Chapter 38
Distribution Analyses

Chapter Table of Contents

Chapter 38
Distribution Analyses

Choosing **Analyze:Distribution (Y)** gives you access to a variety of *distribution analyses*. For nominal **Y** variables, you can generate bar charts, mosaic plots, and frequency counts tables.

For interval variables, you can generate univariate statistics, such as moments, quantiles, confidence intervals for the mean, standard deviation, and variance, tests for location, frequency counts, robust measures of the scale, tests for normality, and trimmed and Winsorized means.

You can use parametric estimation based on normal, lognormal, exponential, or Weibull distributions to estimate density and cumulative distribution functions and to generate quantile-quantile plots. You can also generate nonparametric density estimates based on normal, triangular, or quadratic kernels.

You can use Kolmogorov statistics to generate confidence bands for the cumulative distribution and to test the hypothesis that the data are from a completely specified distribution with known parameters. You can also test the hypothesis that the data are from a specific family of distributions but with unknown parameters.

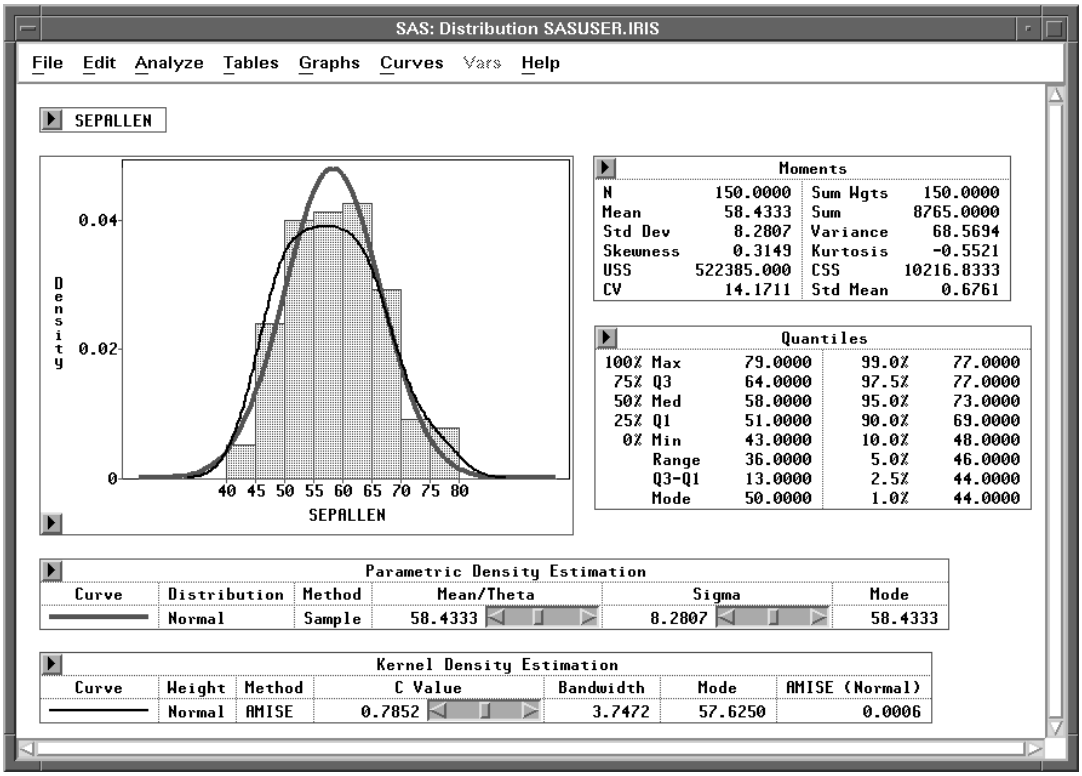


Figure 38.1. Distribution Analysis

Parametric Distributions

A parametric family of distributions is a collection of distributions with a known form that is indexed by a set of quantities called *parameters*. Methods based on parametric distributions of normal, lognormal, exponential, and Weibull are available in a distribution analysis. This section describes the details of each of these distributions. Use of these distributions is described in the sections "Graphs" and "Curves" later in this chapter.

You can use both the density function and the cumulative distribution function to identify the distribution. The density function is often more easily interpreted than the cumulative distribution function.

Normal Distribution

The normal distribution has the probability density function

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

where μ is the mean and σ is the scale parameter.

The cumulative distribution function is

$$F(y) = \Phi\left(\frac{y-\mu}{\sigma}\right)$$

where the function Φ is the cumulative distribution function of the standard normal variable: $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-u^2/2\right) du$

Lognormal Distribution

The lognormal distribution has the probability density function

$$f(y) = \frac{1}{y-\theta} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{\log(y-\theta)-\zeta}{\sigma}\right)^2\right) \quad \text{for } y > \theta$$

where θ is the threshold parameter, ζ is the scale parameter, and σ is the shape parameter.

The cumulative distribution function is

$$F(y) = \Phi\left(\frac{\log(y-\theta)-\zeta}{\sigma}\right) \quad \text{for } y > \theta$$

Exponential Distribution

The exponential distribution has the probability density function

$$f(y) = \frac{1}{\sigma} \exp\left(-\frac{y-\theta}{\sigma}\right) \quad \text{for } y > \theta$$

where θ is the threshold parameter and σ is the scale parameter.

The cumulative distribution function is

$$F(y) = 1 - \exp\left(-\frac{y-\theta}{\sigma}\right) \quad \text{for } y > \theta$$

Weibull Distribution

The Weibull distribution has the probability density function

$$f(y) = \frac{c}{\sigma}\left(\frac{y-\theta}{\sigma}\right)^{c-1} \exp\left(-\left(\frac{y-\theta}{\sigma}\right)^{c}\right) \quad \text{for } y > \theta, c > 0$$

where θ is the threshold parameter, σ is the scale parameter, and c is the shape parameter.

The cumulative distribution function is

$$F(y) = 1 - \exp\left(-\left(\frac{y-\theta}{\sigma}\right)^{c}\right) \quad \text{for } y > \theta$$

Variables

To create a distribution analysis, choose **Analyze:Distribution (Y)**. If you have already selected one or more variables, a distribution analysis for each selected variable appears. If you have not selected any variables, a variables dialog appears.

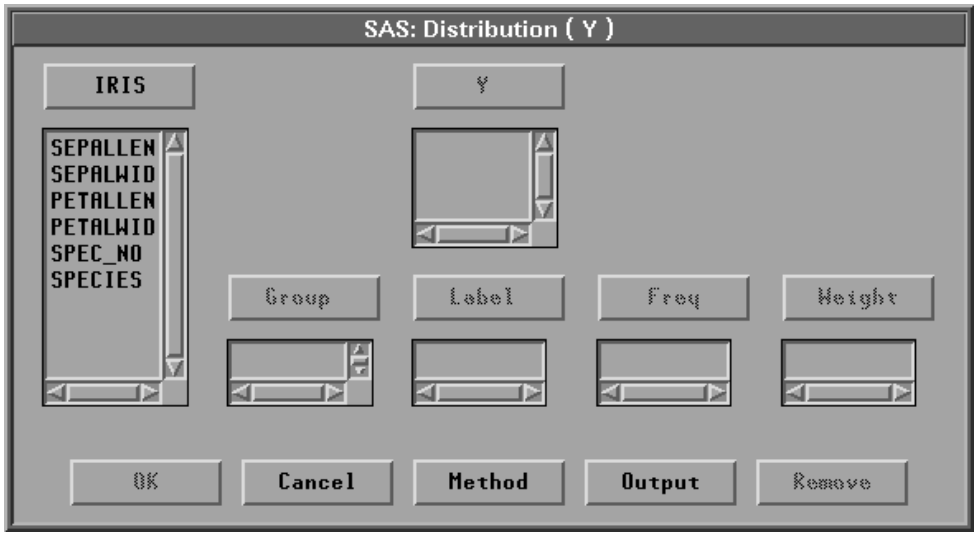


Figure 38.2. Distribution Variables Dialog

Select at least one **Y** variable for each distribution analysis.

You can select one or more **Group** variables if you have grouped data. This creates one distribution analysis for each group.

You can select a **Label** variable to label observations in the plots.

You can select a **Freq** variable. If you select a **Freq** variable, each observation is assumed to represent n observations, where n is the value of the **Freq** variable.

You can select a **Weight** variable to specify relative weights for each observation in the analysis. The details of weighted analyses are explained in the individual sections of this chapter.

Method

Observations with missing values for a **Y** variable are not used in the analysis for that variable. Observations with **Weight** or **Freq** values that are missing or that are less than or equal to zero are not used. Only the integer part of **Freq** values is used.

The following notation is used in the rest of this chapter:

- n is the number of nonmissing values.
- y_i is the ith observed nonmissing value.
- $y_{(i)}$ is the ith ordered nonmissing value, $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$.
- $\overline{y}$ is the sample mean, $\sum_i y_i/n$.
- d is the variance divisor.
- s^2 is the sample variance, $\sum_i (y_i - \overline{y})^2/d$.
- z_i is the standardized value, $(y_i - \overline{y})/s$.

The summation $\sum_i$ represents a summation of $\sum_{i=1}^{n}$.

Based on the variance definition, vardef, the variance divisor d is computed as

- $d = n - 1$ for vardef=**DF**, degrees of freedom
- $d = n$ for vardef=**N**, number of observations

The skewness is a measure of the tendency of the deviations from the mean to be larger in one direction than in the other. The sample skewness is calculated as

- $g_1 = c_{3n} \sum_i z_i^3$ for vardef=**DF**
- $g_1 = \frac{1}{n} \sum_i z_i^3$ for vardef=**N**

where $c_{3n} = \frac{n}{(n-2)} \frac{1}{(n-1)}$.

The kurtosis is primarily a measure of the heaviness of the tails of a distribution. The sample kurtosis is calculated as

- $g_2 = c_{4n} \sum_i z_i^4 - 3c_n$ for vardef=**DF**
- $g_2 = \frac{1}{n} \sum_i z_i^4 - 3$ for vardef=**N**

where $c_{4n} = \frac{n(n+1)}{(n-2)(n-3)} \frac{1}{(n-1)}$ and $c_n = \frac{(n-1)^2}{(n-2)(n-3)}$.

When the observations are independently distributed with a common mean and unequal variances, $\sigma_i^2 = \sigma^2/w_i$, where w_i are individual weights, weighted analyses may be appropriate. You select a **Weight** variable to specify relative weights for each observation in the analysis.

The following notation is used in weighted analyses:

- w_i is the weight associated with y_i.
- $w_{(i)}$ is the weight associated with $y_{(i)}$.
- $\overline{w}$ is the average observation weight, $\sum_i w_i/n$.
- $\overline{y}_w$ is the weighted sample mean, $\sum_i w_i y_i / \sum_i w_i$.
- s_w^2 is the weighted sample variance, $\sum_i w_i(y_i - \overline{y}_w)^2/d$.
- z_{wi} is the standardized value, $(y_i - \overline{y}_w)/(s_w/\sqrt{w_i})$.

In addition to vardef=DF and vardef=N, the variance divisor is also computed as

- $d = \sum_i w_i - 1$ for vardef=**WDF**, sum of weights minus 1
- $d = \sum_i w_i$ for vardef=**WGT**, sum of weights

With $Var(y_i) = \sigma_i^2 = \sigma^2/w_i$, $Var(\overline{y}_w) = \sigma^2 / \sum_i w_i$ and the expected value

$$E\left(\sum_i w_i(y_i - \overline{y}_w)^2\right) = E\left(\sum_i w_i(y_i - \mu)^2 - \sum_i w_i(\overline{y}_w - \mu)^2\right) = (n-1)\sigma^2$$

† **Note:** The use of vardef=WDF/WGT may not be appropriate since it is the weighted average of individual variances, σ_i^2, which have unequal expected values.

For vardef=**DF/N**, s_w^2 is the variance of observations with unit weight and may not be informative in the weighted plots of parametric normal distributions. SAS/INSIGHT software uses the weighted sample variance for an observation with average weight, $s_a^2 = s_w^2/\overline{w}$, to replace s_w^2 in the plots.

The weighted skewness is computed as

- $g_{w1} = c_{3n} \sum_i z_3^{wi} = c_{3n} \sum_i w_i^{\frac{3}{2}} (\frac{y_i - \overline{y}}{s_w})^3$ for **DF**
- $g_{w1} = \frac{1}{n} \sum_i z_3^{wi} = \frac{1}{n} \sum_i w_i^{\frac{3}{2}} (\frac{y_i - \overline{y}}{s_w})^3$ for **N**

The weighted kurtosis is computed as

- $g_{w2} = c_{4n} \sum_i z_4^{wi} - 3c_n = c_{4n} \sum_i w_i^2 (\frac{y_i - \overline{y}}{s_w})^4 - 3c_n$ for **DF**
- $g_{w2} = \frac{1}{n} \sum_i z_4^{wi} - 3 = \frac{1}{n} \sum_i w_i^2 (\frac{y_i - \overline{y}}{s_w})^4 - 3$ for **N**

The formulations are invariant under the transformation $w_i^* = cw_i, c > 0$. The sample skewness and kurtosis are set to missing if vardef=**WDF** or vardef=**WGT**.

To view or change the divisor d used in the calculation of variances, or to view or change the use of observations with missing values, click on the **Method** button from the variables dialog to display the method options dialog.

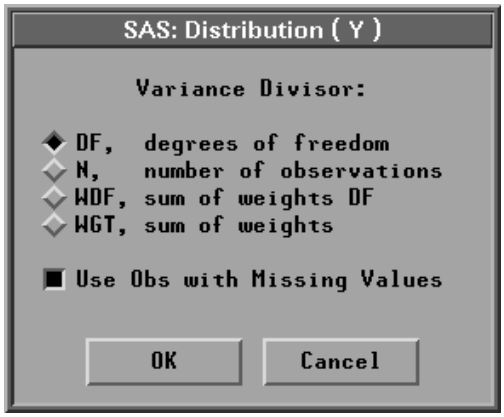


Figure 38.3. Distribution Method Options Dialog

By default, SAS/INSIGHT software uses vardef=**DF, degrees of freedom** to compute the variance divisor.

When multiple **Y** variables are analyzed, and some **Y** variables have missing values, the **Use Obs with Missing Values** option uses all observations with nonmissing values for the **Y** variable being analyzed. If the option is turned off, observations with missing values for *any* **Y** variable are not used for any analysis.

Output

To view or change the options associated with your distribution analysis, click on the **Output** button from the variables dialog. This displays the output options dialog.

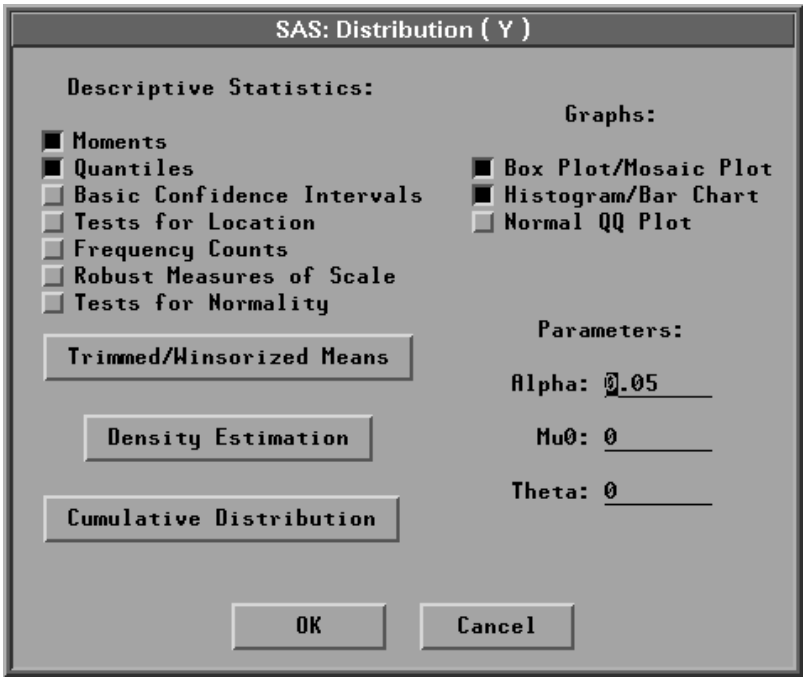


Figure 38.4. Distribution Output Options Dialog

The options you set in this dialog determine which tables and graphs appear in the distribution window. A distribution analysis can include descriptive statistics, graphs, density estimates, and cumulative distribution function estimates. By default, SAS/INSIGHT software displays a moments table, a quantiles tables, a box plot, and a histogram. Individual tables and graphs are described following this section.

You can specify the α coefficient in the **Parameters:Alpha:** entry field. The $100(1-\alpha)\%$ confidence level is used in the basic confidence intervals and the trimmed/Winsorized means tables. You can specify μ_0 in the **Parameters: Mu0:** entry field. μ_0 is used in the tests for location and the trimmed/Winsorized means tables. You can also specify θ in the **Parameters: Theta:** entry field. The parameter θ is used in the parametric density estimation and cumulative distribution for lognormal, exponential, and Weibull distributions.

If you select a **Weight** variable, tables of weighted moments, weighted quantiles, weighted confidence intervals, weighted tests for location, and weighted frequency counts can be generated. Robust measures of scale, tests for normality, and trimmed/Winsorized means are not computed. Graphs of weighted box plot, weighted histogram, and weighted normal QQ plot can also be generated.

The **Trimmed/Winsorized Means** button enables you to view or change the options associated with trimmed and Winsorized means. Click on **Trimmed/Winsorized Means** to display the **Trimmed/Winsorized Means** dialog.

Figure 38.5. Trimmed / Winsorized Means Dialog

In the dialog, you choose the number of observations trimmed or Winsorized in each tail in (1/2)N and the percent of observations trimmed or Winsorized in each tail in (1/2)Percent. If you specify a percentage, the smallest integer greater than or equal to np is trimmed or Winsorized.

The **Density Estimation** button enables you to set the options associated with both parametric density and nonparametric kernel density estimation. Click on **Density Estimation** to display the **Density Estimation** dialog.

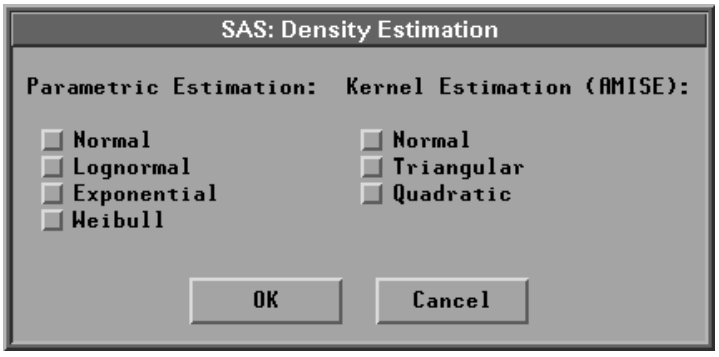


Figure 38.6. Density Estimation Dialog

If you select **Parametric Estimation:Normal**, a normal distribution with the sample mean and standard deviation is created. For the lognormal, exponential, and Weibull distributions, you specify the threshold parameter θ in the **Parameters:Theta:** entry field in the distribution output options dialog, as shown in Figure 38.4, and have the remaining parameters estimated by the maximum-likelihood estimates.

If you select a **Weight** variable, the weighted parametric normal density and weighted kernel density are generated. The parametric lognormal, exponential, and Weibull density are not computed.

The **Cumulative Distribution** button enables you to set the options associated with cumulative distribution estimation. Click on **Cumulative Distribution** to display the **Cumulative Distribution** dialog.

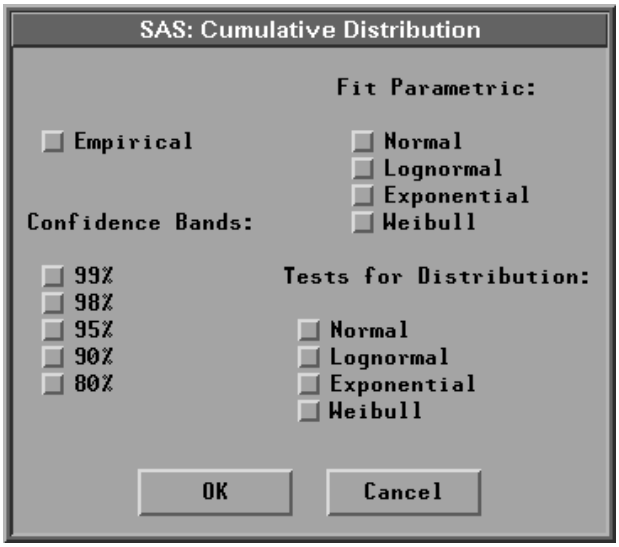


Figure 38.7. Cumulative Distribution Dialog

If you select **Fit Parametric:Normal**, a normal distribution with the sample mean and standard deviation is created. For the lognormal, exponential, and Weibull distributions, you specify the threshold parameter θ in the **Parameters:Theta:** entry field in the distribution output options dialog, as shown in Figure 38.4, and have the remaining parameters estimated by the maximum-likelihood estimates.

If you select a **Weight** variable, weighted empirical and normal cumulative distribution functions can be generated. The confidence bands, the parametric lognormal, exponential, and Weibull cumulative distributions, and tests for distribution are not computed.

Click on **OK** to close the dialogs and create your distribution analysis.

Tables

You can generate distribution tables by setting the options in the output options dialog or by choosing from the **Tables** menu.

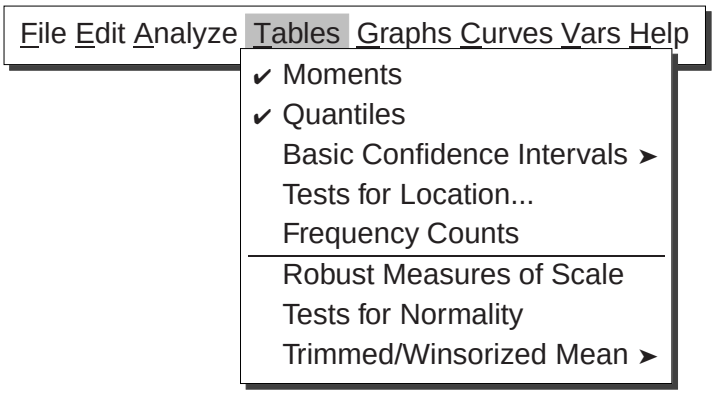


Figure 38.8. Tables Menu

The tables of robust measures of scale, tests for normality, and trimmed/Winsorized mean are not created for weighted analyses.

Moments

The **Moments** table, as shown in Figure 38.9, includes the following statistics:

- **N** is the number of nonmissing values, n.
- **Sum Wgts** is the sum of weights and is equal to n if no **Weight** variable is specified.
- **Mean** is the sample mean, $\overline{y}$.
- **Sum** is the variable sum, $\sum_i y_i$.
- **Std Dev** is the standard deviation, s.
- **Variance** is the variance, s^2.
- **Skewness** is the sample skewness, g_1.
- **Kurtosis** is the sample kurtosis, g_2.
- **USS** is the uncorrected sum of squares, $\sum_i y_i^2$.
- **CSS** is the sum of squares corrected for the mean, $\sum_i (y_i - \overline{y})^2$.
- **CV** is the percent coefficient of variation, $100s/\overline{y}$.
- **Std Mean** is the standard error of the mean, $s/\sqrt{n}$. The value is set to missing if vardef$\neq$**DF**.

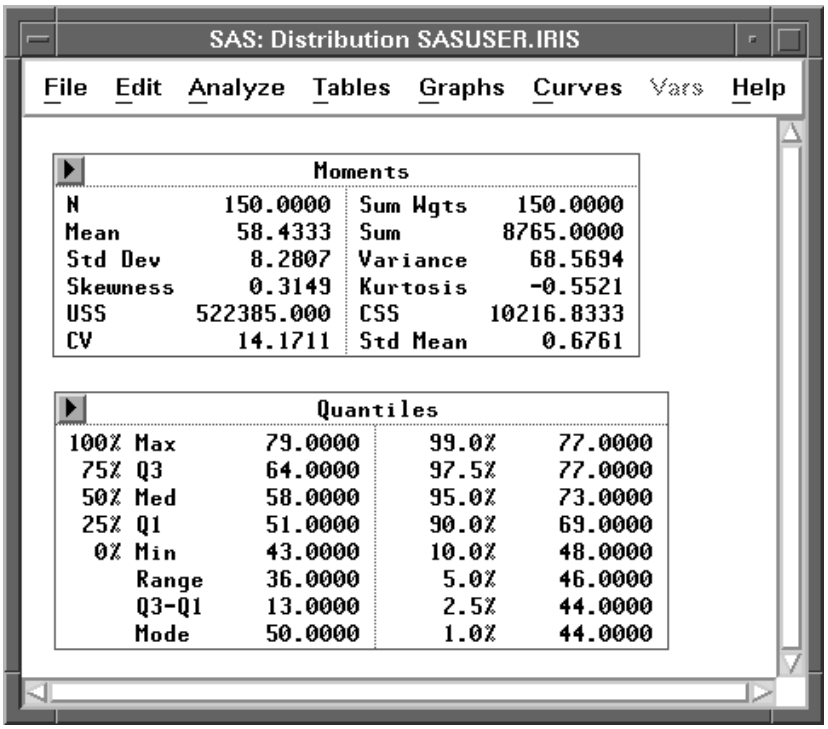


Figure 38.9. Moments and Quantiles Tables

For weighted analyses, the **Weighted Moments** table includes the following statistics:

- **N** is the number of nonmissing values, n.
- **Sum Wgts** is the sum of weights, $\sum_i w_i$.
- **Mean** is the weighted sample mean, $\overline{y}_w$.
- **Sum** is the weighted variable sum, $\sum_i w_i y_i$.
- **Std Dev** is the weighted standard deviation, s_w.
- **Variance** is the weighted variance, s_w^2.
- **Skewness** is the weighted sample skewness, g_{w1}.
- **Kurtosis** is the weighted sample kurtosis, g_{w2}.
- **USS** is the uncorrected weighted sum of squares, $\sum_i w_i y_i^2$.
- **CSS** is the weighted sum of squares corrected for the mean, $\sum_i w_i (y_i - \overline{y}_w)^2$.
- **CV** is the percent coefficient of variation, $100 s_w / \overline{y}_w$.
- **Std Mean** is the standard error of the weighted mean, $s_w / \sum_i w_i$.

  The value is set to missing if vardef$\neq$**DF**.

Quantiles

It is often convenient to subdivide the area under a density curve so that the area to the left of the dividing value is some specified fraction of the total unit area. For a given value of p between 0 and 1, the pth quantile (or $100p$th percentile) is the value such that the area to the left of it is p.

The pth quantile is computed from the empirical distribution function with averaging:

$$y = \begin{cases} \frac{1}{2}(y_{(i)} + y_{(i+1)}) & \text{if } f = 0 \\ y_{(i+1)} & \text{if } f > 0 \end{cases}$$

where i is the integer part and f is the fractional part of $np = i + f$.

If you specify a **Weight** variable, the pth quantile is computed as

$$y = \begin{cases} \frac{1}{2}(y_{(i)} + y_{(i+1)}) & \text{if } \sum_{j=1}^{i} w_{(j)} = p \sum_{j=1}^{n} w_{(j)} \\ y_{(i+1)} & \text{if } \sum_{j=1}^{i} w_{(j)} < p \sum_{j=1}^{n} w_{(j)} < \sum_{j=1}^{i+1} w_{(j)} \end{cases}$$

When each observation has an identical weight, the weighted quantiles are identical to the unweighted quantiles.

The **Quantiles** table, as shown in Figure 38.9, includes the following statistics:

- **100% Max** is the maximum, $y_{(n)}$.
- **75% Q3** is the upper quartile (the 75th percentile).
- **50% Med** is the median.
- **25% Q1** is the lower quartile (the 25th percentile).
- **0% Min** is the minimum, $y_{(1)}$.
- **99%**, **97.5%**, **95%**, **90%**, **10%**, **5%**, **2.5%**, and **1%** give the corresponding percentiles.
- **Range** is the range, $y_{(n)} - y_{(1)}$.
- **Q3-Q1**, the interquartile range, is the difference between the upper and lower quartiles.
- **Mode** is the most frequently occurring value. When there is more than one mode, the lowest mode is displayed. When all the distinct values have frequency one, the value is set to missing.

Basic Confidence Intervals

Assuming that the population is normally distributed, the **Confidence Intervals** table gives confidence intervals for the mean, standard deviation, and variance at the confidence coefficient specified. You specify the confidence intervals either in the distribution output options dialog or from the **Tables** menu.

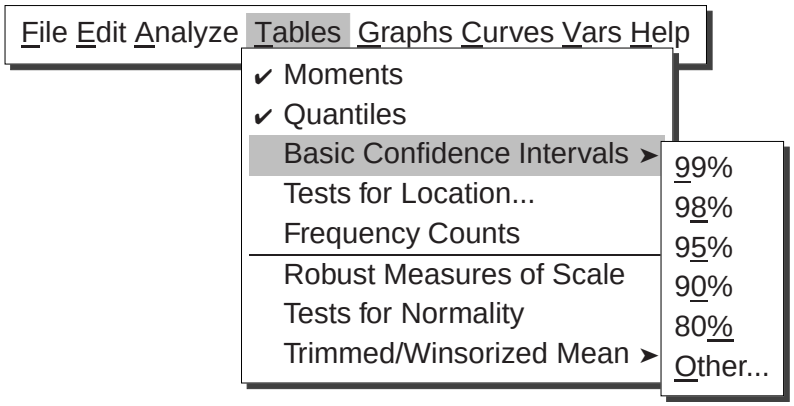


Figure 38.10. Basic Confidence Intervals Menu

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

$$\overline{y} \pm t_{(1-\alpha/2)} \frac{s}{\sqrt{n}}$$

where $t_{(1-\alpha/2)}$ is the $(1-\alpha/2)$ critical value of the Student's t statistic with $n-1$ degrees of freedom.

For weighted analyses, the limits are

$$\overline{y}_w \pm t_{(1-\alpha/2)} \frac{s_w}{\sqrt{\sum_i w_i}}$$

For large values of n, $t_{(1-\alpha/2)}$ acts as $z_{(1-\alpha/2)}$, the $(1-\alpha/2)$ critical value of the standard normal distribution.

The $100(1-\alpha)\%$ confidence interval for the standard deviation has upper and lower limits

$$s\sqrt{\frac{n-1}{c_{\alpha/2}}} \text{ and } \mathrm{s}\sqrt{\frac{\mathrm{n}-1}{\mathrm{c}_{(1-\alpha/2)}}}$$

where $c_{\alpha/2}$ and $c_{(1-\alpha/2)}$ are the $\alpha/2$ and $(1-\alpha/2)$ critical values of the chi-square distribution with $n-1$ degrees of freedom.

For weighted analyses, the limits are

$$s_w\sqrt{\frac{n-1}{c_{\alpha/2}}} \text{ and } \mathrm{s_w}\sqrt{\frac{\mathrm{n}-1}{\mathrm{c}_{(1-\alpha/2)}}}$$

The $100(1-\alpha)\%$ confidence interval for the variance has upper and lower limits equal to the squares of the corresponding upper and lower limits for the standard deviation.

Figure 38.11 shows a table of the 95% confidence intervals for the mean, standard deviation, and variance.

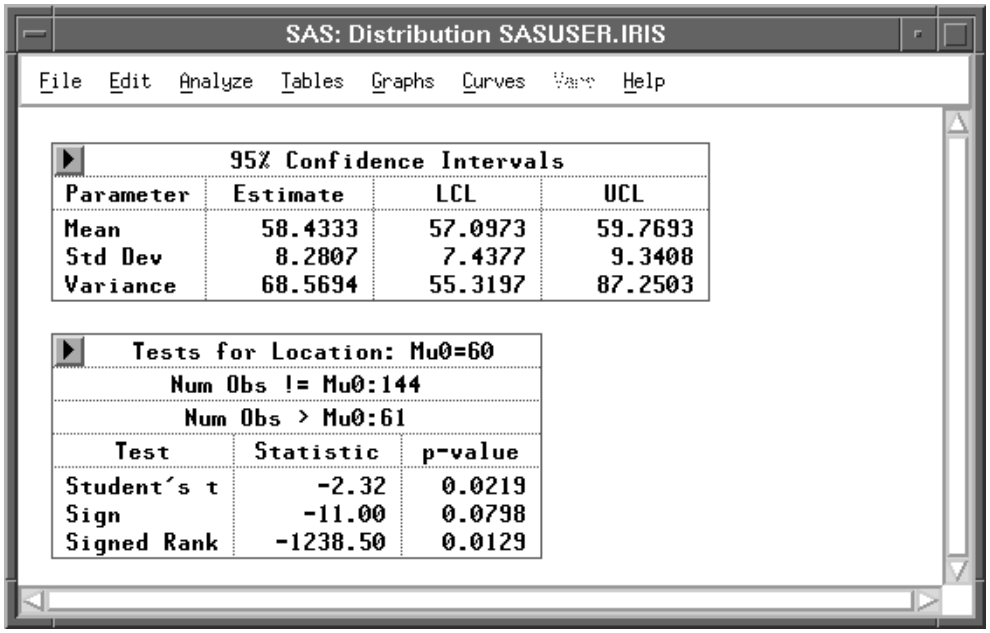


Figure 38.11. Basic Confidence Intervals and Tests for Location Tables

† **Note:** The confidence intervals are set to missing if vardef≠**DF**.

Tests for Location

The location tests include the Student's t, sign, and signed rank tests of the hypothesis that the mean/median is equal to a given value μ against the two-sided alternative that the mean/median is not equal to μ. The Student's t test is appropriate when the data are from an approximately normal population; otherwise, nonparametric tests such as the sign or signed rank test should be used.

The **Student's t** gives a Student's t statistic

$$t = \frac{\overline{y} - \mu_0}{s \,/\, \sqrt{n}}$$

For weighted analyses, the t statistic is computed as

$$t = \frac{\overline{y}_w - \mu_0}{s_w \,/\, \sqrt{\sum_i w_i}}$$

Assuming that the null hypothesis (H_0: mean = μ) is true and the population is normally distributed, the t statistic has a Student's t distribution with $n - 1$ degrees of freedom. The p-value is the probability of obtaining a Student's t statistic greater in absolute value than the absolute value of the observed statistic t.

† **Note:** The t statistic and p-value are set to missing if vardef≠**DF**.

The **Sign** statistic is

$$\mathbf{M} = \frac{1}{2}(n^+ - n^-)$$

where n^+ is the number of observations with values greater than μ, and n^- is the number of observations with values less than μ.

Assuming that the null hypothesis (H_0: median = μ_0) is true, the p-value for the observed statistic M is

$$\mathrm{Prob}\{|\mathbf{M}| >= |\mathrm{M}|\} = (\frac{1}{2})^{\mathrm{n_t}-1} \sum_{\mathrm{i}=0}^{\min(\mathrm{n}^+,\mathrm{n}^-)} \binom{\mathrm{n_t}}{\mathrm{i}}$$

where $n_t = n^+ + n^-$ is the number of y_i values not equal to μ_0.

The **Signed Rank** test assumes that the distribution is symmetric. The signed rank statistic is computed as $S = \Sigma r_i^+ - n_t(n_t + 1)/4$ where r_i^+ is the rank of $|y_i - \mu_0|$ after discarding y_i values equal to μ_0, and the sum is calculated for values of $y_i > \mu_0$. Average ranks are used for tied values.

The p-value is the probability of obtaining a signed rank statistic greater in absolute value than the absolute value of the observed statistic S. If $n_t <= 20$, the p-value of the statistic S is computed from the exact distribution of S. When $n_t > 20$, the significance level of S is computed by treating

$$\sqrt{n_t - 1}\frac{S}{\sqrt{n_t V - S^2}}$$

as a Student's t variate with $n_t - 1$ degrees of freedom, where V is computed as

$$V = \frac{1}{24}\{n_t(n_t + 1)(2n_t + 1) - \frac{1}{2}\sum_{j=1}^{n} t_j(t_j + 1)(t_j - 1)\}.$$

The sum is calculated over groups tied in absolute value, and t_j is the number of tied values in the jth group (Iman 1974, Lehmann 1975).

You can specify location tests either in the distribution output options dialog or in the **Location Tests** dialog after choosing **Tables:Tests for Location** from the menu.

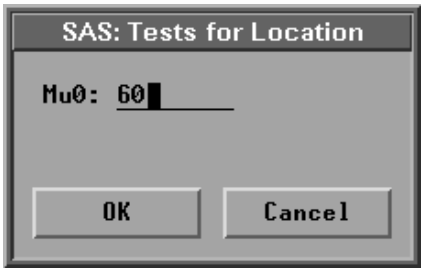


Figure 38.12. Location Tests Dialog

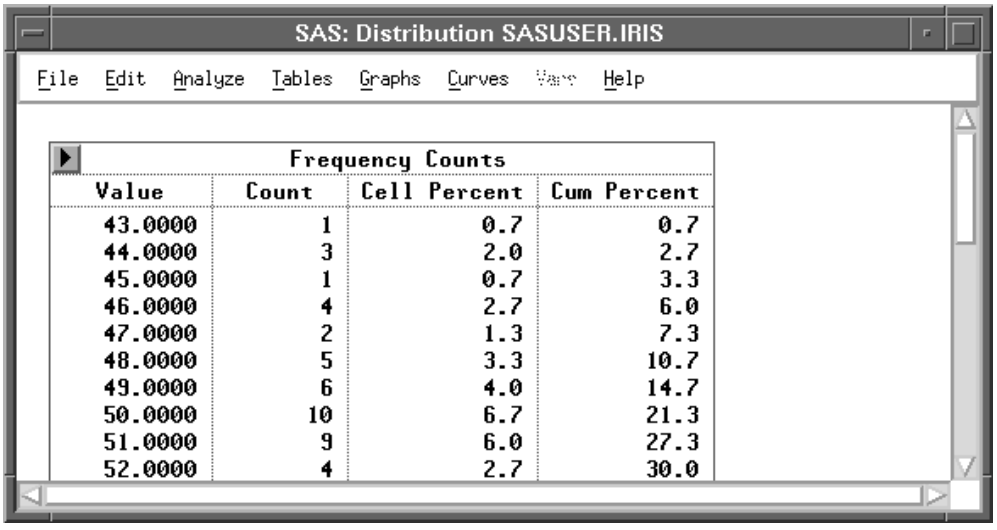
In the dialog, you can specify the parameter μ_0. Figure 38.11 shows a table of the three location tests for $\mu_0 = 60$. Here, **Num Obs != Mu0** is the number of observations with values not equal to μ_0, and **Num Obs > Mu0** is the number of observations with values greater than μ_0.

For weighted analyses, the sign and signed rank tests are not generated.

Frequency Counts

The **Frequency Counts** table, a portion of which is shown in Figure 38.13, includes the variable values, counts, percentages, and cumulative percentages. You can generate frequency tables for both interval and nominal variables.

If you specify a **Weight** variable, the table also includes the weighted counts. These weighted counts are used to compute the percentages and cumulative percentages.



Frequency Counts

Value	Count	Cell Percent	Cum Percent
43.0000	1	0.7	0.7
44.0000	3	2.0	2.7
45.0000	1	0.7	3.3
46.0000	4	2.7	6.0
47.0000	2	1.3	7.3
48.0000	5	3.3	10.7
49.0000	6	4.0	14.7
50.0000	10	6.7	21.3
51.0000	9	6.0	27.3
52.0000	4	2.7	30.0

Figure 38.13. Frequency Counts Table

Robust Measures of Scale

The sample standard deviation is a commonly used estimator of the population scale. However, it is sensitive to outliers and may not remain bounded when a single data point is replaced by an arbitrary number. With robust scale estimators, the estimates remain bounded even when a portion of the data points are replaced by arbitrary numbers.

A simple robust scale estimator is the interquartile range, which is the difference between the upper and lower quartiles. For a normal population, the standard deviation σ can be estimated by dividing the interquartile range by 1.34898.

Gini's mean difference is also a robust estimator of the standard deviation σ. It is computed as

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

If the observations are from a normal distribution, then $\sqrt{\pi}G/2$ is an unbiased estimator of the standard deviation σ.

A very robust scale estimator is the median absolute deviation (*MAD*) about the median (Hampel 1974).

$$MAD = med_i(|y_i - med_j(y_j)|)$$

where the inner median, $med_j(y_j)$, is the median of the n observations and the outer median, med_i, is the median of the n absolute values of the deviations about the median.

For a normal distribution, 1.4826 *MAD* can be used to estimate the standard deviation σ.

The *MAD* statistic has low efficiency for normal distributions and it may not be appropriate for symmetric distributions. Rousseeuw and Croux (1993) proposed two new statistics as alternatives to the *MAD* statistic, S_n and Q_n.

$$S_n = 1.1926\, med_i(\, med_j(|y_i - y_j|))$$

where the outer median, med_i, is the median of the n medians of

$$\{|y_i - y_j|; j = 1, 2, .., n\}.$$

To reduce small-sample bias, $c_{sn}S_n$ is used to estimate the standard deviation σ, where c_{sn} is a correction factor (Croux and Rousseeuw 1992).

The second statistic is computed as

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

where $k = \binom{h}{2}$, $h = [n/2] + 1$ and $[n/2]$ is the integer part of $n/2$. That is, Q_n is 2.2219 times the kth order statistic of the $\binom{n}{2}$ distances between data points.

The bias-corrected statistic $c_{qn}Q_n$ is used to estimate the standard deviation σ, where $c_{qn}$is the correction factor.

A **Robust Measures of Scale** table includes the interquartile range, Gini's mean difference, *MAD*, S_n, and Q_n, with their corresponding estimates of σ, as shown in Figure 38.14.

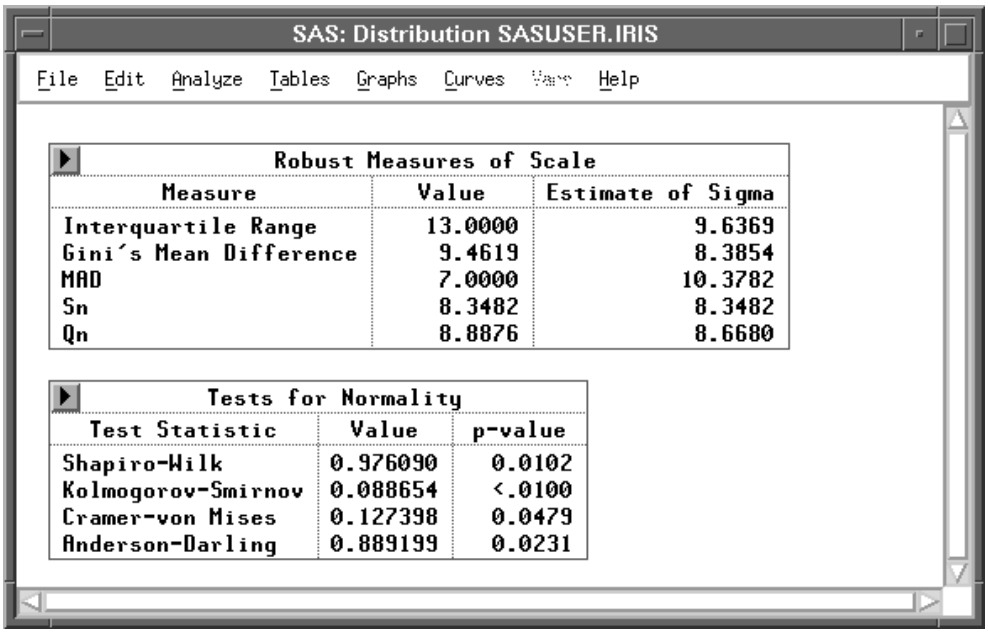


Figure 38.14. Robust Measures of Scale and Tests for Normality

Tests for Normality

SAS/INSIGHT software provides tests for the null hypothesis that the input data values are a random sample from a normal distribution. These test statistics include the Shapiro-Wilk statistic (W) and statistics based on the empirical distribution function: the Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling statistics.

The Shapiro-Wilk statistic is the ratio of the best estimator of the variance (based on the square of a linear combination of the order statistics) to the usual corrected sum of squares estimator of the variance. W must be greater than zero and less than or equal to one, with small values of W leading to rejection of the null hypothesis of normality. Note that the distribution of W is highly skewed. Seemingly large values of W (such as 0.90) may be considered small and lead to the rejection of the null hypothesis.

The W statistic is computed when the sample size is less than or equal to 2000. When the sample size is greater than three, the coefficients for computing the linear combination of the order statistics are approximated by the method of Royston (1992).

With a sample size of three, the probability distribution of W is known and is used to determine the significance level. When the sample size is greater than three, simulation results are used to obtain the approximate normalizing transformation (Royston 1992)

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

where γ, μ, and σ are functions of n, obtained from simulation results, and Z_n is a standard normal variate with large values indicating departure from normality.

The Kolmogorov statistic assesses the discrepancy between the empirical distribution and the estimated hypothesized distribution. For a test of normality, the hypothesized distribution is a normal distribution function with parameters μ and σ estimated by the sample mean and standard deviation. The probability of a larger test statistic is obtained by linear interpolation within the range of simulated critical values given by Stephens (1974).

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

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

and it is computed as

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

where $U_{(i)} = F(y_{(i)})$ is the cumulative distribution function value at $y_{(i)}$, the ith ordered value. The probability of a larger test statistic is obtained by linear interpolation within the range of simulated critical values given by Stephens (1974).

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

$$A^2 = n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \{F(x)(1 - F(x))\}^{-1} dF(x)$$

and it is computed as

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

The probability of a larger test statistic is obtained by linear interpolation within the range of simulated critical values in D'Agostino and Stephens (1986).

A **Tests for Normality** table includes the Shapiro-Wilk, Kolmogorov, Cramer-von Mises, and Anderson-Darling test statistics, with their corresponding p-values, as shown in Figure 38.14.

Trimmed and Winsorized Means

When outliers are present in the data, trimmed and Winsorized means are robust estimators of the population mean that are relatively insensitive to the outlying values. Therefore, trimming and Winsorization are methods for reducing the effects of extreme values in the sample.

The k-times trimmed mean is calculated as

$$\overline{y}_{tk} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} y_{(i)}$$

The trimmed mean is computed after the k smallest and k largest observations are deleted from the sample. In other words, the observations are trimmed at each end.

The k-times Winsorized mean is calculated as

$$\overline{y}_{wk} = \frac{1}{n}\{(k+1)y_{(k+1)} + \sum_{i=k+2}^{n-k-1} y_{(i)} + (k+1)y_{(n-k)}\}$$

The Winsorized mean is computed after the k smallest observations are replaced by the $(k+1)$st smallest observation, and the k largest observations are replaced by the $(k+1)$st largest observation. In other words, the observations are Winsorized at each end.

For a symmetric distribution, the symmetrically trimmed or Winsorized mean is an unbiased estimate of the population mean. But the trimmed or Winsorized mean does not have a normal distribution even if the data are from a normal population.

The Winsorized sum of squared deviations is defined as

$$s^2_{wk} = (k+1)(y_{(k+1)} - \overline{y}_{wk})^2 + \sum_{i=k+2}^{n-k-1} (y_{(i)} - \overline{y}_{wk})^2 + (k+1)(y_{(n-k)} - \overline{y}_{wk})^2$$

A robust estimate of the variance of the trimmed mean $\overline{y}_{tk}$ can be based on the Winsorized sum of squared deviations (Tukey and McLaughlin 1963). The resulting trimmed t test is given by

$$t_{tk} = \frac{\overline{y}_{tk}}{\mathrm{STDERR}(\overline{\mathrm{y}}_{\mathrm{tk}})}$$

where $\mathrm{STDERR}(\overline{\mathrm{y}}_{\mathrm{tk}})$ is the standard error of $\overline{y}_{tk}$:

$$\mathrm{STDERR}(\overline{\mathrm{y}}_{\mathrm{tk}}) = \frac{\mathrm{s_{wk}}}{\sqrt{(\mathrm{n}-2\mathrm{k})(\mathrm{n}-2\mathrm{k}-1)}}$$

A Winsorized t test is given by

$$t_{wk} = \frac{\overline{y}_{wk}}{\mathrm{STDERR}(\overline{\mathrm{y}}_{\mathrm{wk}})}$$

where $\mathrm{STDERR}(\overline{\mathrm{y}}_{\mathrm{wk}})$ is the standard error of $\overline{y}_{wk}$:

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

When the data are from a symmetric distribution, the distribution of the trimmed t statistic t_{tk} or the Winsorized t statistic t_{wk} can be approximated by a Student's t distribution with $n - 2k - 1$ degrees of freedom (Tukey and McLaughlin 1963, Dixon and Tukey 1968).

You can specify the number or percentage of observations to be trimmed or Winsorized from each end either by using the **Trimmed/Winsorized Means** options dialog or by using the **Trimmed/Winsorized Means** dialog after choosing **Tables:Trimmed/Winsorized Mean:(1/2)N** or **Tables:Trimmed/Winsorized Mean:(1/2)Percent** from the menus.

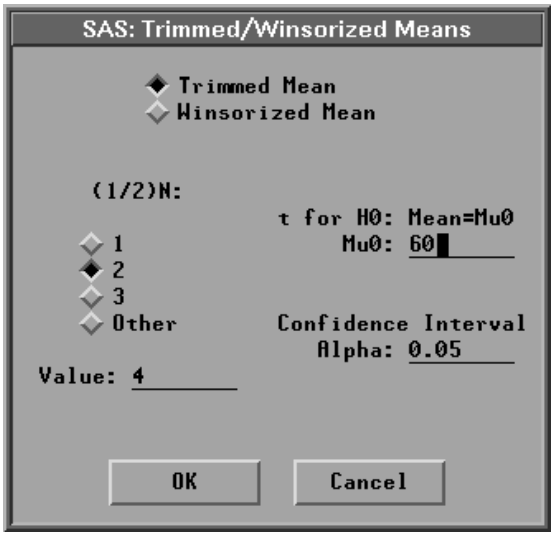


Figure 38.15. (1/2)N Menu

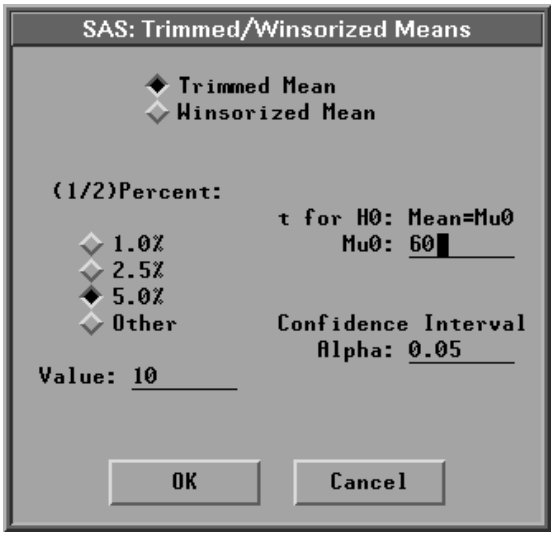


Figure 38.16. (1/2)Percent Menu

If you specify a percentage, $100p\%$, $0 < p < 1$, the smallest integer greater than or equal to np is trimmed or Winsorized from each end.

The **Trimmed Mean** and **Winsorized Mean** tables, as shown in Figure 38.17, contain the following statistics:

- **(1/2)Percent** is the percentage of observations trimmed or Winsorized at each end.
- **(1/2)N** is the number of observations trimmed or Winsorized at each end.
- **Mean** is the trimmed or Winsorized mean.
- **Std Mean** is the standard error of the trimmed or Winsorized mean.
- **DF** is the degrees of freedom used in the Student's t test for the trimmed or Winsorized mean.
- **Confidence Interval** includes **Level (%)**: the confidence level, **LCL**: lower confidence limit, and **UCL**: upper confidence limit.
- **t for H0: Mean=Mu0** includes **Mu0**: the location parameter μ_0, **t Stat**: the trimmed or Winsorized t statistic for testing the hypothesis that the population mean is μ_0, and **p-value**: the approximate p-value of the trimmed or Winsorized t statistic.

SAS: Distribution SASUSER.IRIS

File Edit Analyze Tables Graphs Curves Help

Trimmed Means

					Confidence Interval			t for H0: Mean=Mu0		
(1/2)Percen	(1/2)N	Mean	Std Mean	DF	Level (%)	LCL	UCL	Mu0	t Stat	p-value
1.33	2	58.3699	0.6910	145	95.00	57.0041	59.7356	60.0000	-2.36	0.0197
5.33	8	58.1866	0.7047	133	95.00	56.7927	59.5804	60.0000	-2.57	0.0112

Winsorized Means

					Confidence Interval			t for H0: Mean=Mu0		
(1/2)Percen	(1/2)N	Mean	Std Mean	DF	Level (%)	LCL	UCL	Mu0	t Stat	p-value
1.33	2	58.4267	0.6911	145	95.00	57.0608	59.7926	60.0000	-2.28	0.0243
5.33	8	58.2733	0.7050	133	95.00	56.8790	59.6677	60.0000	-2.45	0.0156

Figure 38.17. Trimmed Means and Winsorized Means Tables

Graphs

You can generate a histogram, a box plot, or a quantile-quantile plot in the distribution output options dialog or from the **Graphs** menu.

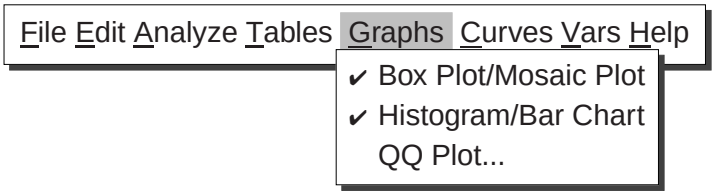


Figure 38.18. Graphs Menu

If you select a **Weight** variable, a weighted box plot/mosaic plot, a weighted histogram/bar chart, and a weighted normal QQ plot can be generated.

Box Plot/Mosaic Plot

The *box plot* is a stylized representation of the distribution of a variable, and it is shown in Figure 38.19. You can also display mosaic plots for nominal variables, as shown in Figure 38.37.

In a box plot, the sample mean and sample standard deviation computed with vardef=**DF** are used in the construction of the mean diamond, as shown in Figure 38.19.

If you select a **Weight** variable, a weighted box plot based on weighted quantiles is created. The weighted sample mean and the weighted sample standard deviation of an observation with average weight for vardef=**DF** is used in the construction of the mean diamond.

⊕ **Related Reading:** Box Plots, Chapter 33.

Histogram/Bar Chart

The histogram is the most widely used density estimator, and it is shown in Figure 38.19. You can also display bar charts for nominal variables, as shown in Figure 38.37.

⊕ **Related Reading:** Bar Charts, Chapter 32.

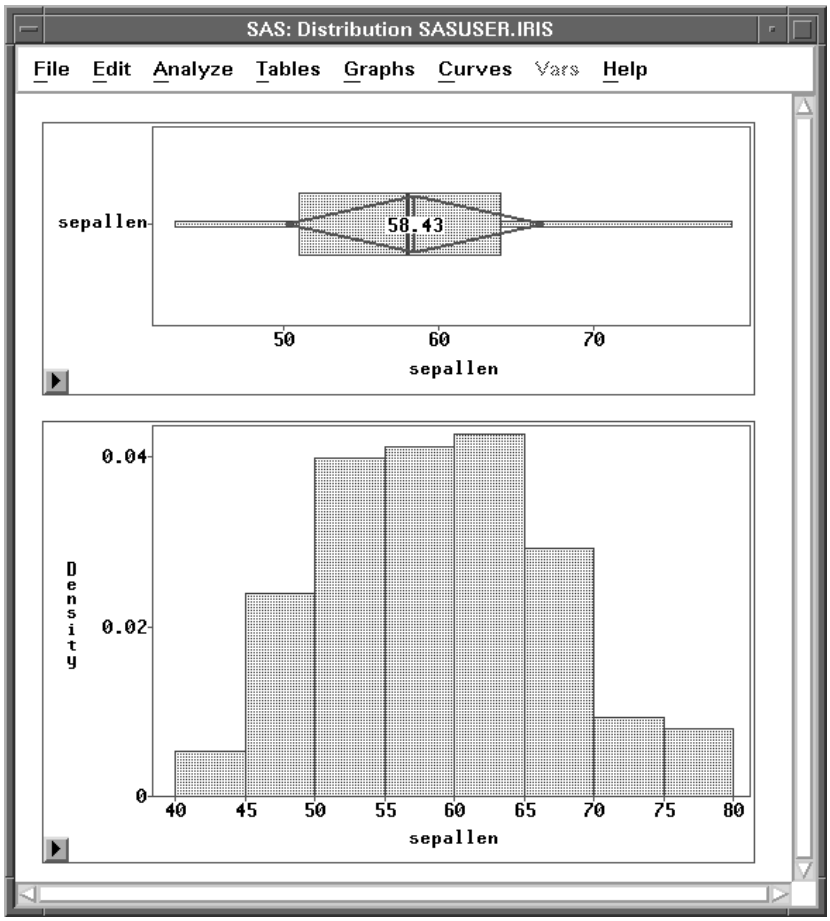


Figure 38.19. Box Plot and Histogram

QQ Plot

A *quantile-quantile plot* (QQ plot) compares ordered values of a variable with quantiles of a specific theoretical distribution. If the data are from the theoretical distribution, the points on the QQ plot lie approximately on a straight line. The normal, lognormal, exponential, and Weibull distributions can be used in the plot.

You can specify the type of QQ plot from the **QQ Plot** dialog after choosing **Graphs:QQ Plot** from the menu.

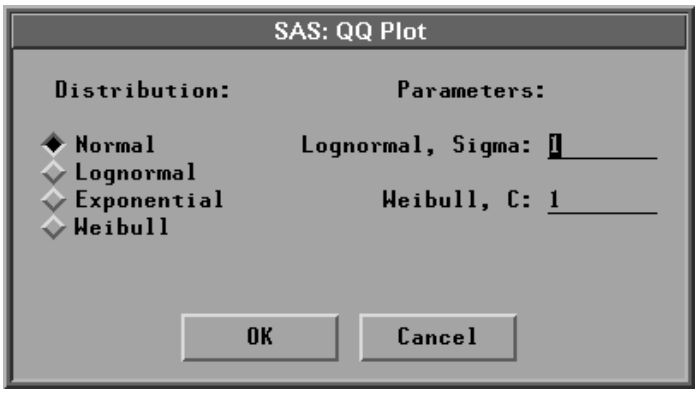


Figure 38.20. QQ Plot Dialog

In the dialog, you must specify a shape parameter for the lognormal or Weibull distribution. The normal QQ plot can also be generated with the graphs options dialog. As described later in this chapter, you can also add a reference line to the QQ plot from the **Curves** menu.

The following expression is used in the discussion that follows:

$$v_i = \frac{i - 0.375}{n + 0.25} \quad \text{for } i = 1, 2, \ldots, n$$

where n is the number of nonmissing observations.

For the normal distribution, the ith ordered observation is plotted against the normal quantile $\Phi^{-1}(v_i)$, where Φ^{-1} is the inverse standard cumulative normal distribution. If the data are normally distributed with mean μ and standard deviation σ, the points on the plot should lie approximately on a straight line with intercept μ and slope σ. The normal quantiles are stored in variables named **N_name** for each variable, where **name** is the **Y** variable name.

For the lognormal distribution, the ith ordered observation is plotted against the lognormal quantile $\exp\left(\sigma\Phi^{-1}(v_i)\right)$ for a given shape parameter σ. If the data are lognormally distributed with parameters θ, σ, and ζ, the points on the plot should lie approximately on a straight line with intercept θ and slope $\exp(\zeta)$. The lognormal quantiles are stored in variables named **L_name** for each variable, where **name** is the **Y** variable name.

For the exponential distribution, the ith ordered observation is plotted against the exponential quantile $-\log(1 - \mathrm{v_i})$. If the data are exponentially distributed with parameters θ and σ, the points on the plot should lie approximately on a straight line with intercept θ and slope σ. The exponential quantiles are stored in variables named **E_name** for each variable, where **name** is the **Y** variable name.

For the Weibull distribution, the ith ordered observation is plotted against the Weibull quantile $(-\log(1 - \mathrm{v_i}))^{\frac{1}{c}}$ for a given shape parameter c. If the data are from a Weibull distribution with parameters θ, σ, and c, the points on the plot should lie approximately on a straight line with intercept θ and slope σ. The Weibull quantiles are stored in variables named **W_name** for each variable, where **name** is the **Y** variable name.

A normal QQ plot is shown in Figure 38.21. You can also add a reference line to the QQ plot from the **Curves** menu. You specify the intercept and slope for the reference line from the **Curves** menu.

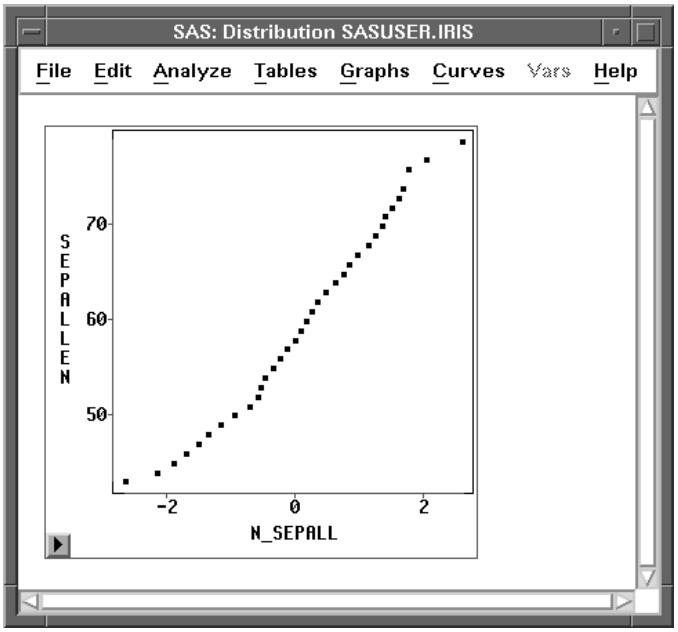


Figure 38.21. Normal QQ Plot

Further information on interpreting quantile-quantile plots can be found in Chambers et al. (1983).

If you select a **Weight** variable, a weighted normal QQ plot can be generated. Log-normal, exponential, and Weibull QQ plots are not computed.

For a weighted normal QQ plot, the ith ordered observation is plotted against the normal quantile $\Phi^{-1}(v_i)$, where

$$v_i = \frac{(\sum_{j=1}^{i} w_{(j)})(1 - 0.375/i)}{W(1 + 0.25/n)}$$

When each observation has an identical weight, $w_{(j)} = w_0$, the formulation reduces to the usual expression in the unweighted normal probability plot

$$v_i = \frac{i - 0.375}{n + 0.25}$$

If the data are normally distributed with mean μ and standard deviation σ and if each observation has approximately the same weight (w_0), then, as in the unweighted normal QQ plot, the points on the plot should lie approximately on a straight line with intercept μ and slope σ for vardef=WDF/WGT and with slope $\sigma/\sqrt{w_0}$ for vardef=DF/N.

Curves

Density estimation is the construction of an estimate of the density function from the observed data. The methods provided for univariate density estimation include parametric estimators and kernel estimators.

Cumulative distribution analyses include the empirical and the parametric cumulative distribution function. The empirical distribution function is a nonparametric estimator of the cumulative distribution function. You can fit parametric distribution functions if the data are from a known family of distributions, such as the normal, lognormal, exponential, or Weibull.

You can use the Kolmogorov statistic to construct a confidence band for the unknown distribution function. The statistic also tests the hypotheses that the data are from a completely specified distribution or from a specified family of distributions with unknown parameters.

You can generate density estimates and cumulative distribution analysis in the output options dialog, as described previously in the section "Output," or by choosing from the **Curves** menu, as shown in Figure 38.22. You can also generate QQ reference lines from the **Curves** menu.

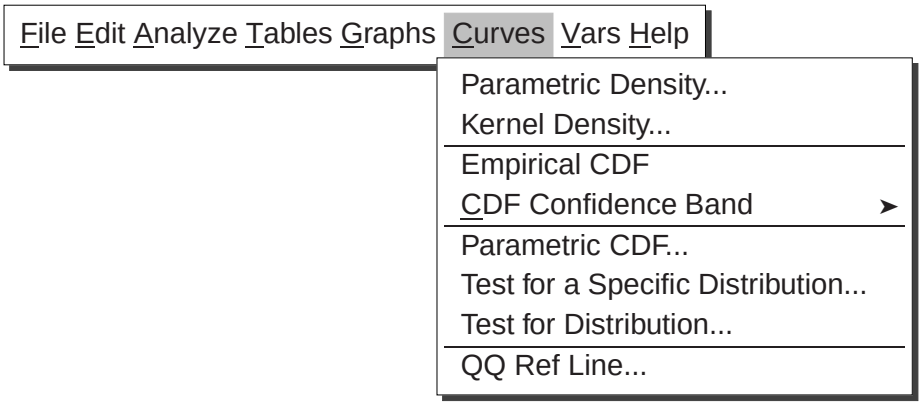


Figure 38.22. Curves Menu

If you select a **Weight** variable, curves of parametric weighted normal density, weighted kernel density, weighted empirical CDF, parametric weighted normal CDF, and weighted QQ reference line (based on weighted least squares) can be generated. CDF confidence band, test for a specific distribution, and test for distribution are not computed.

Parametric Density

Parametric density estimation assumes that the data are from a known family of distributions, such as the normal, lognormal, exponential, and Weibull. After choosing **Curves:Parametric Density** from the menu, you specify the family of distributions in the **Parametric Density Estimation** dialog, as shown in Figure 38.23.

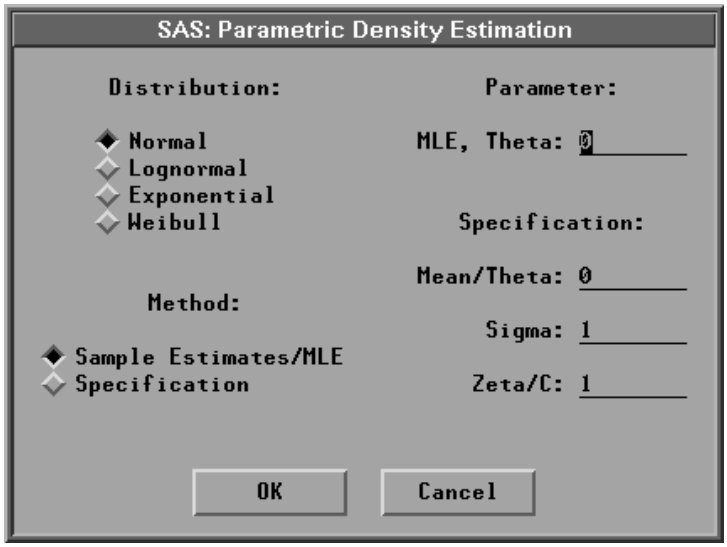


Figure 38.23. Parametric Density Dialog

The default uses a normal distribution with the sample mean and standard deviation as estimates for μ and σ. You can also specify your own μ and σ parameters for the normal distribution by choosing **Method:Specification** in the dialog.

For the lognormal, exponential, and Weibull distributions, you can specify your own threshold parameter θ in the **Parameter:MLE, Theta** entry field and have the remaining parameters estimated by the maximum-likelihood estimates (MLE) by choosing **Method:Sample Estimates/MLE**. Otherwise, you can specify all the parameters in the **Specification** fields and choose **Method:Specification** in the dialog.

If you select a **Weight** variable, only normal density can be created. For **Method:Sample Estimates/MLE**, $\overline{y}_w$ and s_w are used to display the density with vardef=**WDF/WGT**; $\overline{y}_w$ and s_a are used with vardef=**DF/N**. For **Method:Specification**, the values in the entry fields **Mean/Theta** and **Sigma** are used to display the density with vardef=**WDF/WGT**; the values of **Mean/Theta** and **Sigma**/$\sqrt{\overline{w}}$ are used with vardef=**DF/N**.

Figure 38.24 displays a normal density estimate with $\mu = 58.4333$ (the sample mean) and $\sigma = 8.2807$ (the sample standard deviation). It also displays a lognormal density estimate with $\theta = 30$ and with σ and ζ estimated by the MLE.

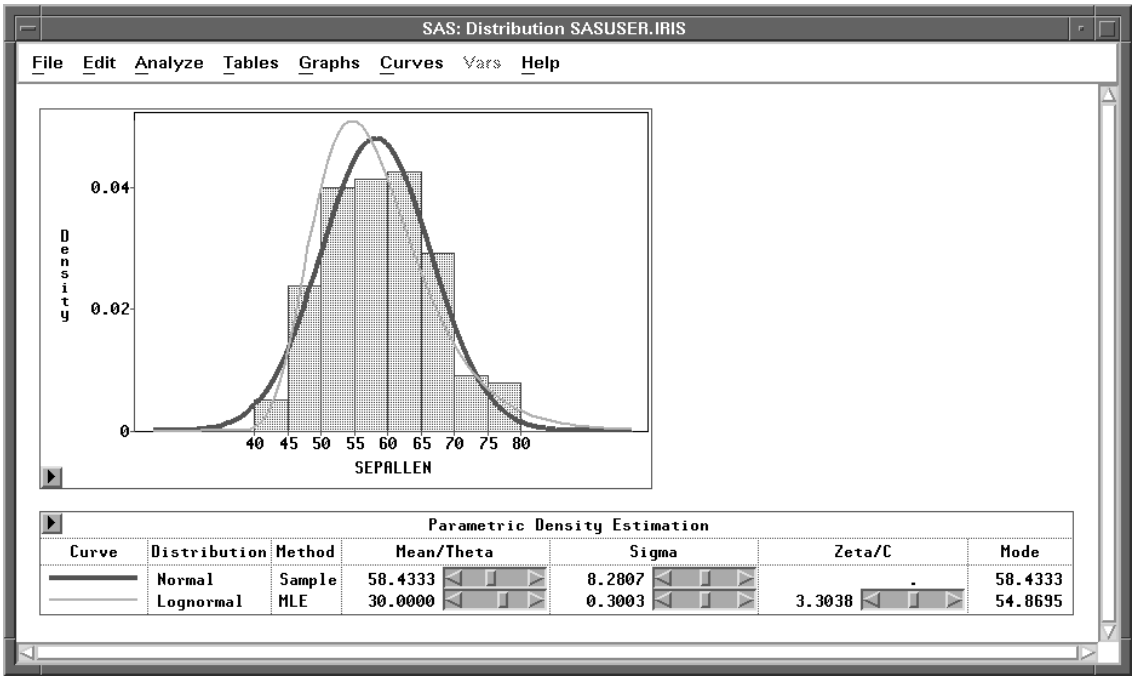


Figure 38.24. Parametric Density Estimation

The **Mode** is the point with the largest estimated density. Use sliders in the table to change the density estimate. When MLE is used for the lognormal, exponential, and Weibull distributions, changing the value of θ in the **Mean/Theta** slider also causes the remaining parameters to be estimated by the MLE for the new θ.

Kernel Density

Kernel density estimation provides normal, triangular, and quadratic kernel density estimators. The general form of a kernel estimator is

$$\hat{f}_\lambda(y) = \frac{1}{n\lambda} \sum_{i=1}^{n} K_0\left(\frac{y - y_i}{\lambda}\right)$$

where K_0 is a kernel function and λ is the bandwidth.

Some symmetric probability density functions commonly used as kernel functions are

- $Normal \quad K_0(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-t^2/2\right) \quad \text{for } -\infty < t < \infty$
- $Triangular \quad K_0(t) = \begin{cases} 1 - |t| & \text{for } |t| \leq 1 \\ 0 & \text{otherwise} \end{cases}$
- $Quadratic \quad K_0(t) = \begin{cases} \frac{3}{4}(1 - t^2) & \text{for } |t| \leq 1 \\ 0 & \text{otherwise} \end{cases}$

Both theory and practice suggest that the choice of a kernel function is not crucial to the statistical performance of the method (Epanechnikov 1969). With a specific kernel function, the value of λ determines the degree of averaging in the estimate of the density function and is called a *smoothing parameter*. You select a bandwidth λ for each kernel estimator by specifying c in the formula

$$\lambda = n^{-\frac{1}{5}} Q c$$

where Q is the sample interquartile range of the **Y** variable. This formulation makes c independent of the units of **Y**.

For a specific kernel function, the discrepancy between the density estimator $\hat{f}_\lambda(y)$ and the true density $f(y)$ can be measured by the mean integrated square error

$$\mathrm{MISE}(\lambda) = \int_{\mathrm{y}} \{\mathrm{E}(\hat{\mathrm{f}}_\lambda(\mathrm{y})) - \mathrm{f}(\mathrm{y})\}^2 \mathrm{dy} + \int_{\mathrm{y}} \mathrm{Var}(\hat{\mathrm{f}}_\lambda(\mathrm{y}))\, \mathrm{dy}$$

which is the sum of the integrated square bias and the integrated variance.

An approximate mean integrated square error based on the bandwidth λ is

$$\mathrm{AMISE}(\lambda) = \frac{1}{4}\lambda^4 \left(\int_{\mathrm{t}} \mathrm{t}^2 \mathrm{K}(\mathrm{t}) \mathrm{dt}\right)^2 \int_{\mathrm{y}} (\mathrm{f}''(\mathrm{y}))^2\, \mathrm{dy} + \frac{1}{\mathrm{n}\lambda} \int_{\mathrm{t}} \mathrm{K}(\mathrm{t})^2 \mathrm{dt}$$

If $f(y)$ is assumed normal, then a bandwidth based on the sample mean and variance can be computed to minimize AMISE. The resulting bandwidth for a specific kernel is used when the associated kernel function is selected in the density estimation options dialog. This is equivalent to choosing **MISE** from the normal, triangular, or quadratic kernel menus. If $f(y)$ is not roughly normal, this choice may not be appropriate.

SAS/INSIGHT software divides the range of the data into 128 evenly spaced intervals, then approximates the data on this grid and uses the fast Fourier transformation (Silverman 1986) to estimate the density.

If you select a **Weight** variable, the kernel estimator is modified to include the individual observation weights.

$$\hat{f}_\lambda(y) = \frac{1}{\sum_i w_i \lambda} \sum_{i=1}^{n} w_i K_0\left(\frac{y - y_i}{\lambda}\right)$$

You can specify the kernel function in the density estimation options dialog or from the **Curves** menu. When you specify the kernel function in the density estimation options dialog, **AMISE** is used. After choosing **Curves:Kernel Density** from the menu, you can specify the kernel function and use either **AMISE** or a specified C value in the **Kernel Density Estimation** dialog.

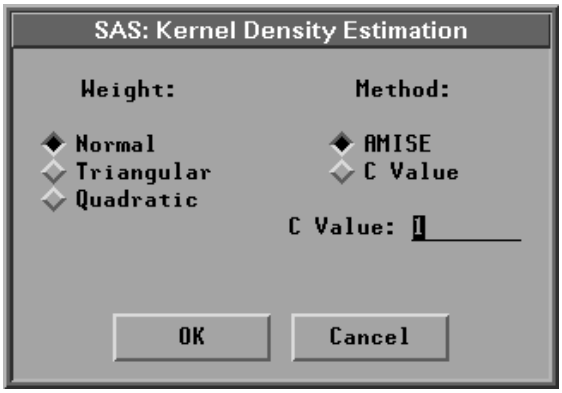


Figure 38.25. Kernel Density Dialog

The default uses a normal kernel density with a c value that minimizes the AMISE. Figure 38.26 displays normal kernel estimates with c = 0.7852 (the AMISE value) and c = 0.25. Small values of c (and hence small values of the smoothing parameter λ) provide jagged estimates as the curve more closely follows the data points. Large values of c provide smoother estimates. The **Mode** is the point with the largest estimated density. Use the slider to change the smoothing parameter, c.

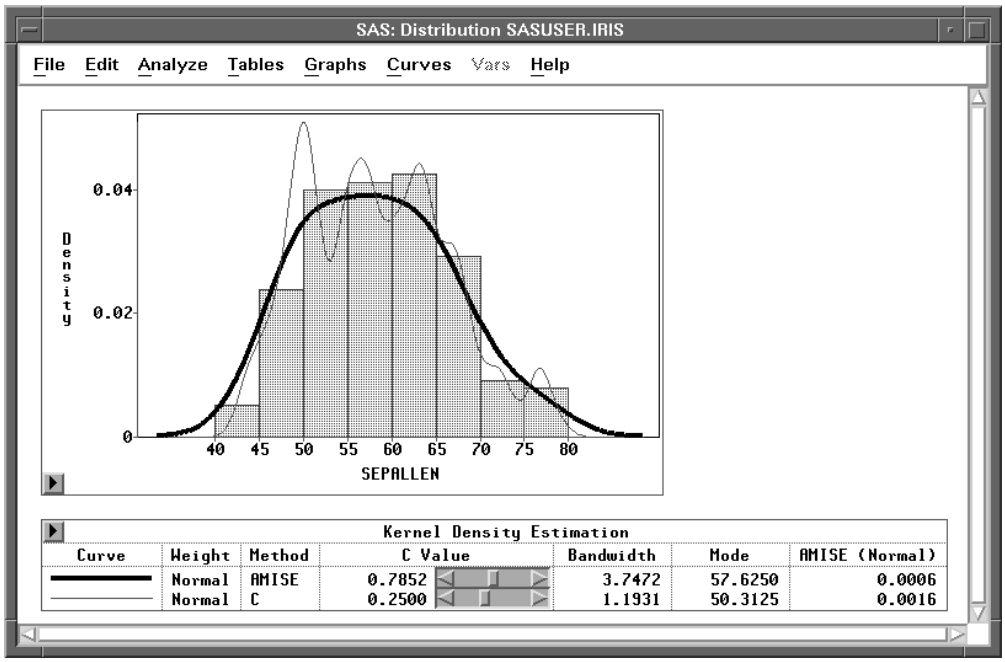


Figure 38.26. Kernel Density Estimation

Empirical CDF

The *empirical distribution function* of a sample, $F_n(y)$, is the proportion of observations less than or equal to y.

$$F_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \leq y)$$

where n is the number of observations, and $I(y_i \leq y)$ is an indicator function with value 1 if $y_i \leq y$ and with value 0 otherwise.

The Kolmogorov statistic D is a measure of the discrepancy between the empirical distribution and the hypothesized distribution.

$$D = \mathrm{Max_y}|\mathrm{F_n(y)} - \mathrm{F(y)}|$$

where $F(y)$ is the hypothesized cumulative distribution function. The statistic is the maximum vertical distance between the two distribution functions. The Kolmogorov statistic can be used to construct a confidence band for the unknown distribution function, to test for a hypothesized completely known distribution, and to test for a specific family of distributions with unknown parameters.

If you select a **Weight** variable, the weighted empirical distribution function is the proportion of observation weights for observations less than or equal to y.

$$F_w(y) = \frac{1}{\sum_i w_i}\sum_{i=1}^{n} w_i I(y_i \leq y)$$

CDF Confidence Band

The *confidence band* gives a confidence region for the population distribution. The critical values given by Feller (1948) for the completely specified hypothesized distribution are used to generate the confidence band. All parameters in the hypothesized distribution are known. The null hypothesis that the population distribution is equal to a given completely specified distribution is rejected if the hypothesized distribution falls outside the confidence band at any point.

You specify the confidence coefficient in the cumulative distribution options dialog or by choosing **Curves:CDF Confidence Band**.

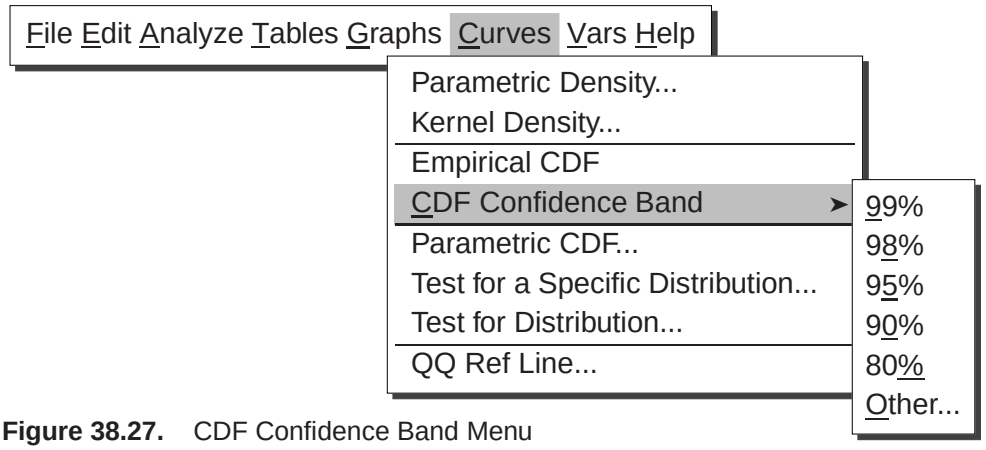


Figure 38.27. CDF Confidence Band Menu

Figure 38.28 displays an empirical distribution function and a 95% confidence band for the cumulative distribution function. Use the **Coefficient** slider to change the coefficient for the confidence band.

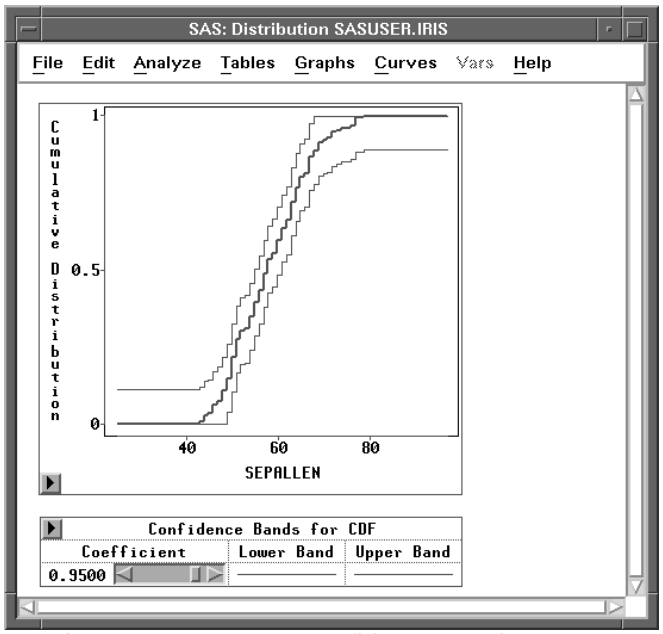


Figure 38.28. CDF Confidence Band

Parametric CDF

You can fit the normal, lognormal, exponential, and Weibull distributions to your data. You specify the family of distributions either in the cumulative distribution options dialog or from the **Parametric CDF Estimation** dialog after choosing **Curves:Parametric CDF** from the menu.

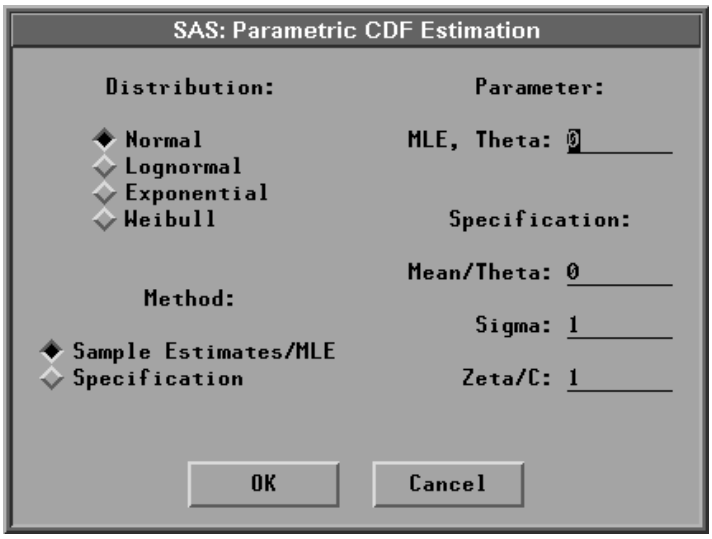


Figure 38.29. Parametric CDF Dialog

For the normal distribution, you can specify your own μ and σ parameters from the **Fit Parametric** menu. Otherwise, you can use the sample mean and standard deviation as estimates for μ and σ by selecting **Fit Parametric:Normal** in the cumulative distribution options dialog or by choosing **Distribution:Normal** and **Method:Sample Estimates/MLE** in the **Parametric CDF Estimation** dialog.

For the lognormal, exponential, and Weibull distributions, you can specify your own threshold parameter θ and have the remaining parameters estimated by the maximum-likelihood method, or you can specify all the distribution parameters in the **Parametric CDF Estimation** dialog. Otherwise, you can have the threshold parameter set to 0 and the remaining parameters estimated by the maximum-likelihood method. To do this, select **Lognormal**, **Exponential**, or **Weibull** in the Cumulative Distribution Output dialog or choose **Method:Sample Estimates/MLE** and **Parameter:MLE, Theta:0** in the **Parametric CDF Estimation** dialog.

If you select a **Weight** variable, only normal CDF can be created. For **Method:Sample Estimates/MLE**, $\overline{y}_w$ and s_w are used to display the cumulative distribution function with vardef=**WDF/WGT**; $\overline{y}_w$ and s_a are used with vardef=**DF/N**. For **Method:Specification**, the values in the entry fields **Mean/Theta** and **Sigma** are used to display the cumulative distribution function with vardef=**WDF/WGT**; the values of **Mean/Theta** and **Sigma**/$\sqrt{\overline{w}}$ are used with vardef=**DF/N**.

Figure 38.30 displays a normal distribution function with μ = 58.4333 (the sample mean) and σ = 8.2807 (the sample standard deviation); it also displays a lognormal distribution function with θ = 30 and σ and ζ estimated by the MLE.

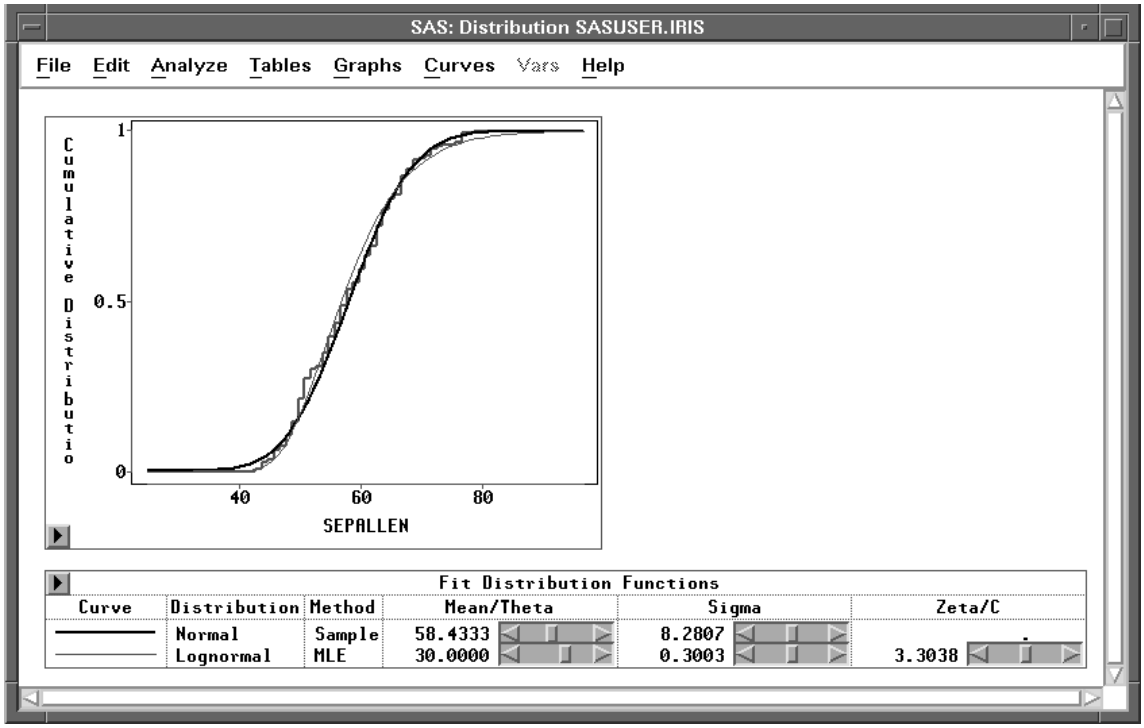


Figure 38.30. Parametric CDF

Use sliders to change the CDF estimate. When MLE is used for the lognormal, exponential, and Weibull distributions, changing the value of θ in the slider also causes the remaining parameters to be estimated by the MLE for the new θ.

Test for a Specific Distribution

You can test whether the data are from a specific distribution with known parameters by using the Kolmogorov statistic. The probability of a larger Kolmogorov statistic is given in Feller (1948). After choosing **Curves:Test for a Specific Distribution** from the menu, you can specify the distribution and its parameters in the **Test for a Specific Distribution** dialog.

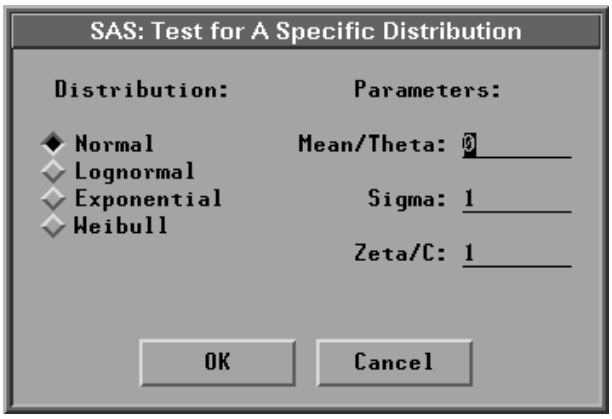


Figure 38.31. Test for a Specific Distribution Dialog

The default tests that the data are from a normal distribution with $\mu = 0$ and $\sigma = 1$. Figure 38.32 shows a test for a specified normal distribution ($\mu = 60$, $\sigma = 10$). Use sliders to change the distribution parameters and have the test results updated accordingly.

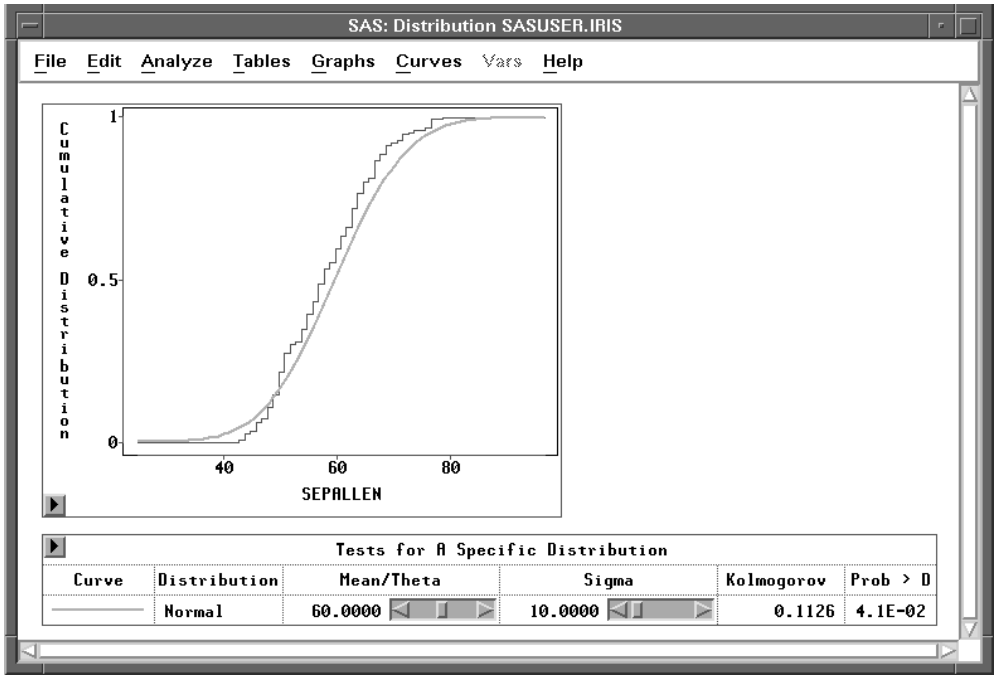


Figure 38.32. Test for a Specific Distribution

Test for Distribution

You can test that the data are from a specific family of distributions, such as the normal, lognormal, exponential, or Weibull distributions. You do not need to specify the distribution parameters except the threshold parameters for the lognormal, exponential, and Weibull distributions. The Kolmogorov statistic assesses the discrepancy between the empirical distribution and the estimated hypothesized distribution F.

For a test of normality, the hypothesized distribution is a normal distribution function with parameters μ and σ estimated by the sample mean and standard deviation. The probability of a larger test statistic is obtained by linear interpolation within the range of simulated critical values given by Stephens (1974).

For a test of whether the data are from a lognormal distribution, the hypothesized distribution is a lognormal distribution function with a given parameter θ and parameters ζ and σ estimated from the sample after the logarithmic transformation of the data, $\log(y - \theta)$. The sample mean and standard deviation of the transformed sample are used as the parameter estimates. The test is therefore equivalent to the test of normality on the transformed sample.

For a test of exponentiality, the hypothesized distribution is an exponential distribution function with a given parameter θ and a parameter σ estimated by $\overline{y} - \theta$. The probability of a larger test statistic is obtained by linear interpolation within the range of simulated critical values given by Stephens (1974).

For a test of whether the data are from a Weibull distribution, the hypothesized distribution is a Weibull distribution function with a given parameter θ and parameters c and σ estimated by the maximum-likelihood method. The probability of a larger test statistic is obtained by linear interpolation within the range of simulated critical values given by Chandra, Singpurwalla, and Stephens (1981).

You specify the distribution in the cumulative distribution options dialog or in the **Test for Distribution** dialog after choosing **Curves:Test for Distribution** from the menu, as shown in Figure 38.33. You can also specify a threshold parameter other than zero for lognormal, exponential, and Weibull distributions.

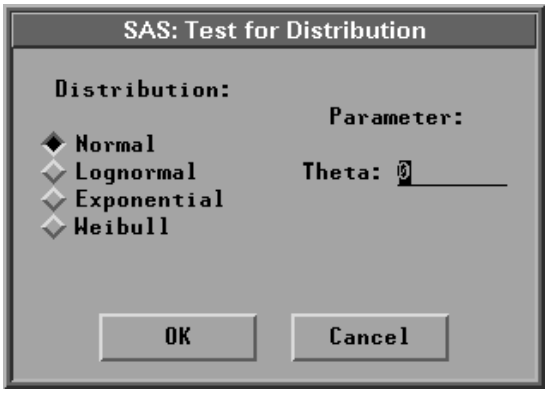


Figure 38.33. Test for Distribution Dialog

The default tests that the data are from a normal distribution. A test for normality and a test for lognormal distribution with $\theta = 30$ are given in Figure 38.34. You can use the **Mean/Theta** slider to adjust the threshold parameter, θ, for lognormal, exponential, and Weibull distributions.

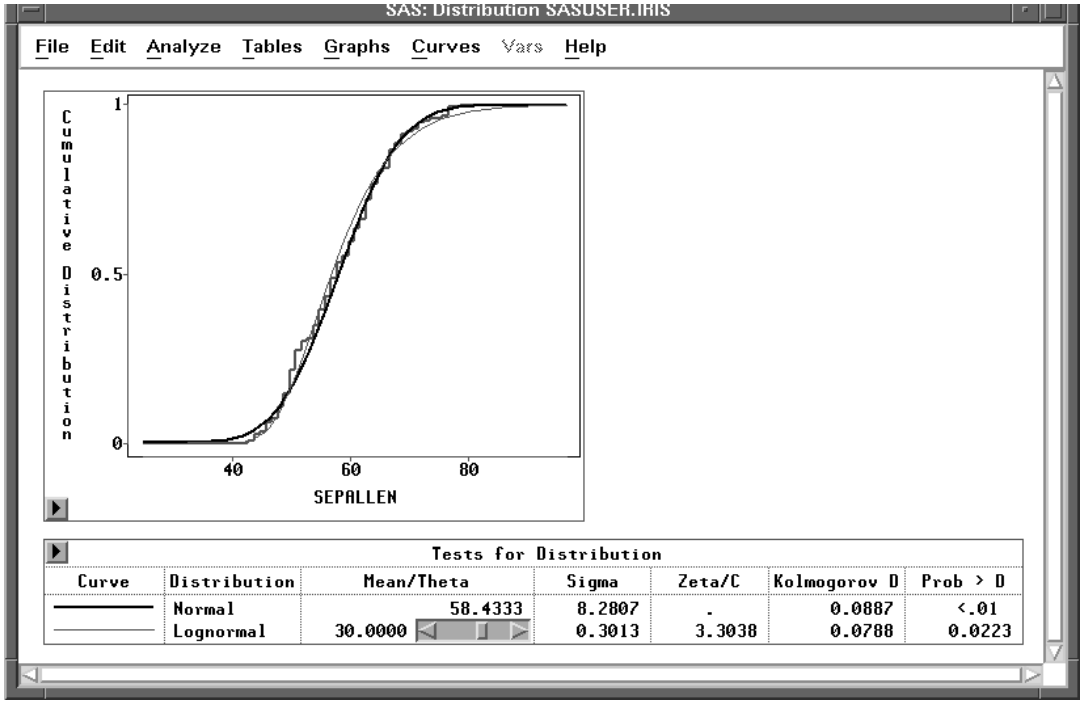


Figure 38.34. Tests for Distribution

QQ Ref Line

After choosing **Curves:QQ Ref Line**, you can use the **QQ Ref Line** dialog to add distribution reference lines to QQ plots.

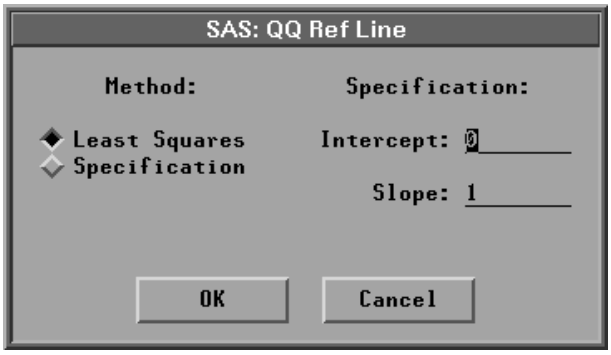


Figure 38.35. QQ Ref Line Dialog

The default adds a least squares regression line. You can also specify your own reference line by choosing **Method:Specification** and specifying both the intercept and slope.

If you select a **Weight** variable, you can add a weighted least squares regression line to the normal QQ plot. If the data are normally distributed with mean μ and standard deviation σ and if each observation has approximately the same weight (w_0), then the least squares regression line has approximately intercept μ and slope σ for vardef=WDF/WEIGHT and slope $\sigma/\sqrt{w_0}$ for vardef=DF/N.

A normal QQ plot with a least squares reference line is shown in Figure 38.36. Use the sliders to change the intercept and slope of the reference line.

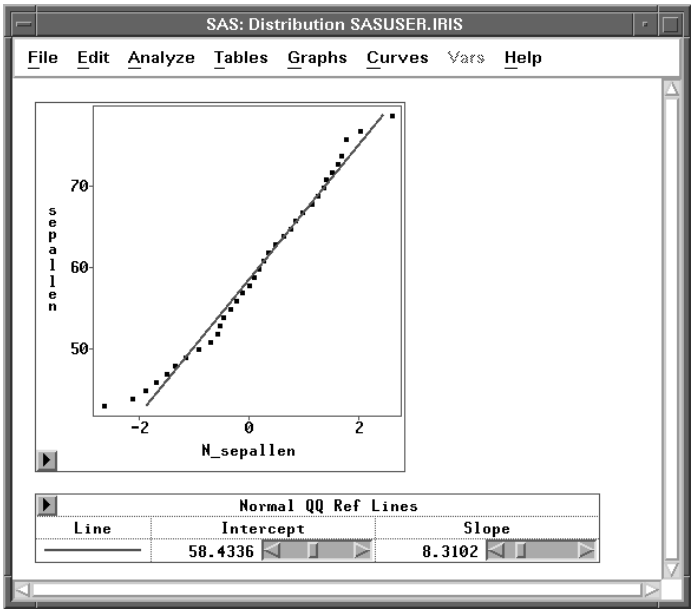


Figure 38.36. Normal QQ Plot with a Reference Line

Analysis for Nominal Variables

You can generate a frequency table, display a bar chart, and display a mosaic plot for each nominal variable in the distribution analysis, as shown in Figure 38.37.

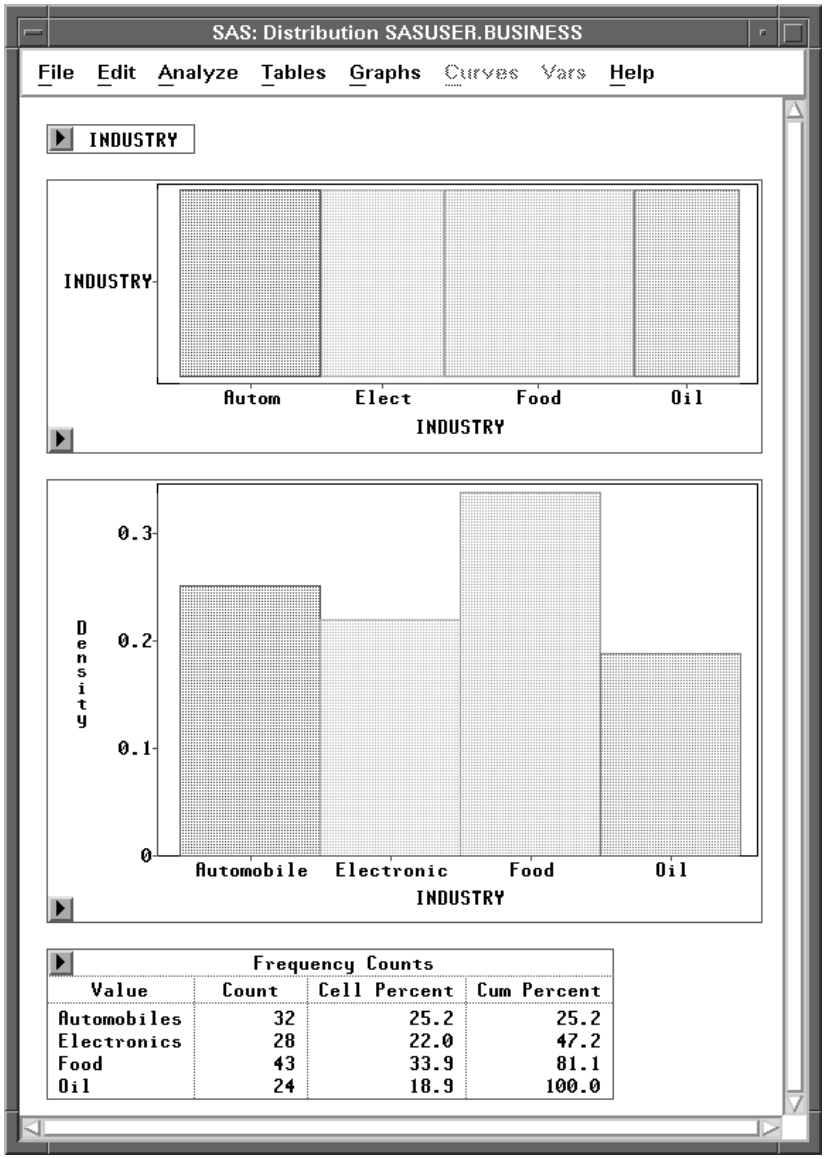


Figure 38.37. Nominal Variable Output

⊕ **Related Reading:** Bar Charts, Chapter 32.

⊕ **Related Reading:** Mosaic Plots, Chapter 33.

References

Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P.A. (1983), *Graphical Methods for Data Analysis*, Belmont, CA: Wadsworth International Group.

Chandra, M., Singpurwalla, N.D., and Stephens, M.A. (1981), "Kolmogorov Statistics for Tests of Fit for the Extreme-Value and Weibull Distributions," *Journal of the American Statistical Association*, 76, 729–731.

Conover, W.J. (1980), *Practical Nonparametric Statistics*, Second Edition, New York: John Wiley & Sons, Inc.

Croux, C. and Rousseeuw, P.J. (1992), "Time-Efficient Algorithms for Two Highly Robust Estimators of Scale," *Computational Statistics*, Volume 1, 411–428.

D'Agostino, R.B. and Stephens, M.A., Eds. (1986), *Goodness-of-Fit Techniques*, New York: Marcel Dekker, Inc.

Dixon, W.J. and Tukey, J.W. (1968), "Approximate Behavior of the Distribution of Winsorized t (Trimming/Winsorization 2)," *Technometrics*, 10, 83–98.

Epanechnikov, V.A. (1969), "Nonparametric Estimation of a Multivariate Probability Density," *Theory of Probability and Its Applications*, 14, 153–158.

Feller, W. (1948), "On the Kolomogorov-Smirnov Limit Theorems for Empirical Distributions," *Annals of Math. Stat.*, 19, 177–189.

Fisher, R.A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7, 179–188.

Hahn, G.J. and Meeker, W.Q. (1991), *Statistical Intervals: A Guide for Practitioners*, New York: John Wiley & Sons, Inc.

Hampel, F.R. (1974), "The Influence Curve and its Role in Robust Estimation," *Journal of the American Statistical Association*, 69, 383–393.

Iman, R.L. (1974), "Use of a t-statistic as an Approximation to the Exact Distribution of the Wilcoxon Signed Ranks Test Statistic," *Communications in Statistics*, 3, 795–806.

Johnson, N.L. and Kotz, S. (1970), *Continuous Univariate Distributions —I*, New York: John Wiley & Sons, Inc.

Lehmann, E.L. (1975), *Nonparametric: Statistical Methods Based on Ranks*, San Francisco: Holden-Day, Inc.

Rosenberger, J.L. and Gasko, M. (1983), "Comparing Location Estimators: Trimmed Means, Medians, and Trimean," in *Understanding Robust and Exploratory Data Analysis*, eds. D.C. Hoaglin, F. Mosteller, and J.W. Tukey, New York: John Wiley & Sons, Inc., 297–338.

Rousseeuw, P.J. and Croux, C. (1993), "Alternatives to the Median Absolute Deviation," *Journal of the American Statistical Association*, 88, 1273–1283.

Royston, P. (1992), "Approximating the Shapiro-Wilk W-Test for non-normality," *Statistics and Computing*, 2, 117–119.

Silverman, B.W. (1982), "Kernel Density Estimation using the Fast Fourier Transform," *Applied Statistics*, 31, 93–99.

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, New York: Chapman and Hall.

Smirnov, N. (1948) "Table for Estimating the Goodness of Fit of Empirical Distributions," *Annals of Math. Stat.*, 19, 279.

Stephens, M.A. (1974), "EDF Statistics for Goodness of Fit and Some Comparisons," *Journal of the American Statistical Association*, 69, 730–737.

Tukey, J.W. (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.

Tukey, J.W. and McLaughlin, D.H. (1963), "Less Vulnerable Confidence and Significance Procedures for Location Based on a Single Sample: Trimming/Winsorization 1," *Sankhya A*, 25, 331–352.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/INSIGHT User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999. 752 pp.

SAS/INSIGHT User's Guide, Version 8

ISBN 1–58025–490–X

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

1st printing, October 1999