Are Histograms Giving You Fits? New SAS® Software for Analyzing Distributions

Nathan A. Curtis, SAS Institute Inc., Cary, NC

Abstract

Exploring and modeling the distribution of a data sample is a key step in many applications of statistics and data mining. This presentation will introduce you to software for creating high-resolution graphics displays of data distributions, including histograms, probability plots, and quantile-quantile plots, which have been added to the UNIVARIATE procedure in Version 8. Comparative versions of these displays are available, along with statistical methods for fitting parametric and nonparametric distributional models. For data involving multiple samples, you can create box plot displays with the new BOXPLOT procedure in SAS/STAT® software.

Introduction

The UNIVARIATE procedure, introduced in the late 1970s, produced "simple descriptive statistics for numeric variables" and differed from other SAS procedures "in the greater detail it provides on the distribution of a variable" (SAS Institute Inc., 1979). Designed to analyze univariate distributions, the procedure computed basic statistics such as the mean and standard deviation, provided detail on quantiles and extreme values, and generated several plots such as stem-and-leaf and probability plots. Because it was developed to print results to a line printer mode of output, PROC UNIVARIATE displayed statistical tables compactly and rendered plots using ASCII characters.

Until now, PROC UNIVARIATE has not provided a number of features for analyzing distributions, including confidence intervals on appropriate statistics, numerous goodness-of-fit statistics, distribution fitting, and a wide range of publication-quality graphics. Some of these capabilities have been available in the CAPABILITY procedure, which was designed for process capability analysis available in SAS/QC® software. However, these features were not available in Base SAS® or SAS/STAT software and were not presented in the context of general data analysis.

Version 8 SAS software addresses these deficiencies by providing many new enhancements for modeling and visualizing distributions. Many new statistics have been added to PROC UNIVARIATE, including numerous confidence limits, robust statistics, and empirical goodness-of-fit tests. New distribution fitting facilities enable you to fit a wide range of parametric distributions including the normal, lognormal, gamma, and Weibull. In addition, user comments and the advent of the Output Delivery System motivated a reorganization of the procedure output.

The UNIVARIATE procedure also provides new statements that enable you to visualize distributions through high-resolution graphic displays including histograms, probability plots, and quantile-quantile plots. In addition, Version 8 SAS/STAT software now includes the BOXPLOT procedure for generating boxplots, a convenient tool for comparing distributions of a quantitative variable across levels of a grouping variable.

Statistical Enhancements

The UNIVARIATE procedure now computes confidence limits for basic statistics and percentiles. You can produce confidence limits for the mean, standard deviation, and variance based on the assumption that the data are normally distributed. You can also request confidence limits for quantiles based on the normal distribution or a method that is distribution free.

If outliers are present in your data, Winsorized and trimmed means provide robust location estimates that are robust to outlying observations. The Winsorized mean is computed after the k smallest and largest observations are replaced with the $(k+1)$ smallest and largest observation, respectively, while the trimmed mean is computed after the k smallest and k largest observations are deleted from the sample. The UNIVARIATE procedure computes both Winsorized and trimmed means as well as robust measures of scale including Gini's mean difference, the mean absolute deviation (MAD), and the S_n and Q_n statistics proposed by Rousseeuw and Croux (1993).

PROC UNIVARIATE provides facilities for fitting and plotting continuous distributions including the normal, lognormal, exponential, gamma, beta, and Weibull. You can estimate distribution parameters by maximum likelihood, or you can specify specific parameter values. In addition, PROC UNIVARIATE can also smooth the data distribution using kernel density estimation.

When you fit a parametric distribution, PROC UNIVARIATE provides a series of goodness-of-fit tests and p-values based on the empirical distribution function (EDF). The EDF tests, including the Kolmogorov-Smimov, Anderson-Darling, and Cramer-von Mises statistics, are based on various measures of discrepancy between the empirical distribution function and the cumulative distribution function based on a specified distribution. The p-values of the goodness-of-fit tests vary depending on which parameters are known and which are estimated by the procedure.

Other new PROC UNIVARIATE options enable you to specify the location parameter value in the null hypothesis of a test of location, request a table of all possible modes, and control the number of listed extreme values and observations.

Example: Octane Requirements

The following example illustrates the basic features of the UNIVARIATE and HISTOGRAM statements. An automotive industry study was carried out to assess the octane requirements of a group of customer-owned cars as determined by trained raters and the customers themselves; refer to Rodriguez and Taniguchi (1980). Based on previous studies, a significant fraction of customers experienced knock on gasoline with an average octane number of 92.6. Consequently, a preliminary stage of the analysis explored the distribution of customer and expert octane requirements.

The following statements create a SAS data set called OCTANE that contains the customer and expert rater requirements.

```
data octane;
   input Expert Customer;
   datalines;
94.5 92.0
94.0 88.0
94.0 90.0
93.0 93.0
88.0 88.0
...  ...
;
run;
```

The following statements invoke the UNIVARIATE procedure and use the Output Delivery System (ODS) to produce a selected list of tables.

```
ods select BasicMeasures
           BasicIntervals
           RobustScale
           Quantiles;

proc univariate data=univar.octane
                CIBASIC
                CIPCTLNORMAL
                ROBUSTSCALE;
   var customer;
run;

ods select all;
```

The ODS SELECT statement identifies four output objects for the selection list of tables printed in the Output window: basic statistical measures, basic confidence limits, quantiles, and robust measures of scale. By displaying only the tables of interest, you can gain a better understanding of the data without wading through tables of irrelevant statistics to find the results you need.

The CIBASIC and CIPCTLNORMAL options request confidence limits for the basic statistics and quantiles, respectively. These confidence limits are based on the assumption that the data are normally distributed. The ROBUSTSCALE option produces a table with robust measures of scale including Gini's mean difference and the IQR. The VAR statement identifies CUSTOMER as the analysis variable, and ODS SELECT ALL resets the selection list to include all tables produced by subsequent analyses.

Basic Statistical Measures

Location		Variability	
Mean	88.36320	Std Deviation	3.90977
Median	90.70000	Variance	15.28634
Mode	88.00000	Range	21.50000
		Interquartile Range	5.00000

Figure 1. Basic Statistical Measures

The UNIVARIATE procedure now prints basic statistical measures of location and variability in the "Basic Statistical Measures" table as displayed in Figure 1. Although statistics such as the mean are also included in the "Moments" table, the new layout provides a more readable display of the basic statistics without being cluttered by other statistics that may be less relevant to your analysis.

Basic Confidence Limits Assuming Normality

Parameter	Estimate	95% Confidence Limits	
Mean	88.36320	85.85635	90.87005
Std Deviation	3.90977	3.58279	4.30295
Variance	15.28634	12.83689	18.51537

Figure 2. Basic Confidence Limits

In response to user requests, the UNIVARIATE procedure now provides confidence intervals for a variety of distributional parameters and quantiles. You can also specify the confidence level and select the type of confidence interval, such as a two-sided, upper, or lower interval.

Robust Measures of Scale

Measure	Value	Estimate of Sigma
Interquartile Range	5.000000	3.706504
Gini's Mean Difference	4.290242	3.890751
MAD	2.300000	3.409980
Sn	3.657060	3.711520
Qn	4.443800	4.417030

Figure 3. Robust Measures of Scale

Figure 3 displays the "Robust Measures of Scale" table that includes the robust statistics with their corresponding estimates of σ. In this case, the robust estimates of σ range from 3.41 to 4.41 and are similar to the traditional estimate of 3.91.

Quantiles (Definition 5)

Quantile	Estimate	95% Confidence Limits Assuming Normality	
100% Max	98.2		
99%	94.0	94.55929	96.51715
95%	92.8	92.07207	93.68087
90%	91.0	90.73372	92.10435
75% Q3	89.0	88.46288	89.58682
50% Median	86.7	85.85635	86.87006
25% Q1	84.0	83.13959	84.26552
10%	80.8	80.62205	81.99269
5%	79.0	79.05554	80.55434
1%	77.0	76.20925	78.16712
0% Min	76.6		

Figure 4. Quantiles with Confidence Limits

The "Quantiles" table displays a wide range of quantiles along with confidence limits based on the assumption that the data are normally distributed. By specifying the CIPCTLDF option, you can also request confidence limits for quantiles based on a method that is distribution-free. The UNIVARIATE procedure uses order statistics (ranks)

to compute distribution-free confidence limits as described in Hahn and Meeker (1991).

Histograms

A *histogram* is a bar chart of a continuous variable where each bar represents observations contained in an interval called a *bin*. Histograms are useful for visualizing a data distribution, suggesting probability models or transformations, and detecting outliers and unusual behavior (Snee and Pfeifer 1983). You can create histograms using high-resolution graphics with the new HISTOGRAM statement in the UNIVARIATE procedure.

The following statements produce a histogram of the customer octane requirements.

```
proc univariate data=octane;
   histogram customer;
run;
```

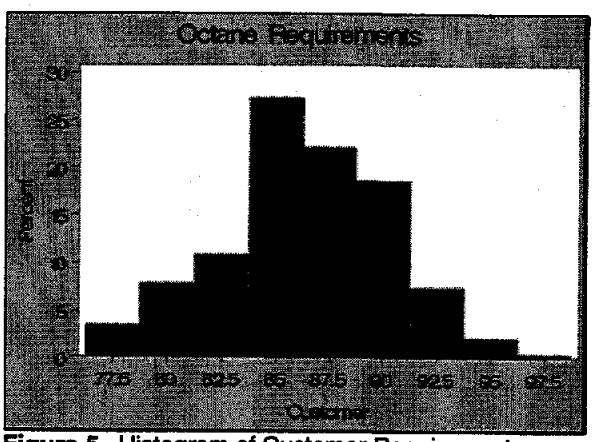


Figure 5. Histogram of Customer Requirements

By default, the UNIVARIATE procedure determines the initial number of midpoints for the data based on the total number of observations using the method of Terrell and Scott (1985). However, changes in bar width and position can greatly affect your perception of the distribution's shape. You can control the width of the histogram bars by specifying midpoint values. The width of the histogram bars is determined by the difference between consecutive midpoints.

```
proc univariate data=octane;
   histogram customer
             / midpoints = 76 to 100 by 2;
run;
```

The MIDPOINTS option specifies how to determine the midpoints for the histogram intervals. The values 76 to 100 by 2 define a range of midpoints that extend at each end by half the bar width and cover the entire data range. Note that PROC UNIVARIATE requires that histogram bins be equally sized.

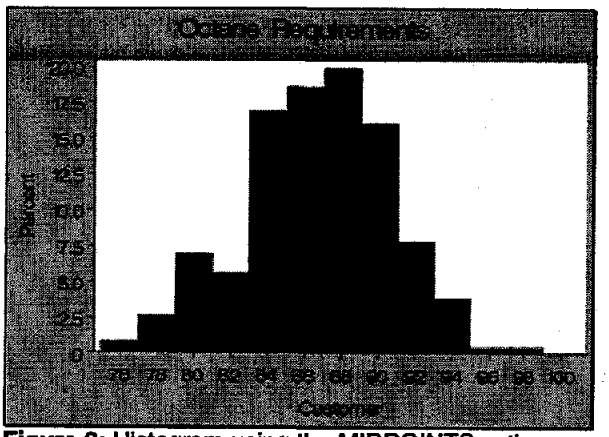


Figure 6: Histogram using the MIDPOINTS option

The two plots exhibit subtle differences, such as the slight depression in the second plot. However, the perturbations are most likely due to sampling variation and both histograms indicate that the data is roughly normally distributed.

Nonparametric Density Estimates

Sometimes, a fitted curve based on a parametric distribution does not effectively summarize a data distribution. For example, a normal curve will not adequately represent a distribution that is multimodel. In this case, you can use nonparametric kernel density estimation to obtain a smooth density estimate that accounts for this characteristic. Kernel density estimation is a technique that averages a kernel function across observations to create a smooth approximation. Using the KERNEL option, you can superimpose kernel density estimates on a histogram to visualize these features using smoother data distribution. The following statements fit a kernel density to both the customer and expert ratings.

```
proc univariate data=octane noprint;
   histogram customer expert /
      kernel ( k = normal
               c = 0.8
               w = 2.5
               color = green )
      normal ( mu = est
               sigma = est
               color = red
               w = 2.5 )
      midpoints = 76 to 100 by 2;
run;
```

The KERNEL option specifies that the UNIVARIATE procedure produce kernel density estimates of the customer and expert ratings. The K=option requests a normal kernel function, the C=0.8 option specifies the standardized bandwidth parameter, and the W=2.5 option defines the width of the density curves. The NORMAL option requests an additional fitted normal curve.

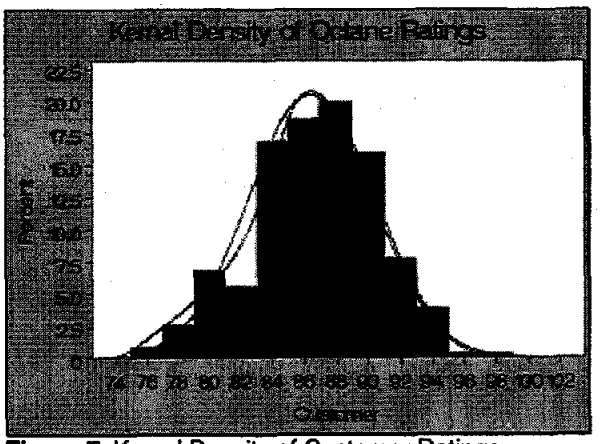


Figure 7. Kernel Density of Customer Ratings

For the customer ratings, the kernel density estimate provides a fit similar to the normal fitted curve. The fact that the kernel estimate does not drop at customer rating equal to 82 suggests that the depressed histogram bar is an artifact of binning and sampling variation.

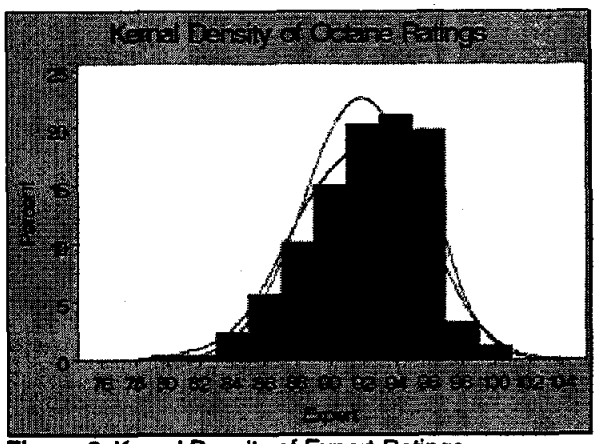


Figure 8. Kernel Density of Expert Ratings

The distribution of expert ratings is clearly skewed to the left, and the normal density estimate fits the distribution poorly. On the other hand, the kernel density estimate fits the distribution more closely than the parametric distribution by providing a smooth estimate that accounts for the skewness.

The UNIVARIATE procedure is designed to model single variables and does not address bivariate and more complex relationships between variables. However, you can compute kernel density estimates of univariate and bivariate densities using the KDE procedure now available in SAS/STAT software. PROC KDE provides several methods for bandwidth selection and enables you to compute contours of the estimated density function for subsequent plotting or analysis. Figure 14 provides an example of a bivariate kernel density estimate of customer and expert ratings using results from the KDE procedure. The display in Figure 9 was produced by the `%surface` macro, available for download from Technical Support at www.sas.com/techsup/download/stat/.

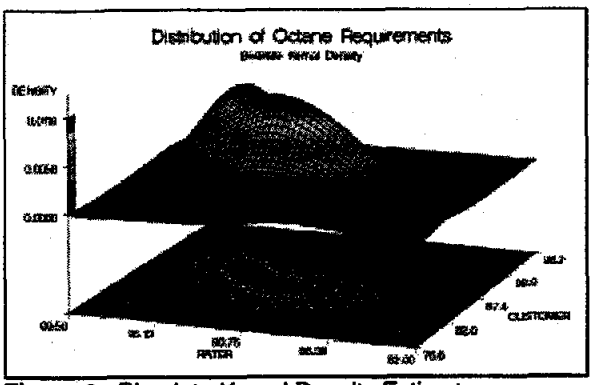


Figure 9. Bivariate Kernel Density Estimate

Modeling Distributions

The UNIVARIATE procedure provides methods for assessing and modeling the distribution of a variable including testing goodness-of-fit, generating graphics, and fitting distributions such as the normal, lognormal, gamma, exponential, and Weibull.

Goodness-of-Fit Tests

When you request a fitted parametric distribution using the HISTOGRAM statement, the procedure computes goodness-of-fit test statistics for the null hypothesis that the values of the analysis variable are a random sample from the specified theoretical distribution. If the sample size is less that 2000, PROC UNIVARIATE computes the ratio of the best estimator of the variance to the usual corrected sum of squares estimator of the variance known as the Shapiro-Wilk statistic (Shapiro and Wilk, 1965). In addition, the procedure also computes the Kolmogorov-Smirnov, Anderson-Darling, and Cramer-von Mises EDF statistics.

Example: Steel Rod Diameters

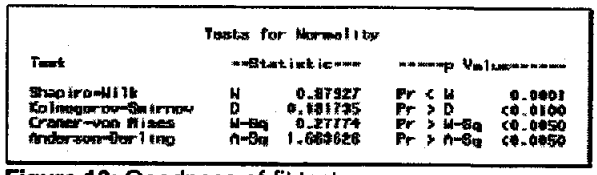
The data for the following example arose from a sample of diameter measurements of 50 steel rods. As a preliminary analysis, you decide to evaluate whether or not the diameter of the rods is normally distributed. The following statements produce the RODS data set that contains the 50 diameter measurements.

```
data rods;
    input diameter @@;
    label diameter='Diameter in mm';
    datalines;
5.501  5.251  5.404  5.366  5.445
5.576  5.607  5.200  5.977  5.177
...
;
run;
```

You can assess whether rod diameters are normally distributed by using the following statements.

```
proc univariate data=rods normal;
   histogram diameter /
      normal (mu=est sigma=est)
      midpoints = 5 to 6.30 by 0.15;
run;
```

The NORMAL option in the PROC UNIVARIATE statement requests tests for normality that include the Shapiro-Wilk test and a series of EDF goodness-of-fit tests. The NORMAL option in the HISTOGRAM statement displays a fitted normal density curve, and the MU=EST and SIGMA=EST options specify that μ and σ are to be estimated from the sample.



Tests for Normality

Test	--Statistic--		----p Value----	
Shapiro-Wilk	W	0.87927	Pr < W	0.0001
Kolmogorov-Smirnov	D	0.181735	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.277774	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.669626	Pr > A-Sq	<0.0050

Figure 10: Goodness-of-fit tests

Based on a Shapiro-Wilk statistic W=0.879 with a p-value of 0.0001, you reject the null hypothesis and conclude that the diameter of the rods are not normally distributed. The Kolmogorov-Smirnov, Anderson-Darling, and Cramer-von Mises statistics also result in p-values less than 0.01, which confirm the conclusion that the data are not normally distributed.

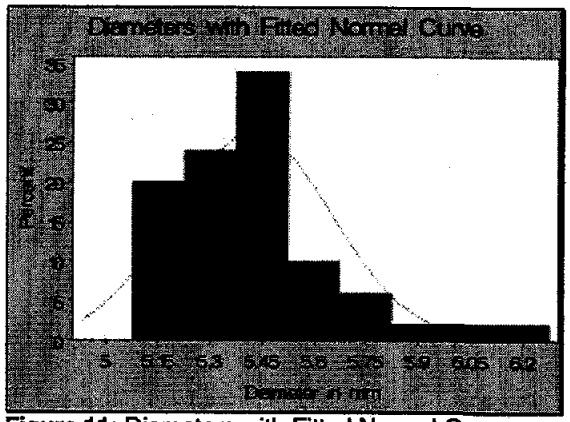


Figure 11: Diameters with Fitted Normal Curve

The histogram of rod diameters is skewed to the right and the normal fitted curve does not fit the histogram well. This also suggests that the diameter measurements are not normally distributed.

Probability Plots and Q-Q Plots

Probability plots and quantile-quantile (Q-Q) plots are useful for comparing ordered values of a variable with percentiles or quantiles of a specified theoretical distribution such as the normal. If the data distribution matches the theoretical distribution, the points on the plot form a linear pattern. Thus, you can use these plots to determine how well a theoretical distribution models a set of measurements. Q-Q plots are more useful than probability plots for graphically estimating location and scale parameters since the x-axis is scaled linearly, while probability plots are convenient for estimating percentiles and probabilities.

There are many reasons why point patterns in probability plots and Q-Q plots may not be linear. The following table summarizes many of the commonly encountered departures from linearity in these plots.

Pattern	Interpretation
All but a few points fall on a line	Outliers in the data
Left end of the pattern is below the line while the right end of the pattern is above the line	Symmetric, long tails at both ends
Left end of the pattern is above the line while the right end of the pattern is below the line	Symmetric, short tails at both ends
Curved pattern with slope increasing from left to right	Skewed to right
Curved pattern with slope decreasing from left to right	Skewed to left
Staircase pattern	Data have been rounded or may be discrete

Table 1: Plot Diagnostics

The UNIVARIATE procedure can create high-resolution graphic probability plots and Q-Q plots and can superimpose a reference line that corresponds to specified or estimated parameters of the theoretical distribution. The following statements produce a probability plot of steel rod diameters.

```
proc univariate data=rods noprint;
   probplot diameter /
      normal (mu=est sigma=est);
run;
```

The PROBPLOT statement creates a probability plot of the DIAMETER variable. The normal option specifies that the percentiles from the normal distribution, and MU=EST and SIGMA=EST suboptions add a reference line corresponding to the normal distribution with mean and standard deviation estimated from the sample.

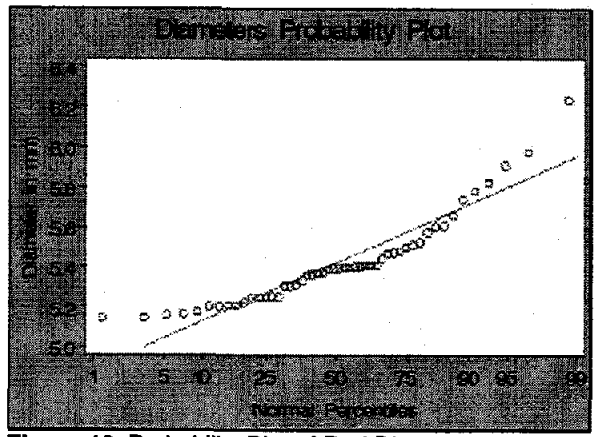


Figure 12. Probability Plot of Rod Diameters

If the rod diameter measurements came from a normal distribution, the points would tend to follow the superimposed distribution reference line. However, the

point pattern is curved with an increasing slope left to right, which indicates that the distribution is skewed to the right.

The lognormal distribution is a useful alternative for fitting data that are skewed to the right, and it may be suitable for modeling the rod diameter data. The lognormal distribution is often used in the biological and physical sciences to model sizes of various quantities and is closely linked to normal distribution. If, for some value θ, $\ln(X-\theta)$ has a normal distribution with mean μ and variance σ^2, then X has a lognormal distribution with threshold parameter θ, scale parameter $\zeta = \exp(\mu)$, and shape parameter σ.

You can create a series of Q-Q plots to visually estimate the shape parameter of the lognormal distribution. An appropriate value for the shape parameter produces an approximately linear pattern with intercept θ and slope $\exp(\zeta)$. The following statements produce a series of Q-Q plots based on increasing values of σ.

```
proc univariate data=rods noprint;
   qqplot diameter / lognormal
      (sigma=0.2 to 0.8 by 0.3
       theta=est zeta=est);
run;
```

The QQPLOT statement with the LOGNORMAL option creates a quantile-quantile plot that compares ordered values of the DIAMETER variable with quantiles of lognormal distribution. By specifying more than one value of σ using the SIGMA= suboption, the procedure creates multiple Q-Q plots that enable you to select the value of σ that most nearly linearizes the point pattern. The THETA=EST and ZETA=EST options add a reference line that corresponds to estimated values of θ and ζ. You can also specify specific values for θ and ζ with the THETA= and ZETA= suboptions.

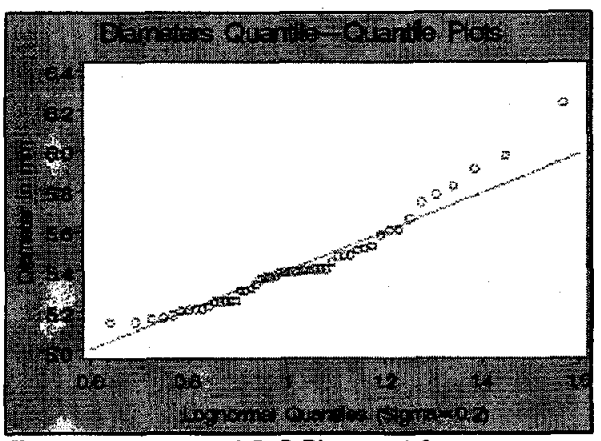


Figure 13: Lognormal Q-Q Plot, $\sigma = 0.2$

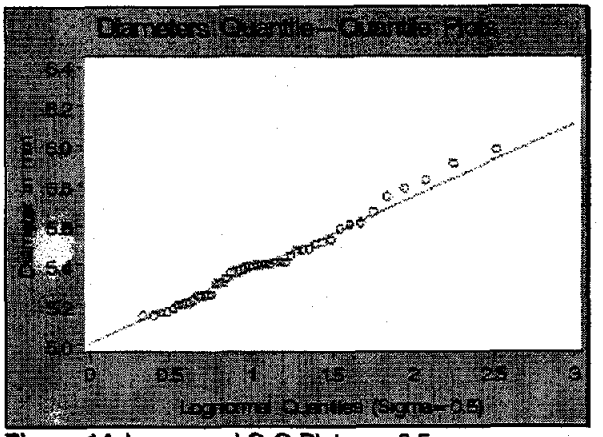


Figure 14: Lognormal Q-Q Plot, $\sigma = 0.5$

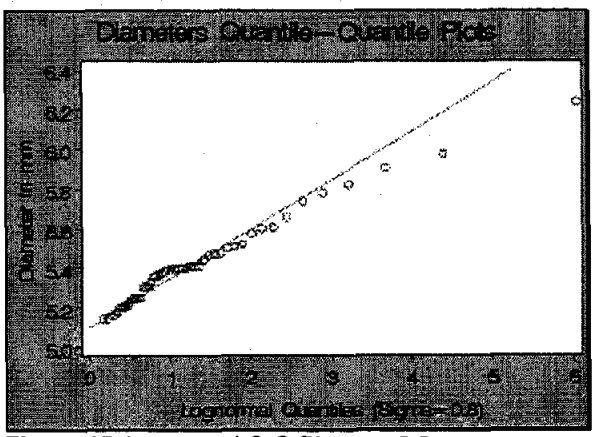


Figure 15: Lognormal Q-Q Plot, $\sigma = 0.8$

The multiple plots provide a convenient tool for visualizing changes in the fitted distribution across increasing values of σ. The Q-Q plot with $\sigma = 0.2$ is nonlinear with an increasing slope and does not fit the data well, while the Q-Q plot with $\sigma = 0.8$ is nonlinear with a decreasing slope and also does not fit the data well. In the Q-Q plot with $\sigma = 0.5$, the observations form an approximately linear pattern, although there are some discrepancies such as observations on the right tail of the distribution that are slightly above the reference line. This pattern indicates that the estimate of σ is close to but perhaps slightly different from 0.5.

Regardless, the close agreement between the observations and reference line indicate that a lognormal distribution with $\sigma = 0.5$ is appropriate. You can compute the sample estimates of θ and ζ by fitting a lognormal distribution using the HISTOGRAM statement with the LOGNORMAL option. This results in an estimated lognormal distribution with parameter estimates of $\sigma = 0.5$, $\theta \cong 5$, and $\zeta \cong -1$. You can also simultaneously estimate all three parameters, which results in similar estimates.

Comparing Distributions

Comparative histograms are useful for visualizing the distribution of a continuous variable simultaneously across levels of a classification variable. You can construct

comparative histograms by using the HISTOGRAM statement in conjunction with the CLASS statement.

Example: Monitoring Change in Blood Pressure

Consider a trial testing the efficacy of a new antihypertensive medication. The study measured the change in blood pressure after nine months for patients receiving either an innovative new medication or a placebo. The following statements create a data set that contains a treatment grouping variable and a variable that contains the change in blood pressure.

```
data BPChange;
   input Treatment $ BPchange;
   datalines;
Placebo  -14.0
Active    -8.0
Active   -23.0
Placebo    2.5
Active   -15.5
Placebo  -12.0
...
;
run;
```

The following statements construct a comparative histogram to compare the change in blood pressure across treatment groups.

```
proc univariate data=BPChange;
   class Treatment;
   histogram BPChange
      / normal
        midpoints  = -50 to 50 by 5
        cfill      = cx153e7e
        cframeside = cxeeeeee;
run;
```

The TREATMENT variable defines the classification levels for the analysis and is specified in the CLASS statement. The CFILL= option specifies the color of the histogram bars, and the CFRAMESIDE= option specifies the color that fills the classification label boxes. These two options are examples of the numerous options for controlling the appearance of graphics available in the UNIVARIATE procedure.

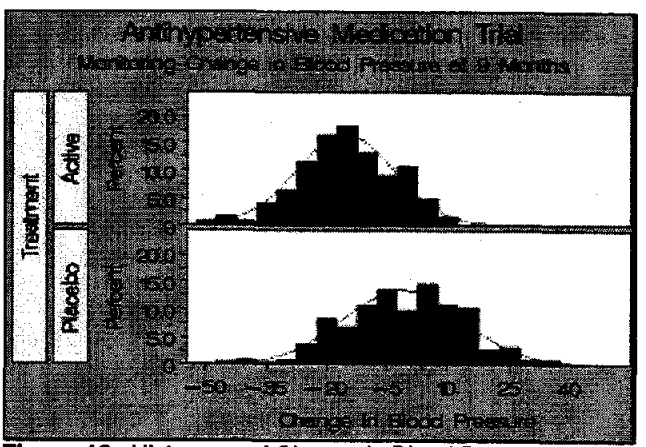


Figure 16. Histogram of Change in Blood Pressure Across Treatment Groups

Both groups appear to have normally distributed values for change in blood pressure. The distribution for the ACTIVE treatment is shifted to the left, indicating that the active treatment reduced blood pressure while the placebo treatment had no effect.

Graphical displays such as histograms and probability plots can be enhanced by an additional box or table (referred to as an *inset*) of summary statistics. A typical application of an inset is to augment a graphic with the sample size, mean, and standard deviation. The INSET statement creates an inset box and includes options for specifying the position of the box, applying a header for the table, and controlling background colors and font characteristics.

The following statements create a comparative histogram with an inset.

```
proc univariate data=BPChange noprint;
   class Treatment;
   var BPChange;
   histogram BPChange
      / normal
        midpoints = -50 to 50 by 5
        href = 0
        chref = cx888888
        vscale = count
        intertile = 1
        cframeside = cxeeeeee;
   inset n="N" (5.0)
         mean="Mean" (5.1)
         std="Std Dev" (5.1)
       / pos=ne height=3;
run;
```

The HREF= option in the HISTOGRAM statement creates a reference line on the x-axis at blood pressure equal to zero and the VSCALE=COUNT option specifies the scale of the vertical axis as the number of observations per bin. The INSET statement positions a table of summary statistics in the northeast corner of the display as defined by POS=NE. The table will include the number of observations, mean, and standard deviation, labeled "N," "Mean," and "Std Dev," respectively. The values in parentheses specify the format for the statistics in the inset, and the HEIGHT= option specifies the height of the text.

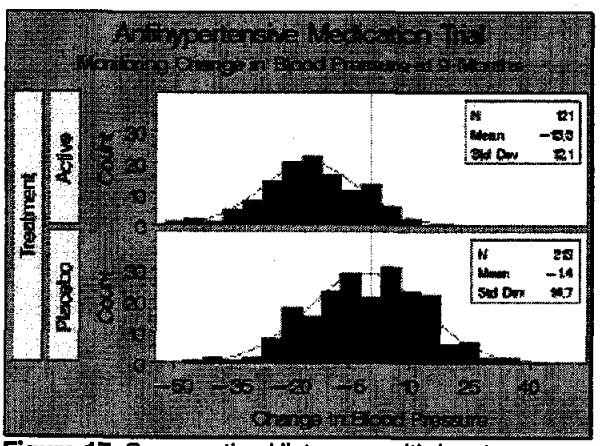


Figure 17. Comparative Histogram with Inset

The inset statistics provide a quick reference for comparing the means and standard deviations of the two treatment groups. On average, the blood pressure of patients receiving the active treatment decreased approximately 15 units, while the average change for the placebo group was nearly zero.

Although the mean difference of the two distributions is visible in the first comparative histogram, the left-shift of the treatment group distribution is graphically enhanced with the addition of the vertical reference line. Here, it is easier to see that the placebo group is nearly centered on zero while the treatment group is substantially shifted to the left. Because the vertical axis is scaled in frequency units, the size of the histogram bars reflects the different group sizes. This was not apparent in the previous histogram and is reinforced by the sample size statistics in the inset.

Box Plots

Box plots, also known as *box-and-whisker plots*, provide a convenient tool for comparing distributions of a quantitative variable across levels of a grouping variable. These plots display a wide range of quantitative information about a variable including its mean, median, quartiles, minimum, maximum, and outlying observations. The BOXPLOT procedure, available in Version 8 SAS/STAT software, produces box plots and enables you to specify different methods for calculating quantiles and control the layout and appearance of the plot.

The BOXPLOT procedure provides a variety of box plot styles. The skeletal box plot contains whiskers that extend from the edges of the box to the extreme values of the group. A skeletal style is requested by specifying BOXSTYLE=SKELETAL in the PLOT statement and is the default for the BOXPLOT procedure. On the other hand, schematic box plots contain whiskers drawn to the largest value within the upper fence and smallest value above the lower fence, and observations beyond the fence are displayed as individual symbols. The fence is one and a half times the interquartile range (IQR) from each edge of the box and is not displayed. You can request a schematic boxplot by specifying the BOXSTYLE=SCHEMATIC option.

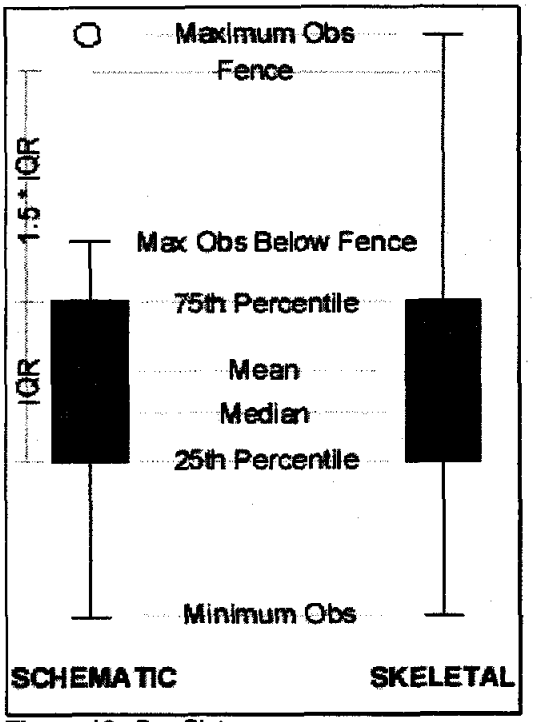


Figure 18. Box Plots

Figure 11 illustrates a few of the main features of schematic and skeletal box plots. You can also produce variations of the schematic box plot by using the SCHEMATICID and SCHEMATICIDFAR box plot styles. If you specify BOXSTYLE=SCHEMATICID, a schematic box plot is displayed in which the value of the first variable listed in the ID statement is used to label the symbol marking each observation outside the lower and upper fences. If you request the SCHEMATICIDFAR style, a schematic box plot is displayed where only observations outside the *far* fences are labeled. The lower and upper far fences are located 3×IQR below the 25th and 75th percentiles, respectively.

Example: Webpage Downloads

This example uses box plots to analyze the number of webpage downloads of the Statistics and Operations Research subsite of www.sas.com. Daily webpage downloads were recorded from late December 1998 through April 1999. The goal of the analysis is to compare the number of downloads across days of the week to monitor trends and diagnose heavy and light periods of web activity.

The following DATA step creates a SAS data set named WEBHITS that contains daily download counts with the corresponding date and day of the week.

```
data webhits;
   input Date $ Day $ Hits;
   datalines;
21Dec Monday    1003
22Dec Tuesday    851
23Dec Wednesday  757
24Dec Thursday   703
...
;
run;
```

The BOXPLOT procedure is used to create a box plot of web downloads across days of the week.

```
proc boxplot data=webhits;
   plot HITS*DAY
        / boxstyle = SKELETAL
          cboxes   = CX153e7e
          cboxfill = CX1589ff;
run;
```

The PLOT statement creates a boxplot of the HITS variable across the levels of the grouping variable DAY. The BOXSTYLE=SKELETAL option specifies the skeletal style, while the CBOXES= and CBOXFILL= options define the colors for the box outline and fill, respectively.

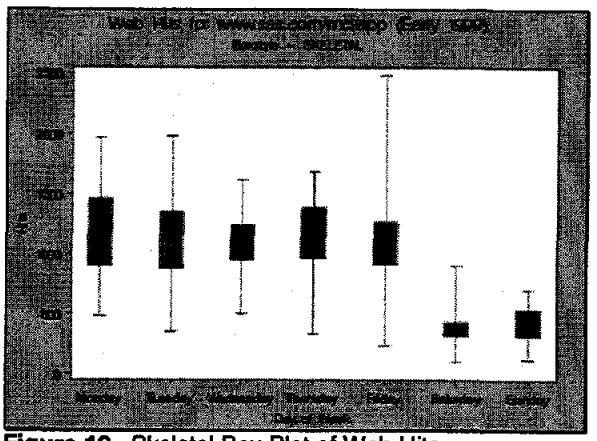


Figure 19. Skeletal Box Plot of Web Hits

Webpage download counts have similar distributions Monday through Friday, and there is a noticeable decrease in hits over the weekend, as expected. Note that even the maximum observations recorded on Saturday and Sunday, indicated by the serifs of the upper whiskers, are less than the median and mean of any weekday.

Although it is useful for visualizing the spread of a distribution, a skeletal box plot does not highlight outlying observations. You can highlight outliers by producing a schematic box plot. The following statements produce the desired plot.

```
proc boxplot data=webhits;
   plot HITS*DAY
        / caxis     = BLACK
          cframe    = CXFFFFFF
          ctext     = BLACK
          cboxes    = CX153E7E
          cboxfill  = CX1589FF
          boxstyle  = SCHEMATICID
          idcolor   = BLUE
          idsymbol  = CIRCLE;
   id DATE;
run;
```

The BOXSTYLE=SCHEMATICID option requests a schematic box plot with symbols marking each observation outside the lower and upper fences. These symbols are labeled using the values of the DATE variable specified in the ID statement. The IDCOLOR=BLUE and IDSYMBOL=CIRCLE options specify that the ID observations are to be displayed as blue circles.

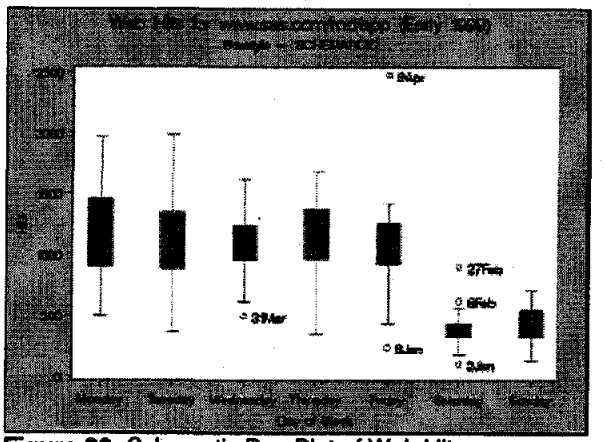


Figure 20: Schematic Box Plot of Web Hits

Although the display is generally similar in appearance to the skeletal box plot in Figure 20, the schematic box plot does provide a different perspective of the data. For example, the long upper whisker of the box plot for Friday was a result of a single extreme observation. Nearly 2500 downloads occurred on April 9th, which is at least 500 downloads more than any other day. Not coincidentally, April 9th was the last Friday before the SUGI 24 conference held in Miami. Perhaps many conference attendees visited the site on April 9th to read about the latest news and enhancements on their last day of work before coming to Florida.

Conclusion

The distributional modeling and visualization tools in SAS software have been significantly improved with enhancements to PROC UNIVARIATE and the new PROC BOXPLOT. Additional statistics, high-resolution graphics, and integration with the Output Delivery System provide users with additional tools for exploring data distributions. You can find more information about the new capabilities and other analytical SAS software on the R&D website at **www.sas.com/rnd/app/.**

Acknowledgements

I am grateful to Bob Rodriguez and Bucky Ransdell for their guidance and assistance and to Maura Stokes and Donna Sawyer for their editorial comments.

References

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

Rodriguez, R. N. and Taniguchi, B. Y. (1980), "A New Statistical Model for Predicting Customer Octane Satisfaction Using Trained-Rater Observations," SAE

Technical Paper 801356, *Transactions of the Society of Automotive Engineers*, 4213-4240.

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

SAS Institute Inc. (1979), *SAS User's Guide, 1979 Edition*, Cary, NC: SAS Institute Inc.

Shapiro, S. S. and Wilk, M. B. (1965), "An Analysis of Variance Test for Normality (complete samples)," *Biometrika*, 52, 591–611.

Snee, R. D. and Pfeifer, C. G. (1983), *Encyclopedia of Statistical Sciences*, 3:635, New York: John Wiley & Sons, Inc.

Terrell, G. R. and Scott, D. W. (1985), "Oversmoothed Nonparametric Density Estimates," *Journal of the American Statistical Association*, 69, 730–737.

Author

Nathan Curtis, SAS Institute Inc., SAS Campus Drive, Cary, NC, 27513. Phone: (919) 677–8000 ext. 5155. FAX: (919) 677-4444. Nathan.Curtis@sas.com.

SAS, Base SAS, SAS/QC and SAS/STAT are registered trademarks of SAS Institute Inc. in the USA and in other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.