Chapter 12
Examining Distributions

Chapter Table of Contents

Chapter 12
Examining Distributions

In Chapter 4, “Exploring Data in One Dimension,” you examined distributions using bar charts and box plots. In this chapter, you examine the distribution of an interval variable using graphs and statistical tables.

You can examine box plots and histograms of the data along with **Moments** and **Quantiles** tables. You can superimpose density curves on the histogram. You can carry out tests to determine whether the data are from specific parametric distributions, such as normal or lognormal.

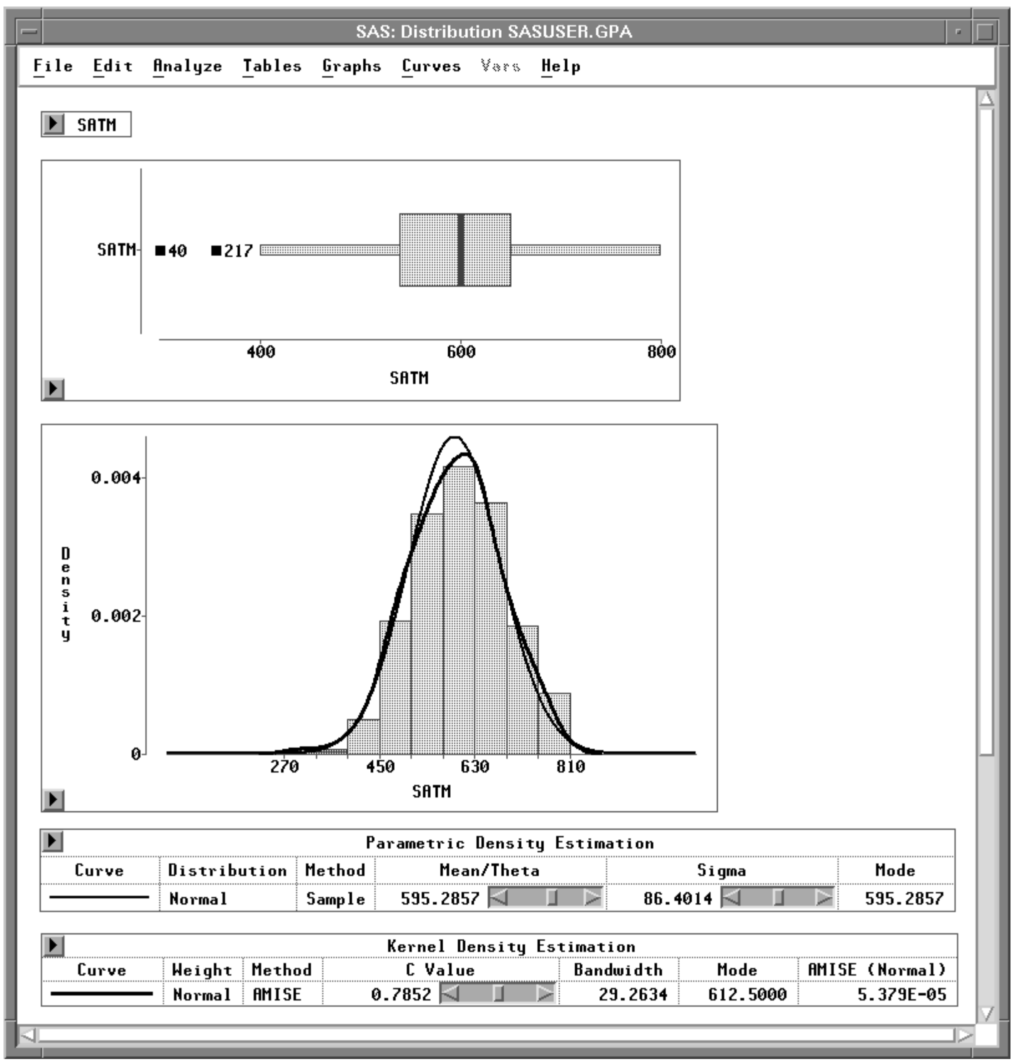


Figure 12.1. Distribution Analysis

Creating the Distribution Analysis

The *distribution* of a variable is the pattern of variation of its numerical values (Moore and McCabe 1989). In this example, you examine a distribution of scores on the mathematics portion of the SAT exam.

⟹ **Open the GPA data set.**

⟹ **Select the variable SATM by clicking on its name in the data window.**

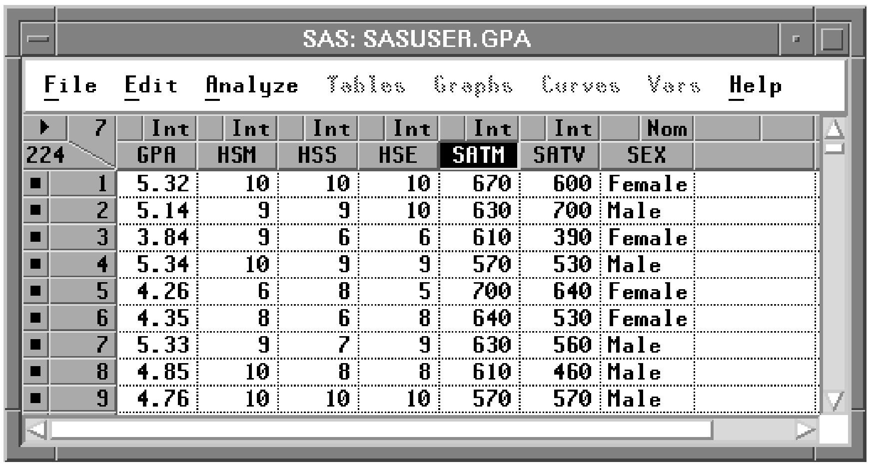


Figure 12.2. Data Window with **SATM** Selected

⟹ **Choose Analyze:Distribution (Y).**

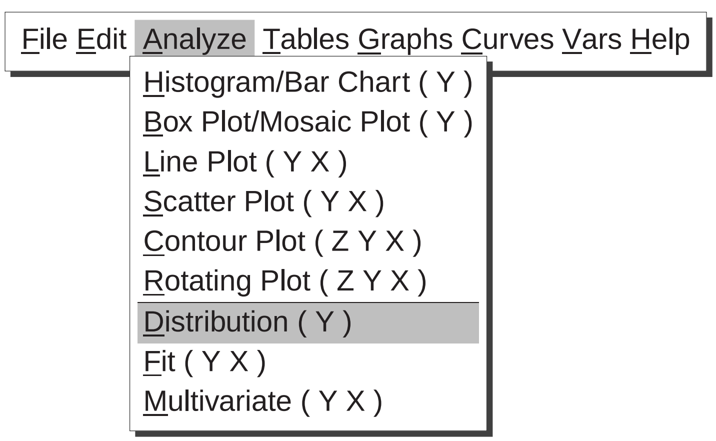


Figure 12.3. Analyze Menu

This creates a distribution window, as shown in Figure 12.4. A box plot, histogram, **Moments** table, and **Quantiles** table appear by default. With these graphs and tables, you can examine important features of a distribution.

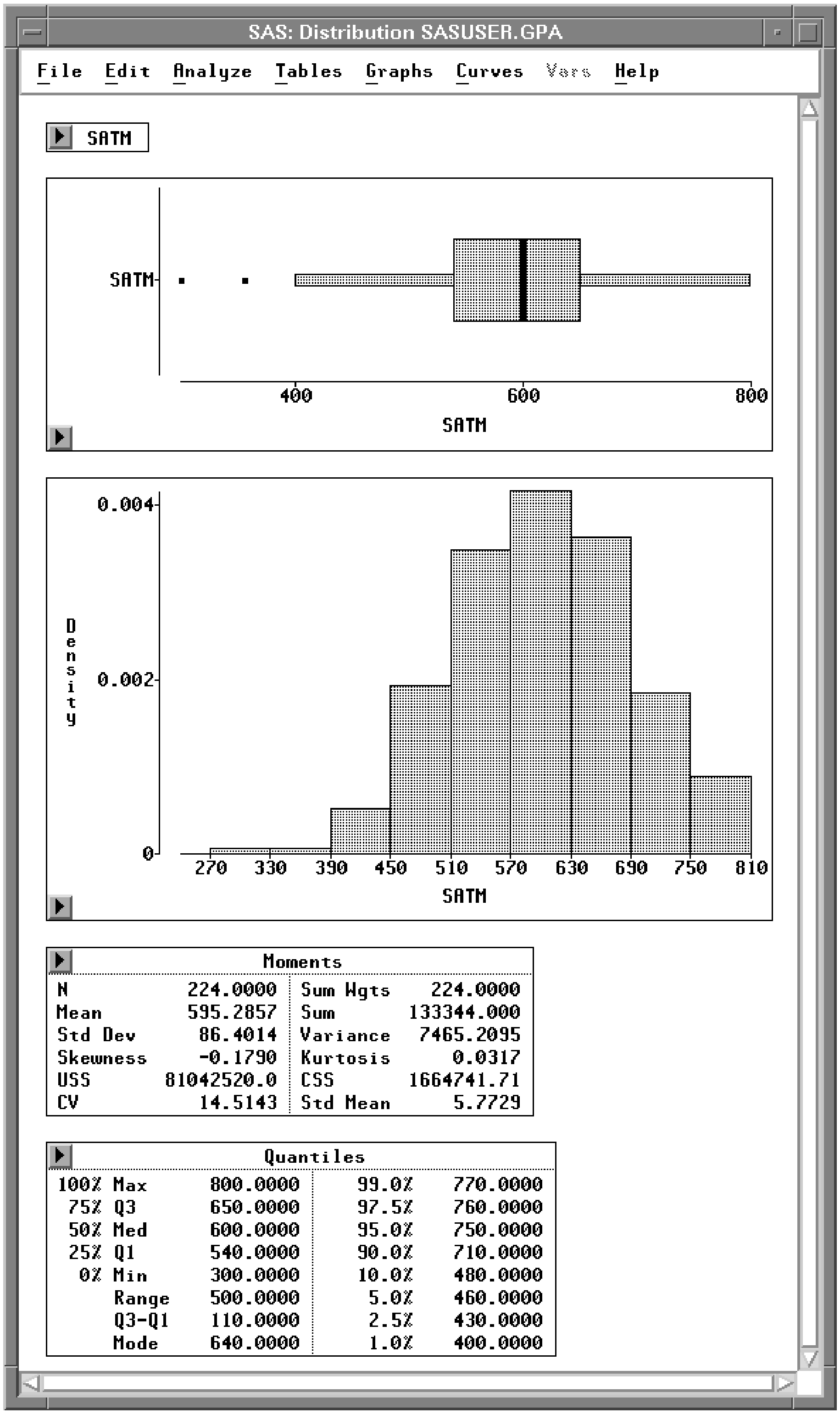


Figure 12.4. Distribution Analysis

Box Plot

A box plot is a schematic representation of a distribution. The vertical lines in the box mark the 25th, 50th, and 75th percentiles of the data. The pth *percentile* of a distribution is the value such that p percent of the observations fall at or below it. The 50th percentile is also called the *median*, and the 25th and 75th percentiles are called *quartiles*.

The narrow boxes extending to the left and right are called *whiskers*. Whiskers extend from the quartiles to the farthest observation not farther than 1.5 times the distance between the quartiles (the *interquartile range*). Beyond the whiskers, extreme observations are plotted individually.

The box plot gives a concise picture of the distribution and emphasizes any extreme values. This particular box plot appears fairly symmetric, with median around 600. You can see two extreme values.

⟹ **Identify the extreme observations by clicking on them.**

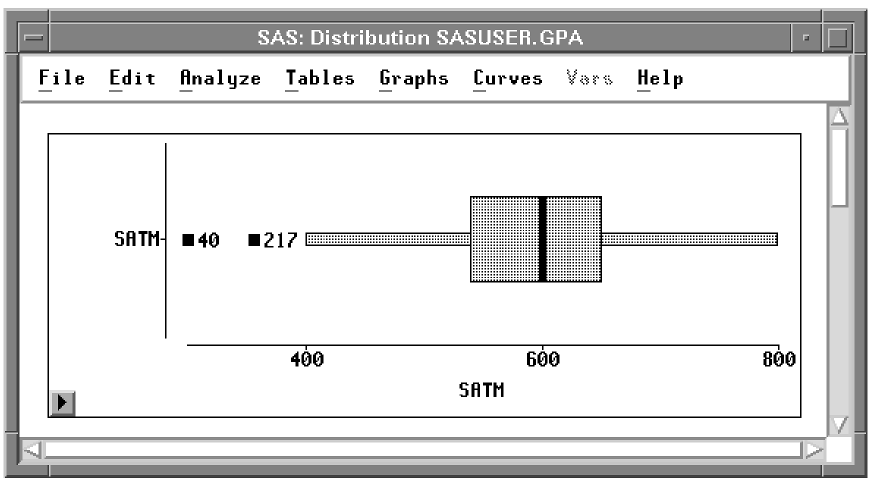


Figure 12.5. Identifying Extreme Observations

These are observations 40 and 217. When you click on them, the observations are selected in the box plot, the histogram, and the data window as well.

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

⟹ **Click in the upper left corner of the data window.**
This displays the data pop-up menu.

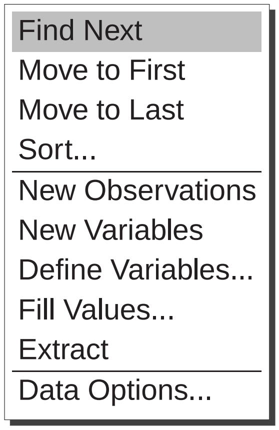


Figure 12.6. Data Pop-up Menu

⟹ **Choose Find Next from the pop-up menu.**
This scrolls the data window to the next selected observation, as shown in Figure 12.7. By choosing **Find Next** again, you can examine all values for the extreme observations.

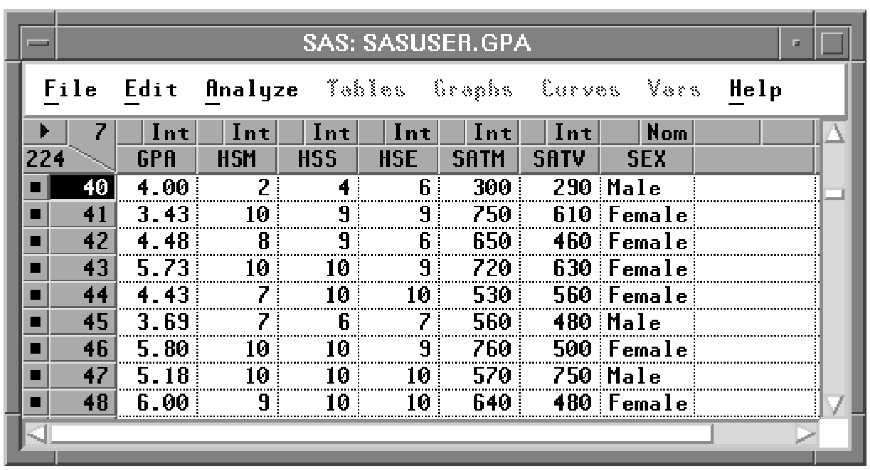


Figure 12.7. Extreme Observation in Data Window

Histogram

A *histogram* is a bar chart of an interval variable. In a histogram, the interval represented by a bar is called a *bin*. Instead of a frequency axis, histograms in a distribution analysis use a *density* axis to measure the fractional distribution over a given interval.

Examine the histogram of **SATM**. The shape of the distribution is fairly symmetric except for slight skewing in the left tail. The distribution's center is around 600.

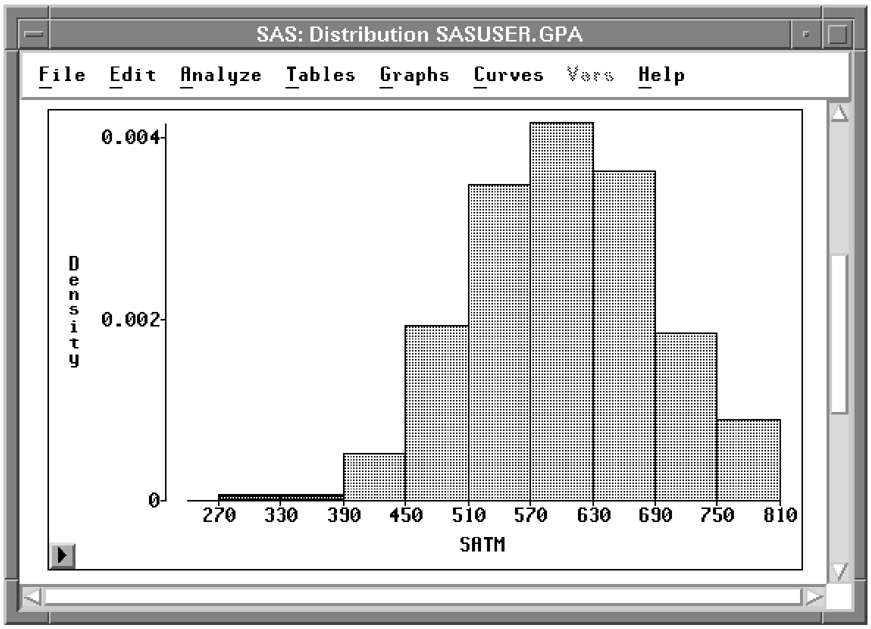


Figure 12.8. Histogram of **SATM**

A histogram is a good tool for visually examining the distribution. However, changes in the width and position of the bars can greatly affect your perception of the shape of the distribution. The histogram illustrated in Figure 12.8 is only one representation of the distribution of **SATM**. It is easy to change the bar widths and positions with SAS/INSIGHT software to explore many different histograms.

⟹ **Choose Edit:Windows:Tools.**
This displays the tools window, as shown in Figure 12.9.

⟹ **Click on the hand in the tools window.**
The cursor changes shape from an arrow to a hand.

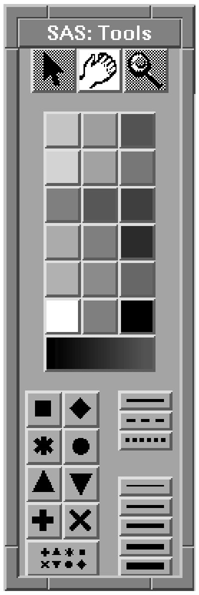


Figure 12.9. Tools Window

⟹ **Move the cursor back to the distribution window and click on the histogram.** This changes the width of the bars in proportion to the distance of the hand tool from the base of the bars. If the hand tool is close to the base of the bars, the bars are wide, as shown in Figure 12.10.

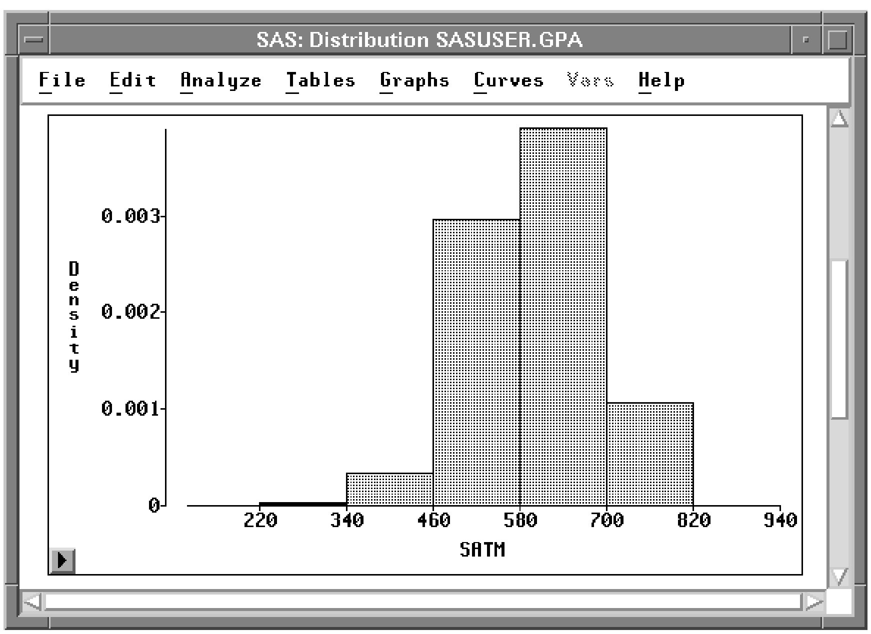


Figure 12.10. Clicking Close to the Base of the Bars

If the hand tool is far from the base of the bars, clicking makes the bars narrow, as shown in Figure 12.11.

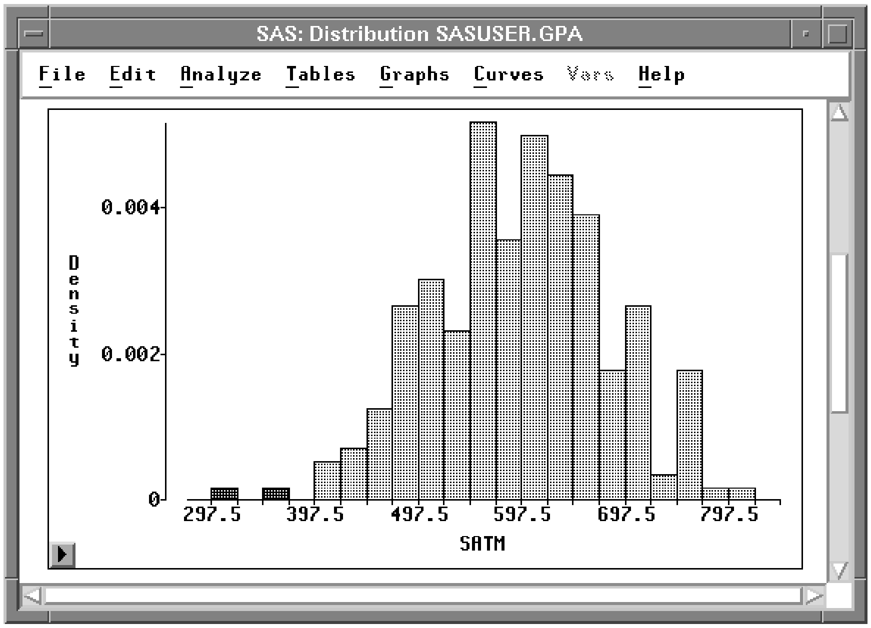


Figure 12.11. Clicking Far from the Base of the Bars

⟹ **Press the mouse button and hold it down as you move horizontally over the bars.**
Notice how the histogram changes as you move the hand. As you move horizontally, the bin width does not change, but the bins start at different locations. When the hand is at the left of the histogram, the bins start at an integral multiple of the bin width. When the hand moves toward the right, the bins are *offset* an amount proportional to the distance of the hand across the histogram.

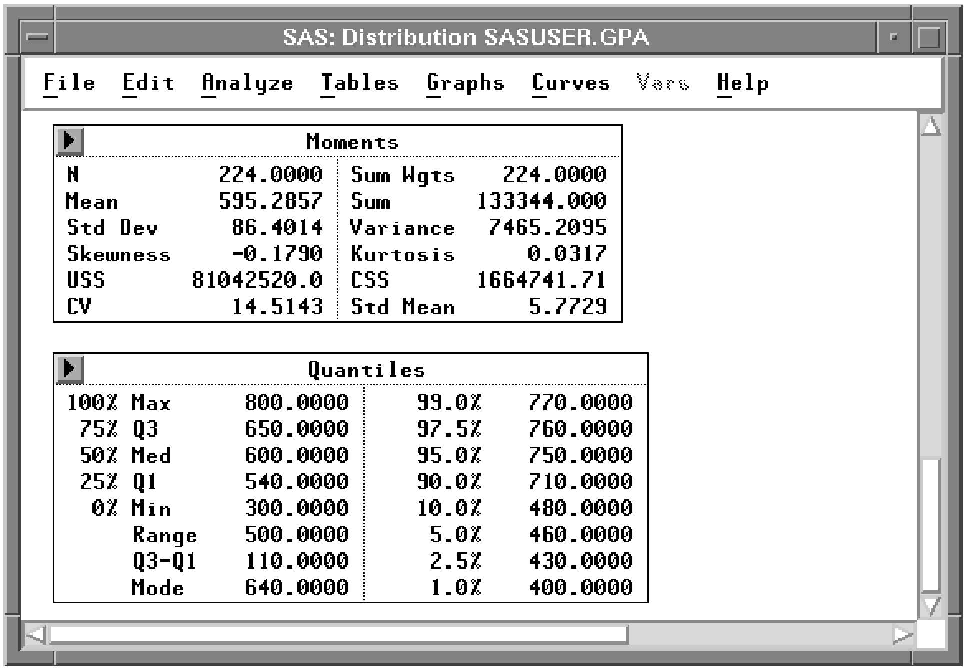
⟹ **Drag the hand horizontally and vertically in the histogram.**
Release the mouse button when you find a histogram that captures the dominant shape of the distribution.

⟹ **Click on the arrow in the tools window before proceeding.**

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

Moments and Quantiles Tables

The **Moments** and **Quantiles** tables give descriptive information that quantifies what you observe in the box plot and histogram.



Moments

N	224.0000	Sum Wgts	224.0000
Mean	595.2857	Sum	133344.000
Std Dev	86.4014	Variance	7465.2095
Skewness	-0.1790	Kurtosis	0.0317
USS	81042520.0	CSS	1664741.71
CV	14.5143	Std Mean	5.7729

Quantiles

100% Max	800.0000	99.0%	770.0000
75% Q3	650.0000	97.5%	760.0000
50% Med	600.0000	95.0%	750.0000
25% Q1	540.0000	90.0%	710.0000
0% Min	300.0000	10.0%	480.0000
Range	500.0000	5.0%	460.0000
Q3-Q1	110.0000	2.5%	430.0000
Mode	640.0000	1.0%	400.0000

Figure 12.12. Moments and Quantiles Tables

In the **Moments** table, **N** is the number of nonmissing observations, **Mean** is the arithmetic mean, **Std Dev** is the standard deviation, and **Variance** is the variance. **Skewness** and **Kurtosis** are both measures of the shape of the distribution.

Skewness is a measure of the tendency of the deviations from the mean to be larger in one direction than in the other. A positive value for **Skewness** indicates that the data are skewed to the right. A negative value indicates that the data are skewed to the left. The distribution of **SATM** is skewed slightly to the left, as you observed previously; thus, the value for **Skewness** is negative.

Kurtosis is primarily a measure of the heaviness of the tails of a distribution. Large values of **Kurtosis** indicate that the distribution has heavy tails. This statistic is standardized so that a normal distribution has a kurtosis of 0.

The **Quantiles** table gives information about the variability in the data as well as about the center of the data. Two distributions having the same center can look quite different if the variability in the two distributions is different. This variability is shown by the percentiles in the **Quantiles** table. The **Quantiles** table also shows the **Range** of the data, the interquartile range **Q3-Q1**, and the **Mode**.

Adding Density Estimates

A *cumulative distribution function* gives the proportion of the data less than each possible value. A *density function* is the derivative of the cumulative distribution function. *Density estimation* is the construction of an estimate of the density function from the observed data.

Histograms are one type of density estimation. You can also plot the density function to construct density curves. Density curves are sometimes preferred because they do not contain the discontinuous steps present in histograms.

Distribution (Y) provides two types of density estimation: parametric and kernel. In parametric estimation, the data are assumed to be from a known parametric family of distributions. The normal distribution is one of the most commonly used parametric distributions. Others include lognormal, exponential, and Weibull.

In kernel estimation, little is assumed about the functional form of the data. The data more completely determine the shape of the density curve. Kernel estimation is a type of nonparametric estimation.

Normal Density Curve

Begin by adding a normal density curve.

⟹ **Choose Curves:Parametric Density.**

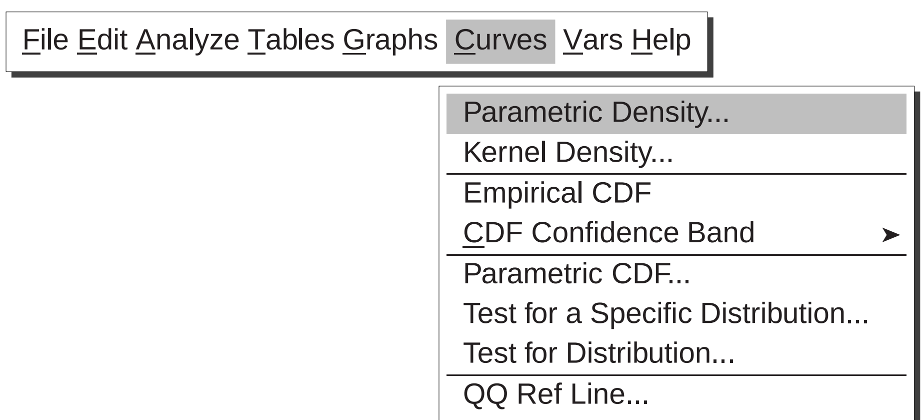


Figure 12.13. Normal Density Menu

This displays the parametric density estimation dialog in Figure 12.14. You can select one of four distribution families, and you can use sample parameter estimates or you can specify your own.

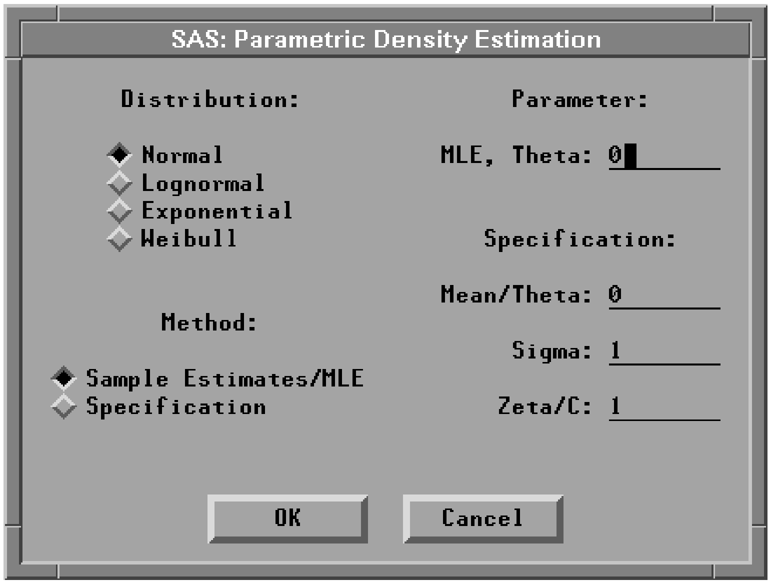


Figure 12.14. Parametric Density Estimation Dialog

⟹ **Click OK in the dialog.**

This requests the default density estimate: a normal distribution using the sample estimates as parameter values. The density curve is superimposed on the histogram, as illustrated in Figure 12.15.

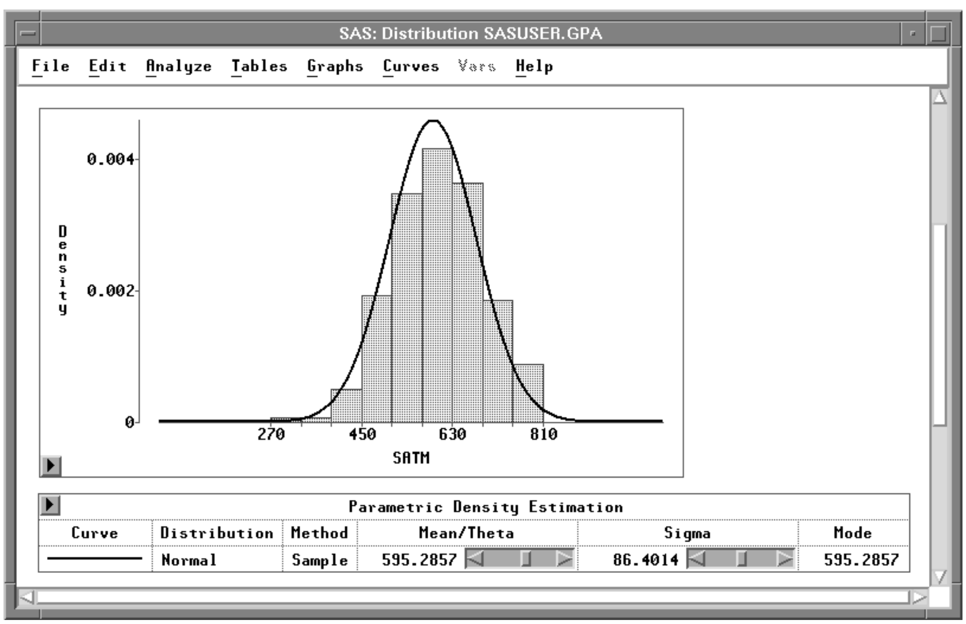


Figure 12.15. Parametric Normal Density Estimation

In addition, a **Parametric Density Estimation** table that contains parameter information appears in the window. You can change the specified parameters and the corresponding curve using the sliders next to the parameter values.

Note that the values of **Mean / Theta** and **Sigma** are equal to the sample **Mean** and **Std Dev** displayed in the **Moments** table illustrated in Figure 12.12. The density curve follows the shape of the distribution fairly well.

⟹ **Select the density curve.**
You can select the curve by clicking on either the curve in the histogram or the legend on the table. Both the curve and the legend become highlighted.

⟹ **Choose Edit:Delete.**

The selected curve and its associated table are deleted from the window.

Kernel Density Curve

A kernel density curve may follow the shape of the distribution more closely. To construct a normal kernel density curve, one parameter is required: the bandwidth λ. The value of λ determines the degree of smoothing in the estimate of the density function. You can either specify a value of λ, or you can let SAS/INSIGHT software find a value based on minimizing an estimate of the mean integrated square error (MISE).

⟹ **Choose Curves:Kernel Density.**

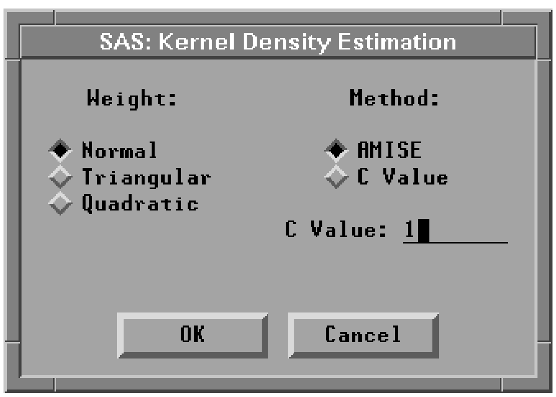


Figure 12.16. Kernel Density Estimation Dialog

⟹ **Click OK in the dialog.**

The kernel density curve is constructed with a bandwidth based on the approximated mean integrated square error (**AMISE**), and it provides a good visual representation of the distribution, as illustrated in Figure 12.17. A table containing the bandwidth and the **AMISE** is also added to the window.

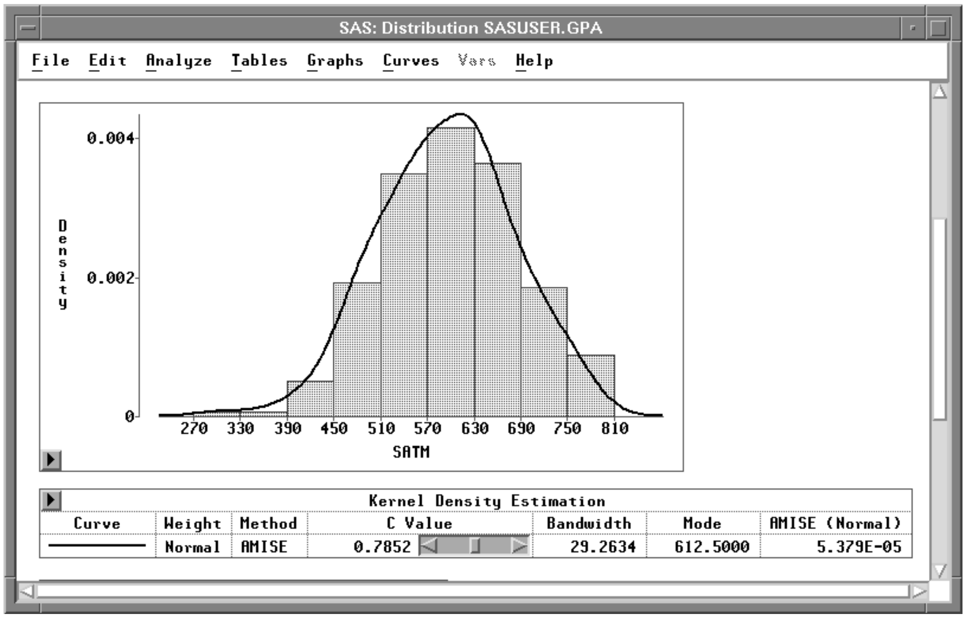


Figure 12.17. Kernel Density Estimate

The **C Value** slider in the table can be used to change the **C** value of the kernel estimate. You can use the slider in three ways:

- click the arrow buttons
- click within the slider
- drag within the slider

⟹ **Click the left arrow button in the slider.**
This decreases the **C** value by half. As the **C** value decreases, the density estimate becomes less smooth, as illustrated in Figure 12.18.

⟹ **Click within the slider, just to the right of the slider control.**
This moves the slider control to the position where you click. The **C** value is set to a value proportional to the slider position. On most personal computers, clicking within the slider is the fastest way to adjust a curve.

⟹ **Drag the slider control left and right.**
When you drag the slider, its speed depends on the number of data points, the type of curve, and the speed of your host. Depending on your host, you may be able to improve the speed of the dynamic graphics with an alternate drawing algorithm. To try this, choose **Edit:Windows:Graph Options**, and set the **Fast Draw** option.

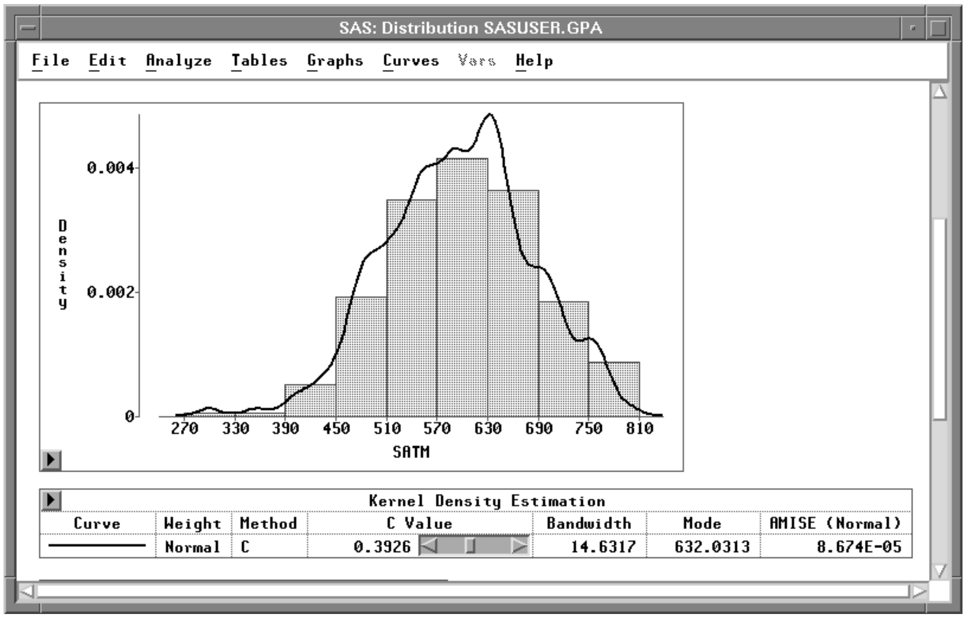


Figure 12.18. Kernel Density Estimate with a Smaller C Value

Testing Distributions

You can add a graph to examine the cumulative distribution function, and you can test for distributions by using the Kolmogorov statistic.

⟹ **Choose Curves:CDF Confidence Band:95%.**

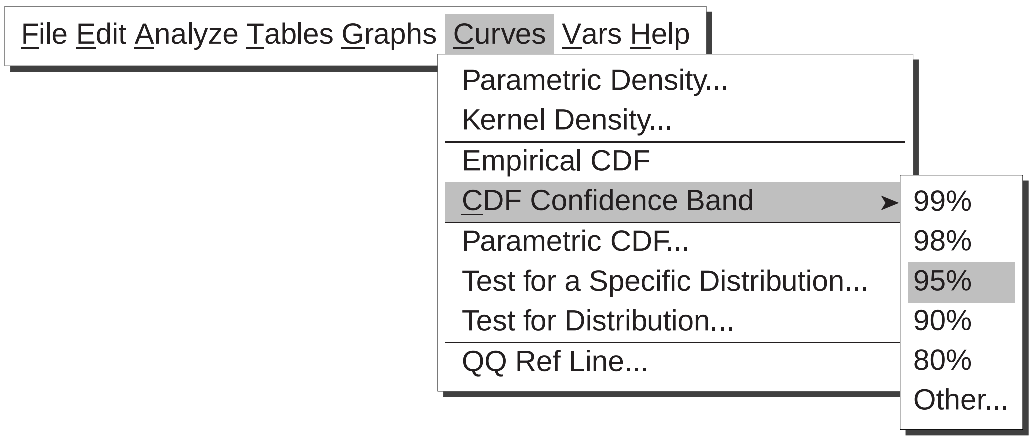


Figure 12.19. Confidence Band Menu

This adds a graph of the cumulative distribution function with 95% confidence bands, as illustrated in Figure 12.20.

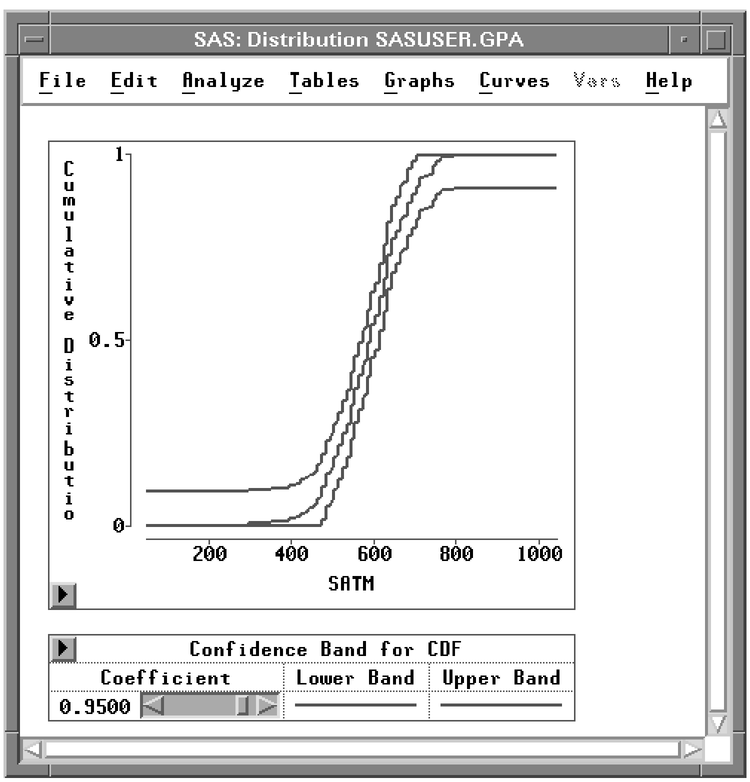


Figure 12.20. Cumulative Distribution Function

⟹ **Choose Curves:Test for Distribution.**
This displays the test for distribution dialog. The default settings test whether the data are from a normal distribution.

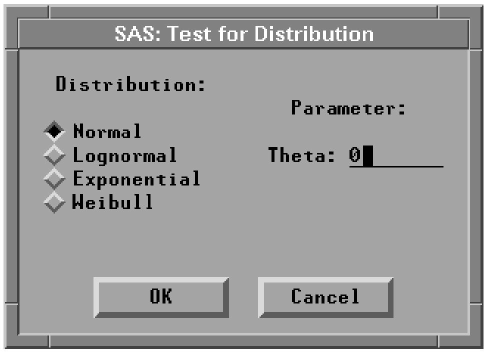


Figure 12.21. Test for Distribution Dialog

⟹ **Click OK in the dialog.**
This adds a curve to the graph and a **Test for Distribution** table to the window, as illustrated in Figure 12.22.

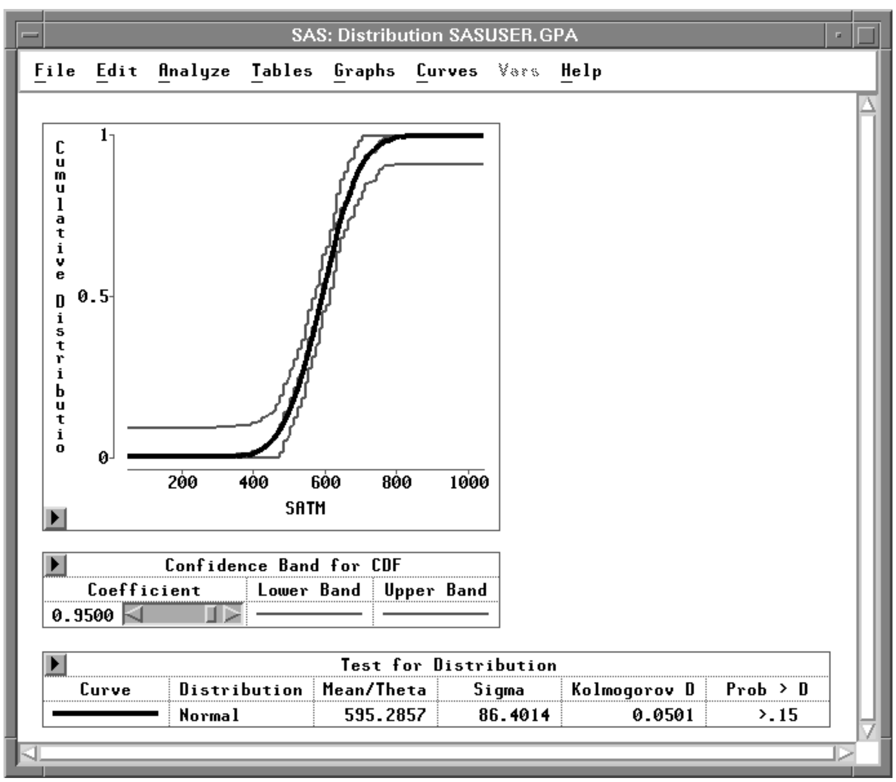


Figure 12.22. Test for Normal Distribution

The smooth curve in the graph represents the fitted normal distribution. It lies quite close to the irregular curve representing the empirical distribution function. The **Test for Distribution** table contains the mean (**Mean / Theta**) and standard deviation (**Sigma**) for the data along with the results of Kolmogorov's test for normality. This tests the null hypothesis that the data come from a normal distribution with unknown mean and variance. The *p*-value (**Prob > D**), also referred to as the *probability value* or *observed significance level*, is the probability of obtaining a *D* statistic greater than the computed *D* statistic when the null hypothesis is true. The smaller the *p*-value, the stronger the evidence against the null hypothesis. The computed *p*-value is large (**>0.15**), so there is no reason to conclude that these data are not normally distributed.

⊕ **Related Reading:** Distributions, Chapter 38.

References

Moore, D.S. and McCabe, G.P. (1989), *Introduction to the Practice of Statistics*, New York: W.H. Freeman and Company.

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