# Chapter 8
# PPPLOT Statement

## Chapter Table of Contents

# Chapter 8
# PPPLOT Statement

## Overview

The PPPLOT statement creates a probability-probability plot (also referred to as a P-P plot or percent plot), which compares the empirical cumulative distribution function (ecdf) of a variable with a specified theoretical cumulative distribution function such as the normal. If the two distributions match, the points on the plot form a linear pattern that passes through the origin and has unit slope. Thus, you can use a P-P plot to determine how well a theoretical distribution models a set of measurements.

You can specify one of the following theoretical distributions with the PPPLOT statement:

- beta
- exponential
- gamma
- lognormal
- normal
- Weibull

You can use options in the PPPLOT statement to

- specify or estimate parameters for the theoretical distribution
- request graphical enhancements

**Note:** Probability-probability plots should not be confused with probability plots, which compare a set of ordered measurements with *percentiles* from a specified distribution. You can create probability plots with the PROBPLOT statement.

# Getting Started

The following example illustrates the basic syntax of the PPPLOT statement. For complete details of the PPPLOT statement, see the "Syntax" section on page 254.

## Creating a Normal Probability-Probability Plot

See CAPPP1 in the SAS/QC Sample Library

The distances between two holes cut into 50 steel sheets are measured and saved as values of the variable DISTANCE in the following data set:*

```
data sheets;
   input distance @@;
   label distance='Distance in cm';
   datalines;
 9.80 10.20 10.27  9.70  9.76
10.11 10.24 10.20 10.24  9.63
 9.99  9.78 10.10 10.21 10.00
 9.96  9.79 10.08  9.79 10.06
10.10  9.95  9.84 10.11  9.93
10.56 10.47  9.42 10.44 10.16
10.11 10.36  9.94  9.77  9.36
 9.89  9.62 10.05  9.72  9.82
 9.99 10.16 10.58 10.70  9.54
10.31 10.07 10.33  9.98 10.15
;
```

The cutting process is in statistical control. As a preliminary step in a capability analysis of the process, it is decided to check whether the distances are normally distributed. The following statements create a P-P plot, shown in Figure 8.1, which is based on the normal distribution with mean $\mu = 10$ and standard deviation $\sigma = 0.3$:

```
title 'Normal Probability-Probability Plot for Hole Distance';
proc capability data=sheets noprint;
   ppplot distance / normal(mu=10 sigma=0.3 color=red)
                     square;
run;
```

The NORMAL option in the PPPLOT statement requests a P-P plot based on the normal cumulative distribution function, and the MU= and SIGMA= *normal-options* specify $\mu$ and $\sigma$. Note that a P-P plot is always based on a *completely specified* distribution, in other words, a distribution with specific parameters. In this example, if you did not specify the MU= and SIGMA= *normal-options*, the sample mean and sample standard deviation would be used for $\mu$ and $\sigma$.

*These data are also used to create Q-Q plots in Chapter 10, "QQPLOT Statement." See pages 309–310, 323–324, and 344.

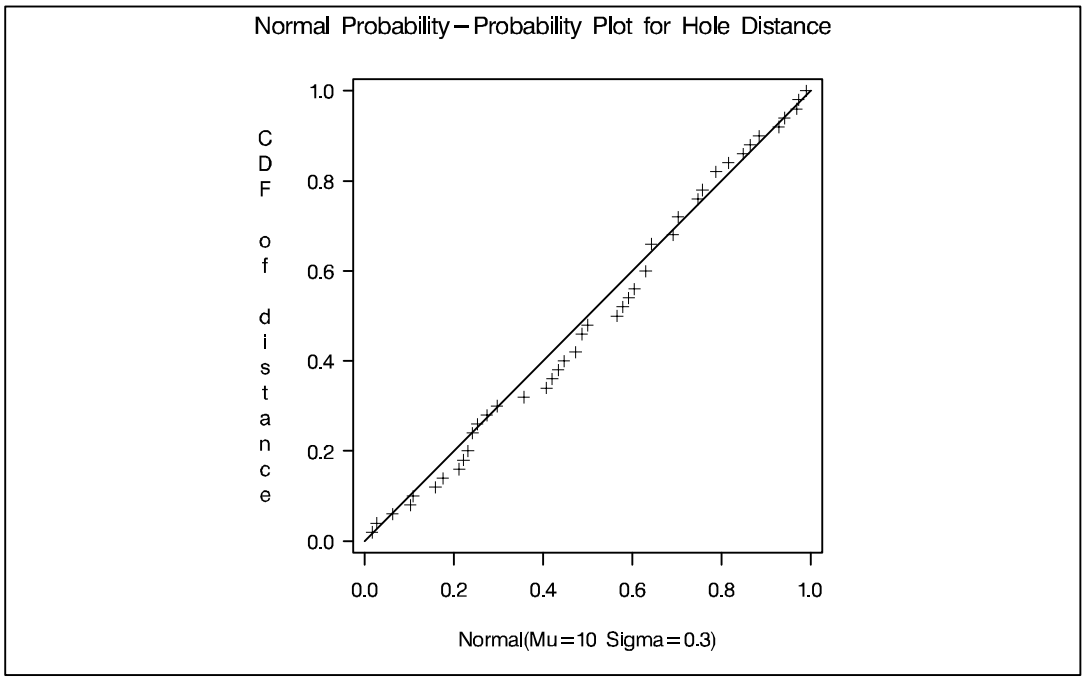


**Figure 8.1.** Normal P-P Plot with Diagonal Reference Line

The linearity of the pattern in Figure 8.1 is evidence that the measurements are normally distributed with mean 10 and standard deviation 0.3. The COLOR= *normal-option* specifies the color for the diagonal reference line, and the SQUARE option displays the plot in a square format.

# Syntax

The syntax for the PPPLOT statement is as follows:

**PPPLOT** < *variables* > < **/** *options* >**;**

You can specify the keyword PP as an alias for PPPLOT, and you can use any number of PPPLOT statements in the CAPABILITY procedure. The components of the PPPLOT statement are described as follows.

*variables*
are the process variables for which to create P-P plots. If you specify a VAR statement, the *variables* must also be listed in the VAR statement. Otherwise, the *variables* can be any numeric variables in the input data set. If you do not specify a list of *variables*, then by default, the procedure creates a P-P plot for each variable listed in the VAR statement or for each numeric variable in the input data set if you do not specify a VAR statement. For example, each of the following PPPLOT statements produces two P-P plots, one for LENGTH and one for WIDTH:

```
proc capability data=measures;
   var length width;
   ppplot;
run;

proc capability data=measures;
   ppplot length width;
run;
```

*options*
specify the theoretical distribution for the plot or add features to the plot. If you specify more than one variable, the options apply equally to each variable. Specify all *options* after the slash (/) in the PPPLOT statement. You can specify only one option naming a distribution, but you can specify any number of other options. The distributions available are the beta, exponential, gamma, lognormal, normal, and Weibull. By default, the procedure produces a P-P plot based on the normal distribution.

In the following example, the NORMAL, MU= and SIGMA= options request a P-P plot based on the normal distribution with mean 10 and standard deviation 0.3. The SQUARE option displays the plot in a square frame, and the CTEXT= option specifies the text color.

```
proc capability data=measures;
   ppplot length width / normal(mu=10 sigma=0.3)
                         square
                         ctext=blue;
run;
```

## Summary of Options

The following tables list the PPPLOT statement options by function. For complete descriptions, see the "Dictionary of Options" section on page 257.

### Distribution Options

Table 8.1 summarizes the options for requesting a specific theoretical distribution.

**Table 8.1.** Options for Specifying the Theoretical Distribution

| | |
|---|---|
| BETA(*beta-options*) | specifies beta P-P plot |
| EXPONENTIAL(*exponential-options*) | specifies exponential P-P plot |
| GAMMA(*gamma-options*) | specifies gamma P-P plot |
| LOGNORMAL(*lognormal-options*) | specifies lognormal P-P plot |
| NORMAL(*normal-options*) | specifies normal P-P plot |
| WEIBULL(*Weibull-options*) | specifies Weibull P-P plot |

Table 8.2 through Table 8.8 summarize options that specify distribution parameters and control the display of the diagonal distribution reference line. Specify these options in parentheses after the distribution option. For example, the following statements use the NORMAL option to request a normal P-P plot:

```
proc capability data=measures;
   ppplot length / normal(mu=10 sigma=0.3 color=red);
run;
```

The MU= and SIGMA= *normal-options* specify $\mu$ and $\sigma$ for the normal distribution, and the COLOR= *normal-option* specifies the color for the line.

**Table 8.2.** Distribution Reference Line Options

| | |
|---|---|
| COLOR=*color* | specifies color of distribution reference line |
| L=*linetype* | specifies line type of distribution reference line |
| NOLINE | suppresses the distribution reference line |
| SYMBOL='*character*' | specifies plotting character for line printer |
| W=$n$ | specifies width of distribution reference line |

**Table 8.3.** Beta-Options

| | |
|---|---|
| ALPHA=*value* | specifies shape parameter $\alpha$ |
| BETA=*value* | specifies shape parameter $\beta$ |
| SIGMA=*value* | specifies scale parameter $\sigma$ |
| THETA=*value* | specifies lower threshold parameter $\theta$ |

**Table 8.4.** Exponential-Options

| | |
|---|---|
| SIGMA=*value* | specifies scale parameter $\sigma$ |
| THETA=*value* | specifies threshold parameter $\theta$ |

**Table 8.5.** Gamma-Options

| | |
|---|---|
| ALPHA=*value* | specifies shape parameter $\alpha$ |
| SIGMA=*value* | specifies scale parameter $\sigma$ |
| THETA=*value* | specifies threshold parameter $\theta$ |

**Table 8.6.** Lognormal-Options

| | |
|---|---|
| SIGMA=*value* | specifies shape parameter $\sigma$ |
| THETA=*value* | specifies threshold parameter $\theta$ |
| ZETA=*value* | specifies scale parameter $\zeta$ |

**Table 8.7.** Normal-Options

| | |
|---|---|
| MU=*value* | specifies mean $\mu$ |
| SIGMA=*value* | specifies standard deviation $\sigma$ |

**Table 8.8.** Weibull-Options

| | |
|---|---|
| C=*value* | specifies shape parameter $c$ |
| SIGMA=*value* | specifies scale parameter $\sigma$ |
| THETA=*value* | specifies threshold parameter $\theta$ |

### *General Options*

Table 8.9 through Table 8.11 list options that control the appearance of the plots.

**Table 8.9.** General Plot Layout Options

| | |
|---|---|
| HREF=*value-list* | specifies reference lines perpendicular to the horizontal axis |
| HREFLABELS='*label1*' . . . '*labeln*' | specifies line labels for HREF= lines |
| NOFRAME | suppresses frame around plotting area |
| SQUARE | displays P-P plot in square format |
| VREF=*value-list* | specifies reference lines perpendicular to the vertical axis |
| VREFLABELS='*label1*' . . . '*labeln*' | specifies line labels for VREF= lines |

**Table 8.10.** Options to Enhance Plots Produced On Line Printers

| | |
|---|---|
| HREFCHAR='*character*' | specifies line character for HREF= lines |
| NOOBSLEGEND | suppresses legend for hidden points |
| PPSYMBOL='*character*' | specifies character for plotted points |
| VREFCHAR='*character*' | specifies line character for VREF= lines |

**Table 8.11.** Options to Enhance Plots Produced On Graphics Devices

| | |
|---|---|
| ANNOTATE=*SAS-data-set* | provides an annotate data set |
| CAXIS=*color* | specifies color for axis |
| 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* | identifies AXIS statement for horizontal axis |
| HMINOR=$n$ | specifies number of minor tick marks on horizontal axis |
| LHREF=*linetype* | specifies line type for HREF= lines |
| LVREF=*linetype* | specifies line type for VREF= lines |
| NAME='*string*' | specifies name for plot in graphics catalog |
| VAXIS=*name* | identifies AXIS statement for vertical axis |
| VMINOR=*value* | specifies number of minor tick marks on vertical axis |

## Dictionary of Options

The following entries provide detailed descriptions of options for the PPPLOT statement. The marginal notes *Graphics* and *Line Printer* identify options that apply to graphics devices and line printers, respectively.

**ALPHA=***value*

specifies the shape parameter $\alpha$ $(\alpha > 0)$ for P-P plots requested with the BETA and GAMMA options. For examples, see the entries for the BETA and GAMMA options.

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

*Graphics*

specifies an input data set containing annotate variables as described in *SAS/GRAPH Software: Reference*. You can use this data set to add features to the plot. The ANNOTATE= data set specified in the PPPLOT 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.

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

creates a beta P-P plot. To create the plot, the $n$ nonmissing observations are ordered from smallest to largest:

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

The $y$-coordinate of the $i^{\text{th}}$ point is the empirical cdf value $\frac{i}{n}$. The $x$-coordinate is the theoretical beta cdf value

$$B_{\alpha\beta}\left(\frac{x_{(i)}-\theta}{\sigma}\right) = \int_{\theta}^{x_{(i)}} \frac{(t-\theta)^{\alpha-1}(\theta+\sigma-t)^{\beta-1}}{B(\alpha,\beta)\sigma^{(\alpha+\beta-1)}}dt$$

where $B_{\alpha\beta}(\cdot)$ is the normalized incomplete beta function, $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ , and

$\theta =$ lower threshold parameter
$\sigma =$ scale parameter $(\sigma > 0)$
$\alpha =$ first shape parameter $(\alpha > 0)$
$\beta =$ second shape parameter $(\beta > 0)$

You can specify $\alpha$, $\beta$, $\sigma$, and $\theta$ with the ALPHA=, BETA=, SIGMA=, and THETA= *beta-options*, as illustrated in the following example:

```
proc capability data=measures;
   ppplot width / beta(theta=1 sigma=2 alpha=3 beta=4);
run;
```

If you do not specify values for these parameters, then by default, $\theta = 0$, $\sigma = 1$, and maximum likelihood estimates are calculated for $\alpha$ and $\beta$.

**IMPORTANT:** If the default unit interval (0,1) does not adequately describe the range of your data, then you should specify THETA=$\theta$ and SIGMA=$\sigma$ so that your data fall in the interval $(\theta, \theta + \sigma)$.

If the data are beta distributed with parameters $\alpha$, $\beta$, $\sigma$, and $\theta$, then the points on the plot for ALPHA=$\alpha$, BETA=$\beta$, SIGMA=$\sigma$, and THETA=$\theta$ tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified beta distribution is a good fit. You can specify the SCALE= option as an alias for the SIGMA= option and the THRESHOLD= option as an alias for the THETA= option.

**BETA=***value*
specifies the shape parameter $\beta$ $(\beta > 0)$ for P-P plots requested with the BETA distribution option. See the preceding entry for the BETA distribution option for an example.

**C=***value*
specifies the shape parameter $c$ $(c > 0)$ for P-P plots requested with the WEIBULL option. See the entry for the WEIBULL option for examples.

**CAXIS=***color*
**CAXES=***color*
specifies the color for the axes. This option overrides any COLOR= specifications in an AXIS statement. The default is the first color in the device color list. *Graphics*

**CFRAME=***color*
**CFR=***color*
specifies a fill color for the area enclosed by the axes and frame. By default, this area is not filled. *Graphics*

**CHREF=***color*
**CH=***color*
specifies the color for reference lines requested by the HREF= option. The default is the first color in the device color list. *Graphics*

**COLOR=***color*
specifies the color for the diagonal reference line. For example, the following statements request a blue line: *Graphics*

```
proc capability data=measures;
   ppplot length / normal(mu=10 sigma=0.25 color=blue);
run;
```

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

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

**CVREF=***color*
**CV=***color*
specifies the color for reference lines requested by the VREF= option. The default is the first color in the device color list. *Graphics*

**DESCRIPTION='***string***'**
**DES='***string***'**
specifies a description, up to 40 characters, that appears in the PROC GREPLAY master menu. The default string is the variable name. *Graphics*

**EXPONENTIAL**<**(***exponential-options***)**>
**EXP**<**(***exponential-options***)**>
creates an exponential P-P plot. To create the plot, the $n$ nonmissing observations are ordered from smallest to largest:

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

The $y$-coordinate of the $i^{\text{th}}$ point is the empirical cdf value $\frac{i}{n}$. The $x$-coordinate is the theoretical exponential cdf value

$$F(x_{(i)}) = 1 - \exp\left(-\frac{x_{(i)}-\theta}{\sigma}\right)$$

where

$\theta =$ threshold parameter
$\sigma =$ scale parameter $(\sigma > 0)$

You can specify $\sigma$ and $\theta$ with the SIGMA= and THETA= *exponential-options*, as illustrated in the following example:

```
proc capability data=measures;
   ppplot width / exponential(theta=1 sigma=2);
run;
```

If you do not specify values for these parameters, then by default, $\theta = 0$ and a maximum likelihood estimate is calculated for $\sigma$.

**IMPORTANT:** Your data must be greater than or equal to the lower threshold $\theta$. If the default $\theta = 0$ is not an adequate lower bound for your data, specify $\theta$ with the THETA= option.

If the data are exponentially distributed with parameters $\sigma$ and $\theta$, the points on the plot for SIGMA=$\sigma$ and THETA=$\theta$ tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified exponential distribution is a good fit. You can specify the SCALE= option as an alias for the SIGMA= option and the THRESHOLD= option as an alias for the THETA= option.

**FONT=***font*

*Graphics*

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

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

creates a gamma P-P plot. To create the plot, the $n$ nonmissing observations are ordered from smallest to largest:

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

The $y$-coordinate of the $i^{\text{th}}$ point is the empirical cdf value $\frac{i}{n}$. The $x$-coordinate is the theoretical gamma cdf value

$$G_\alpha\left(\frac{x_{(i)}-\theta}{\sigma}\right) = \int_\theta^{x_{(i)}} \frac{1}{\sigma\Gamma(\alpha)} \left(\frac{t-\theta}{\sigma}\right)^{\alpha-1} \exp\left(-\frac{t-\theta}{\sigma}\right) dt$$

where $G_\alpha(\cdot)$ is the normalized incomplete gamma function, and

$\theta =$ threshold parameter
$\sigma =$ scale parameter $(\sigma > 0)$
$\alpha =$ shape parameter $(\alpha > 0)$

You can specify $\alpha$, $\sigma$, and $\theta$ with the ALPHA=, SIGMA=, and THETA= *gamma-options*, as illustrated in the following example:

```
proc capability data=measures;
   ppplot width / gamma(alpha=1 sigma=2 theta=3);
run;
```

If you do not specify values for these parameters, then by default, $\theta = 0$ and maximum likelihood estimates are calculated for $\alpha$ and $\sigma$.

**IMPORTANT:** Your data must be greater than or equal to the lower threshold $\theta$. If the default $\theta = 0$ is not an adequate lower bound for your data, specify $\theta$ with the THETA= option.

If the data are gamma distributed with parameters $\alpha$, $\sigma$, and $\theta$, the points on the plot for ALPHA=$\alpha$, SIGMA=$\sigma$, and THETA=$\theta$ tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified gamma distribution is a good fit. You can specify the SHAPE= option as an alias for the ALPHA= option, the SCALE= option as an alias for the SIGMA= option, and the THRESHOLD= option as an alias for the THETA= option.

**HAXIS=***name*
specifies the name of an AXIS statement describing the horizontal axis. *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 also the HREFCHAR=, CHREF=, and LHREF= options.

**HREFCHAR='***character***'**
specifies the character used to form the reference lines requested by the HREF= option for a line printer. The default is the vertical bar (|). *Line Printer*

**HREFLABELS=***'label1' . . . 'labeln'*
**HREFLABEL=***'label1' . . . 'labeln'*
**HREFLAB=***'label1' . . . 'labeln'*
specifies labels for the reference lines requested by the HREF= option. The number of labels must equal the number of lines. Enclose each label in quotes. Labels can be up to 16 characters.

**L=***linetype*

*Graphics*

specifies the line type for the diagonal distribution reference line. For example,

```
proc capability data=measures;
   ppplot length / normal(mu=10 sigma=0.25 l=2);
run;
```

The default is 1, which produces a solid line.

**LHREF=***linetype*
**LH=***linetype*

*Graphics*

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

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

creates a lognormal P-P plot. To create the plot, the $n$ nonmissing observations are ordered from smallest to largest:

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

The $y$-coordinate of the $i^{\text{th}}$ point is the empirical cdf value $\frac{i}{n}$. The $x$-coordinate is the theoretical lognormal cdf value

$$\Phi\left(\frac{\log(x_{(i)}-\theta)-\zeta}{\sigma}\right)$$

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

$\theta =$ threshold parameter
$\zeta =$ scale parameter
$\sigma =$ shape parameter $(\sigma > 0)$

You can specify $\theta$, $\zeta$, and $\sigma$ with the THETA=, ZETA=, and SIGMA= *lognormal-options*, as illustrated in the following example:

```
proc capability data=measures;
   ppplot width / lognormal(theta=1 zeta=2);
run;
```

If you do not specify values for these parameters, then by default, $\theta = 0$ and maximum likelihood estimates are calculated for $\sigma$ and $\zeta$.

**IMPORTANT:** Your data must be greater than the lower threshold $\theta$. If the default $\theta = 0$ is not an adequate lower bound for your data, specify $\theta$ with the THETA= option.

If the data are lognormally distributed with parameters $\sigma$, $\theta$, and $\zeta$, the points on the plot for SIGMA=$\sigma$, THETA=$\theta$, and ZETA=$\zeta$ tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the

point pattern is evidence that the specified lognormal distribution is a good fit. You can specify the SHAPE= option as an alias for the SIGMA=option, the SCALE= option as an alias for the ZETA= option, and the THRESHOLD= option as an alias for the THETA= option.

**LVREF=***linetype*
**LV=***linetype*
specifies the line type for reference lines requested by the VREF= option. The default is 2, which produces a dashed line. *Graphics*

**MU=***value*
specifies the mean $\mu$ for a normal P-P plot requested with the NORMAL option. For examples, see Figure 8.1 on page 253, or Figure 8.2 on page 268 and Figure 8.3 on page 268. By default, the sample mean is used for $\mu$.

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

**NOFRAME**
suppresses the frame around the subplot area.

**NOLINE**
suppresses the diagonal reference line.

**NOOBSLEGEND**
**NOOBSL**
suppresses the legend that indicates the number of hidden observations. *Line Printer*

**NORMAL**<**(***normal-options* **)**>
**NORM**<**(***normal-options* **)**>
creates a normal P-P plot. By default, if you do not specify a distribution option, the procedure displays a normal P-P plot. To create the plot, the $n$ nonmissing observations are ordered from smallest to largest:

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

The $y$-coordinate of the $i^{\text{th}}$ point is the empirical cdf value $\frac{i}{n}$. The $x$-coordinate is the theoretical normal cdf value

$$\Phi\left(\frac{x_{(i)}-\mu}{\sigma}\right) = \int_{-\infty}^{x_{(i)}} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt$$

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

$\mu =$ location parameter or mean
$\sigma =$ scale parameter or standard deviation $(\sigma > 0)$

You can specify $\mu$ and $\sigma$ with the MU= and SIGMA= *normal-options*, as illustrated in the following example:

```
proc capability data=measures;
   ppplot width / normal(mu=1 sigma=2);
run;
```

By default, the sample mean and sample standard deviation are used for $\mu$ and $\sigma$.

If the data are normally distributed with parameters $\mu$ and $\sigma$, the points on the plot for MU=$\mu$ and SIGMA=$\sigma$ tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified normal distribution is a good fit. For an example, see Figure 8.1 on page 253.

**PPSYMBOL='***character***'**

*Line Printer*

specifies the character used to plot the points when the P-P plot is produced on a line printer. The default is the plus sign (+).

**SCALE=***value*

is an alias for the SIGMA= option with the BETA, EXPONENTIAL, GAMMA, and WEIBULL options and an alias for the ZETA= option with the LOGNORMAL option. See the entries for the SIGMA= and ZETA= options.

**SHAPE=***value*

is an alias for the ALPHA= option with the GAMMA option, for the SIGMA= option with the LOGNORMAL option, and for the C= option with the WEIBULL option. See the entries for the ALPHA=, C=, and SIGMA= options.

**SIGMA=***value*

specifies the parameter $\sigma$, where $\sigma > 0$. When used with the BETA, EXPONENTIAL, GAMMA, NORMAL, and WEIBULL options, the SIGMA= option specifies the scale parameter. When used with the LOGNORMAL option, the SIGMA= option specifies the shape parameter. For an example of the SIGMA= option used with the NORMAL option, see Figure 8.1 on page 253.

**SQUARE**

displays the P-P plot in a square frame. The default is a rectangular frame. See Figure 8.1 on page 253 for an example.

**SYMBOL='***character***'**

*Line Printer*

specifies the character used to plot the diagonal reference line for a line printer. The default character is the first letter of the distribution option keyword.

**THETA=***value*

specifies the lower threshold parameter $\theta$ for plots requested with the BETA, EXPONENTIAL, GAMMA, LOGNORMAL, and WEIBULL options.

**THRESHOLD=***value*

is an alias for the THETA= option.

**VAXIS=***name*

*Graphics*

specifies the name of an AXIS statement describing the vertical axis. For an example, see Figure 8.2 on page 268 and Figure 8.3 on page 268.

**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 0. *Graphics*

**VREF=***value-list*

draws reference lines perpendicular to the vertical axis at the values specified. See the entries for the VREFCHAR=, CVREF=, and LVREF= options.

**VREFCHAR='***character***'**

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

**VREFLABELS='***label1' . . . 'labeln'*
**VREFLABEL='***label1' . . . 'labeln'*
**VREFLAB='***label1' . . . 'labeln'*

specifies labels for the reference lines requested by the VREF= option. The number of labels must equal the number of lines. Enclose each label in quotes. Labels can be up to 16 characters.

**W=**$n$

specifies the width in pixels for the diagonal reference line. Specify the W= option in parentheses following a distribution option keyword. For a similar syntax example, see the entry for the L= option. The default is 1. *Graphics*

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

creates a Weibull P-P plot. To create the plot, the $n$ nonmissing observations are ordered from smallest to largest:

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

The $y$-coordinate of the $i^{\text{th}}$ point is the empirical cdf value $\frac{i}{n}$. The $x$-coordinate is the theoretical Weibull cdf value

$$F(x_{(i)}) = 1 - \exp\left(-\left(\frac{x_{(i)}-\theta}{\sigma}\right)^{c}\right)$$

where

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

You can specify $c$, $\sigma$, and $\theta$ with the C=, SIGMA=, and THETA= *Weibull-options*, as illustrated in the following example:

```
proc capability data=measures;
   ppplot width / weibull(theta=1 sigma=2);
run;
```

If you do not specify values for these parameters, then by default $\theta = 0$ and maximum likelihood estimates are calculated for $\sigma$ and $c$.

**IMPORTANT:** Your data must be greater than or equal to the lower threshold $\theta$. If the default $\theta = 0$ is not an adequate lower bound for your data, you should specify $\theta$ with the THETA= option.

If the data are Weibull distributed with parameters $c$, $\sigma$, and $\theta$, the points on the plot for C=$c$, SIGMA=$\sigma$, and THETA=$\theta$ tend to fall on or near the diagonal line $y = x$, which is displayed by default. Agreement between the diagonal line and the point pattern is evidence that the specified Weibull distribution is a good fit. You can specify the SHAPE= option as an alias for the C= option, the SCALE= option as an alias for the SIGMA= option, and the THRESHOLD= option as an alias for the THETA= option.

**ZETA=***value*

specifies a value for the scale parameter $\zeta$ for lognormal P-P plots requested with the LOGNORMAL option.

# Details

This section provides details on the following topics:

- construction and interpretation of P-P plots
- comparison of P-P plots with Q-Q plots
- distributions supported by the PPPLOT statement
- graphical enhancements of P-P plots

## Construction and Interpretation of P-P Plots

A P-P plot compares the empirical cumulative distribution function (ecdf) of a variable with a specified theoretical cumulative distribution function $F(\cdot)$. The ecdf, denoted by $F_n(x)$, is defined as the proportion of nonmissing observations less than or equal to $x$, so that $F_n(x_{(i)}) = \frac{i}{n}$.

To construct a P-P plot, the $n$ nonmissing values are first sorted in increasing order:

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

Then the $i^{\text{th}}$ ordered value $x_{(i)}$ is represented on the plot by the point whose $x$-coordinate is $F(x_{(i)})$ and whose $y$-coordinate is $\frac{i}{n}$.

Like Q-Q plots and probability plots, P-P plots can be used to determine how well a theoretical distribution models a data distribution. If the theoretical cdf reasonably models the ecdf in all respects, including location and scale, the point pattern on the P-P plot is linear through the origin and has unit slope.

Unlike Q-Q and probability plots, P-P plots are not invariant to changes in location and scale. For example, the data in the "Getting Started" section on page 252 are reasonably described by a normal distribution with mean 10 and standard deviation 0.3. It is instructive to display these data on normal P-P plots with a different mean and standard deviation, as created by the following statements:

See CAPPP2 in the SAS/QC Sample Library

```
title 'Normal Probability-Probability Plot for Hole Distance';
proc capability data=sheets noprint;
   ppplot distance / normal(mu=9.5 sigma=0.3 color=red)
                     square
                     vaxis=axis1;
   ppplot distance / normal(mu=10 sigma=0.5 color=red)
                     square
                     vaxis=axis1;
   axis1 label=(a=90 r=0);
run;
```

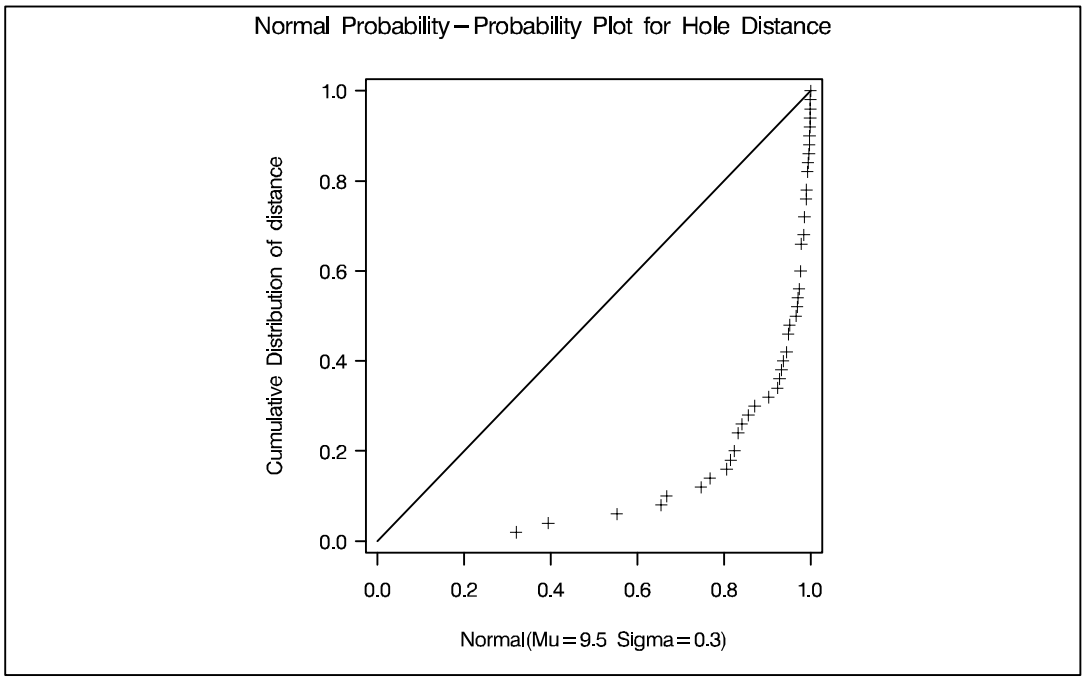


**Figure 8.2.** Normal P-P Plot with Mean Specified Incorrectly

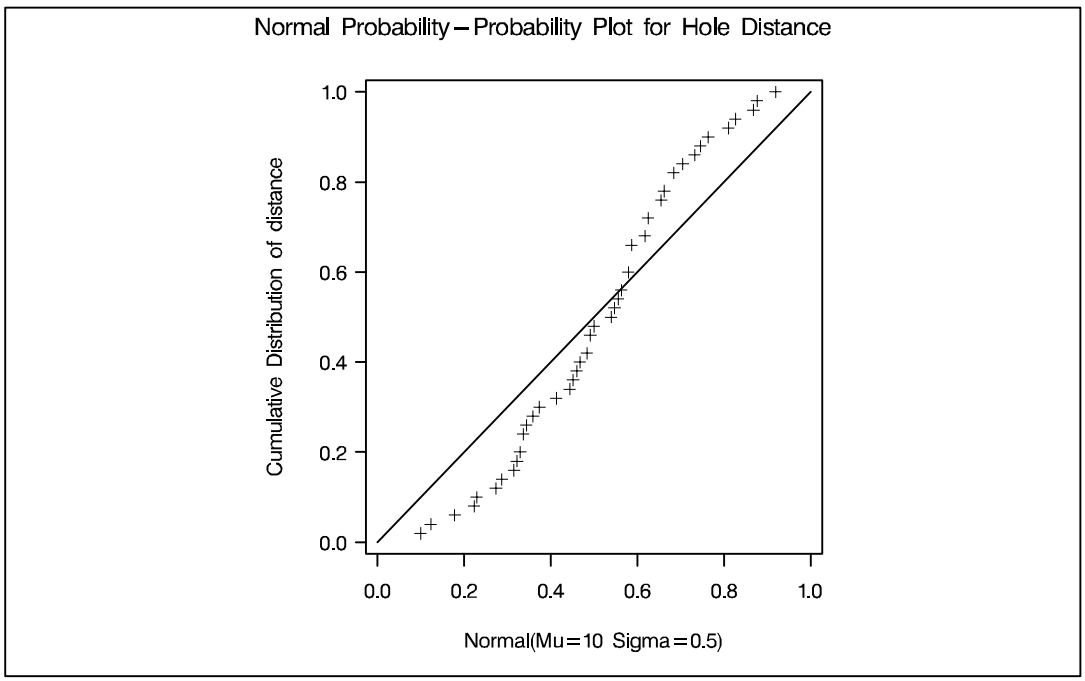


**Figure 8.3.** Normal P-P Plot with Standard Deviation Specified Incorrectly

Specifying a mean of 9.5 instead of 10 results in the plot shown in Figure 8.2, while specifying a standard deviation of 0.5 instead of 0.3 results in the plot shown in Figure 8.3. Both plots clearly reveal the model misspecification.

## Comparison of P-P Plots and Q-Q Plots

A P-P plot compares the empirical cumulative distribution function of a data set with a specified theoretical cumulative distribution function $F(\cdot)$. A Q-Q plot compares the quantiles of a data distribution with the quantiles of a standardized theoretical distribution from a specified family of distributions. There are three important differences in the way P-P plots and Q-Q plots are constructed and interpreted:

- The construction of a Q-Q plot does not require that the location or scale parameters of $F(\cdot)$ be specified. The theoretical quantiles are computed from a standard distribution within the specified family. A linear point pattern indicates that the specified family reasonably describes the data distribution, and the location and scale parameters can be estimated visually as the intercept and slope of the linear pattern. In contrast, the construction of a P-P plot requires the location and scale parameters of $F(\cdot)$ to evaluate the cdf at the ordered data values.
- The linearity of the point pattern on a Q-Q plot is unaffected by changes in location or scale. On a P-P plot, changes in location or scale do not necessarily preserve linearity.
- On a Q-Q plot, the reference line representing a particular theoretical distribution depends on the location and scale parameters of that distribution, having intercept and slope equal to the location and scale parameters. On a P-P plot, the reference line for any distribution is always the diagonal line $y = x$.

Consequently, you should use a Q-Q plot if your objective is to compare the data distribution with a family of distributions that vary only in location and scale, particularly if you want to estimate the location and scale parameters from the plot.

An advantage of P-P plots is that they are discriminating in regions of high probability density, since in these regions the empirical and theoretical cumulative distributions change more rapidly than in regions of low probability density. For example, if you compare a data distribution with a particular normal distribution, differences in the middle of the two distributions are more apparent on a P-P plot than on a Q-Q plot.

For further details on P-P plots, refer to Gnanadesikan (1997) and Wilk and Gnanadesikan (1968).

## Summary of Theoretical Distributions

You can use the PPPLOT statement to request P-P plots based on the theoretical distributions summarized in the following table:

**Table 8.12.** Distributions and Parameters

| Family | Distribution Function $F(x)$ | Range | Parameters | | |
|---|---|---|---|---|---|
| | | | Location | Scale | Shape |
| Beta | $\int_{\theta}^{x} \frac{(t-\theta)^{\alpha-1}(\theta+\sigma-t)^{\beta-1}}{B(\alpha,\beta)\sigma^{(\alpha+\beta-1)}}dt$ | $\theta < x < \theta+\sigma$ | $\theta$ | $\sigma$ | $\alpha$, $\beta$ |
| Exponential | $1-\exp\left(-\frac{x-\theta}{\sigma}\right)$ | $x \geq \theta$ | $\theta$ | $\sigma$ | |
| Gamma | $\int_{\theta}^{x} \frac{1}{\sigma\Gamma(\alpha)}\left(\frac{t-\theta}{\sigma}\right)^{\alpha-1}\exp\left(-\frac{t-\theta}{\sigma}\right)dt$ | $x > \theta$ | $\theta$ | $\sigma$ | $\alpha$ |
| Lognormal | $\int_{\theta}^{x} \frac{1}{\sigma\sqrt{2\pi}(t-\theta)}\exp\left(-\frac{(\log(t-\theta)-\zeta)^2}{2\sigma^2}\right)dt$ | $x > \theta$ | $\theta$ | $\zeta$ | $\sigma$ |
| Normal | $\int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)dt$ | all $x$ | $\mu$ | $\sigma$ | |
| Weibull | $1-\exp\left(-\left(\frac{x-\theta}{\sigma}\right)^{c}\right)$ | $x > \theta$ | $\theta$ | $\sigma$ | $c$ |

You can request these distributions with the BETA, EXPONENTIAL, GAMMA, LOGNORMAL, NORMAL, and WEIBULL options, respectively. If you do not specify a distribution option, a normal P-P plot is created.

To create a P-P plot, you must provide all of the parameters for the theoretical distribution. If you do not specify parameters, then default values or estimates are substituted, as summarized by the following table:

**Table 8.13.** Defaults for Parameters

| Family | Default Values | Estimated Values |
|---|---|---|
| Beta | $\theta = 0$, $\sigma = 1$ | maximum likelihood estimates for $\alpha$ and $\beta$ |
| Exponential | $\theta = 0$ | maximum likelihood estimate for $\sigma$ |
| Gamma | $\theta = 0$ | maximum likelihood estimates for $\sigma$ and $\alpha$ |
| Lognormal | $\theta = 0$ | maximum likelihood estimates for $\sigma$ and $\zeta$ |
| Normal | None | sample estimates for $\mu$ and $\sigma$ |
| Weibull | $\theta = 0$ | maximum likelihood estimates for $\sigma$ and $c$ |

## Specification of Symbol Markers

If you produce the P-P plot on a graphics device, you can use options in the SYMBOL1 statement to specify the appearance of the symbol marker for the points. The V= option specifies the symbol, the C= option specifies the color, and the H= option specifies the height. Refer to *SAS/GRAPH Software: Reference* for details concerning these options. If you produce the plot on a line printer, you can use the PPSYMBOL= option in the PPPLOT statement to specify the character used to plot the points.

## Specification of the Distribution Reference Line

If you produce the P-P plot on a graphics device, you can control the color, type, and width of the diagonal distribution reference line by specifying the COLOR=, L=, and W= options in parentheses after the distribution option in the PPPLOT statement. Alternatively, you can control these features with the C=, L=, and W= options in the SYMBOL4 statement. Refer to *SAS/GRAPH Software: Reference* for details concerning these options. If you produce the plot on a line printer, you can specify the character used for the line with the SYMBOL= option enclosed in parentheses after the distribution option in the PPPLOT statement.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/QC® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999. 1994 pp.

**SAS/QC® User's Guide, Version 8**



ISBN 1–58025–493–4



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

1st printing, October 1999