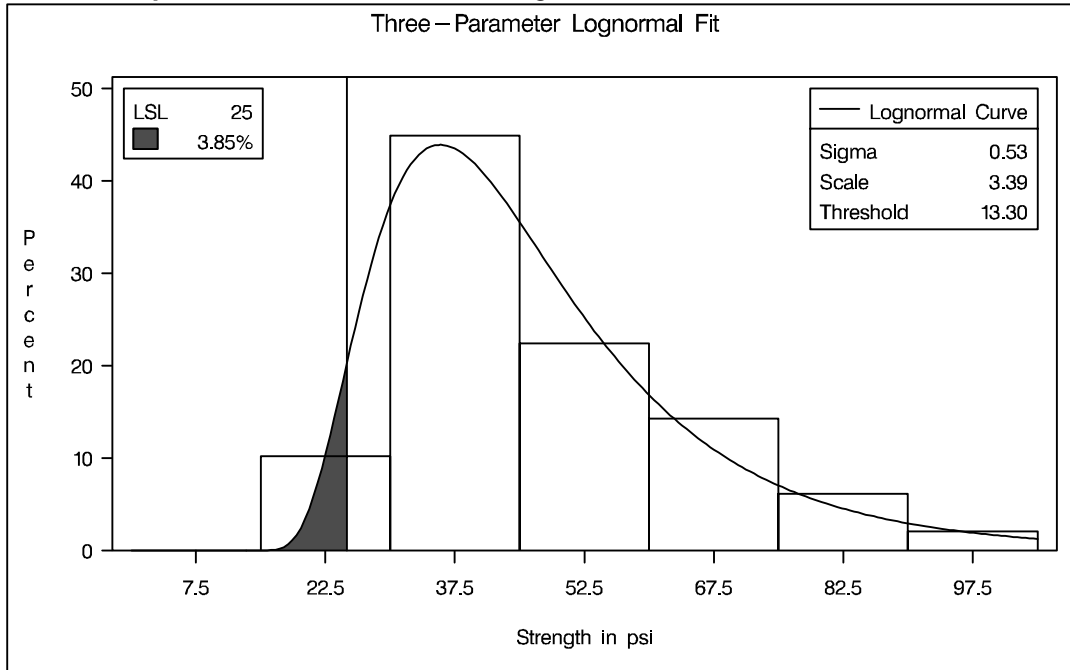
Example 4.7. Annotating a Folded Normal Curve

See FNORM2
in the SAS/QC
Sample Library

This example shows how to display a fitted curve that is not supported by the HIS-
TOGRAM 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

$$\hat{\sigma}_0 \ = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2}{1 + \hat{\theta}^2}}$$
$$\hat{\mu}_0 \ = \hat{\theta} \cdot \hat{\sigma}_0$$

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;

* 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;
```

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 = 0.39894*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 < 0.6374 ) then do;
      print "Warning: Estimates may be unreliable";
      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;
```

The preliminary estimates are saved in the data set PRELIM, as shown in Output 4.7.1.

Output 4.7.1. 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;
```

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 4.7.2.

Output 4.7.2. 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.

```
proc capability data=assembly noprint;
   histogram offset / outfit=out normal(noprint) noplot;
run;
```

Output 4.7.3 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 164.

Output 4.7.3. The OUTFIT= Data Set OUT

```
                        OUTFIT= Data Set OUT

_VAR_   _CURVE_ _LOCATN_ _SCALE_ _CHISQ_ _DF_ _PCHISQ_ _MIDPT1_ _WIDTH_

offset NORMAL    7.62     5.24   31.17    5      0       1.5       3

_MIDPTN_  _EXPECT_  _ESTSTD_  _ADASQ_  _ADP_  _CVMWSQ_  _CVMP_  _KSD_  _KSP_

  22.5     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_ - 0.5*_width_;
   right    = _midptn_ + 0.5*_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 4.7.4:

```
title 'Folded Normal Distribution';
proc capability data=assembly noprint;
   spec usl=27 cusl=black lusl=2 wusl=2;
   histogram offset / annotate = anno
                      cbarline = black
                      cfill    = ligr;
run;
```

Output 4.7.4. Histogram with Annotated Folded Normal Curve
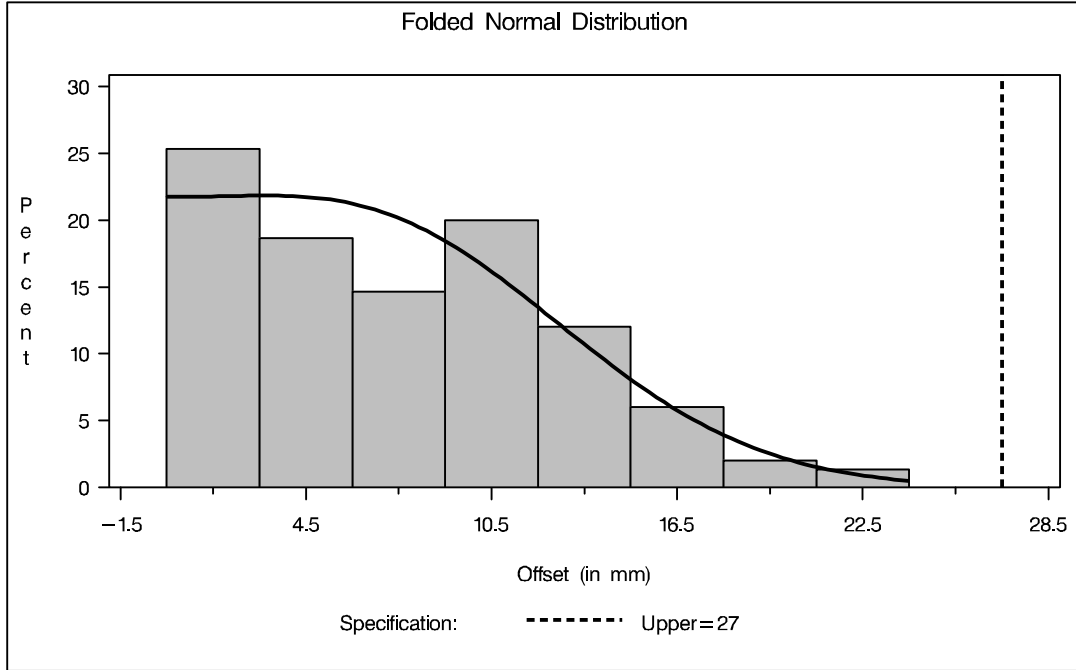
Folded Normal Distribution

SAS/QC® User's Guide, Version 8