Chapter 7

Descriptive Statistics

Chapter Table of Contents

Chapter 7
Descriptive Statistics

Introduction

Descriptive statistics and plots are often used in the initial phase of a statistical analysis. These tools enable you to identify relationships in the data and to determine directions for further analysis.

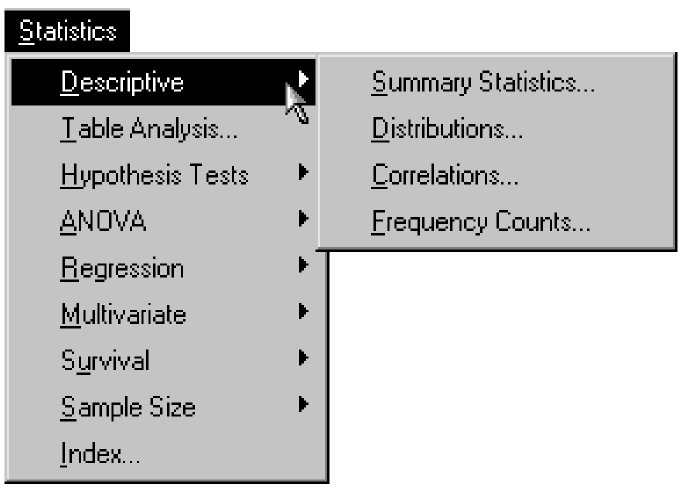


Figure 7.1. Descriptive Menu

The Analyst Application provides several types of descriptive statistics and graphical displays. The Summary Statistics task provides the following information: mean, median, standard error and standard deviation, variance, minimum, maximum, range, sum, skewness and kurtosis, student's *t* and probability value, coefficient of variation, and sums of squares. Graphics in this task include histograms and box-and-whisker plots.

The Distributions task produces statistics such as moments and quantiles as well as measures of location and variability. You can request fitted distributions from the normal, lognormal, Weibull, and exponential distributions. Plots included are the box-and-whisker plot, histogram, probability plot, and quantile-quantile plots. Histograms can be superimposed with fitted curves from the distribution families. Probability and quantile-quantile plots are available for each of the distributions.

The Correlations task gives you the choice of Pearson and Spearman correlations as well as Cronbach's alpha, Kendall's tau-*b*, and Hoeffding's D. Scatter plots with optional confidence ellipses are available.

The Frequency Counts task provides one-way frequency tables, which include frequencies, percentages, and cumulative frequencies and percentages. Horizontal and vertical bar charts are also available.

The examples in this chapter demonstrate how you can use the Analyst Application to compute one-way frequency tables, obtain summary statistics, examine the distribution of your data, and compute correlations.

Producing One-Way Frequencies

The data set analyzed in the following sections is taken from the 1995 Statistical Abstract of the United States. The data are measures of the birth rate and infant mortality rate for 1992 in the United States. Information is provided for the 50 states and the District of Columbia. The states are grouped by region. Here, these data are considered to be a sample of yearly data.

Suppose you want to determine the frequency of occurrence of the various regions. In the following example, a listing of the frequencies and a bar chart are produced.

In the Frequency Counts task, you can compute one-way frequency tables for the variables in your data set. For each value of your analysis variable, Analyst produces the frequency, cumulative frequency, and cumulative percentage. You can control the order in which the values appear and specify group and count variables.

Open the Bthdth92 Data Set

The data are provided in the Analyst Sample Library. To open the Bthdth92 data set, follow these steps:

1. Select **Tools** → **Sample Data** . . .
2. Select Bthdth92.
3. Click **OK** to create the sample data set in your Sasuser directory.
4. Select **File** → **Open By SAS Name** . . .
5. Select Sasuser from the list of **Libraries**.
6. Select Bthdth92 from the list of members.
7. Click **OK** to bring the Bthdth92 data set into the data table.

Request Frequency Counts

To request frequency counts, follow these steps:

1. Select **Statistics** → **Descriptive** → **Frequency Counts**. . .
2. Select region as the frequencies variable from the candidate list.

The default analysis provides the information desired. Note that you can use the Input dialog to select the specific ordering by which the variable values are listed.

Figure 7.2 displays the Frequency Counts dialog with region specified as the frequencies variable.

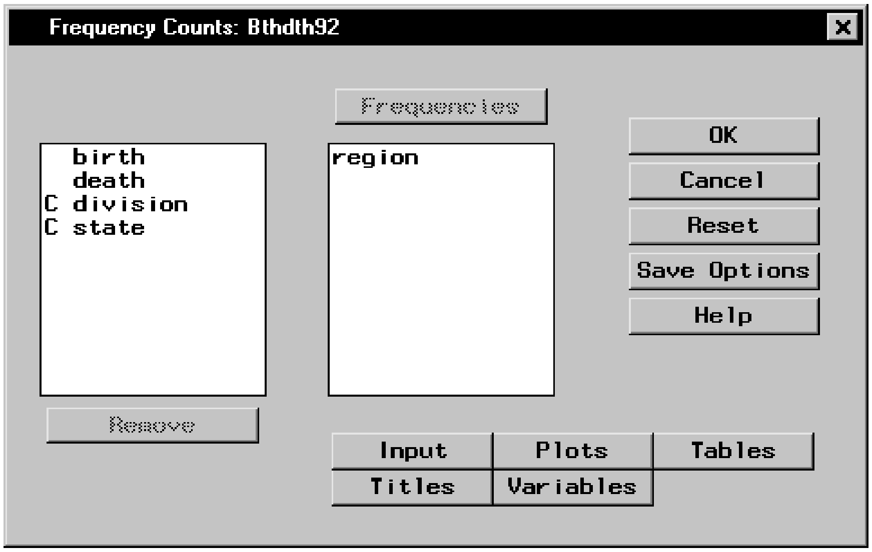


Figure 7.2. Frequency Counts Dialog

Request a Horizontal Bar Chart

To produce a horizontal bar chart in addition to the frequency counts, follow these steps:

1. Click on the **Plots** button.
2. Select **Horizontal**, as displayed in Figure 7.3.
3. Click **OK** to close the Plots dialog.

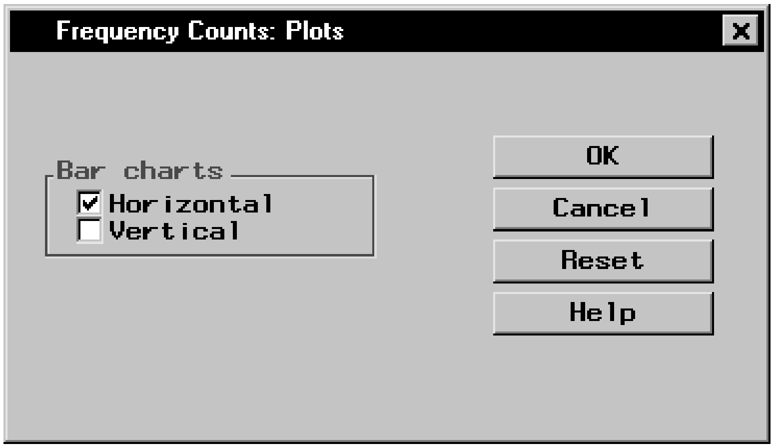


Figure 7.3. Frequency Counts: Plots Dialog

Click **OK** in the Frequency Counts main dialog to perform the analysis.

Review the Results

The results are presented in the project tree under the **Frequency Counts** folder, as displayed in Figure 7.4. The three nodes represent the frequency counts output, the horizontal bar chart, and the SAS programming statements (labeled **Code**) that generate the output.

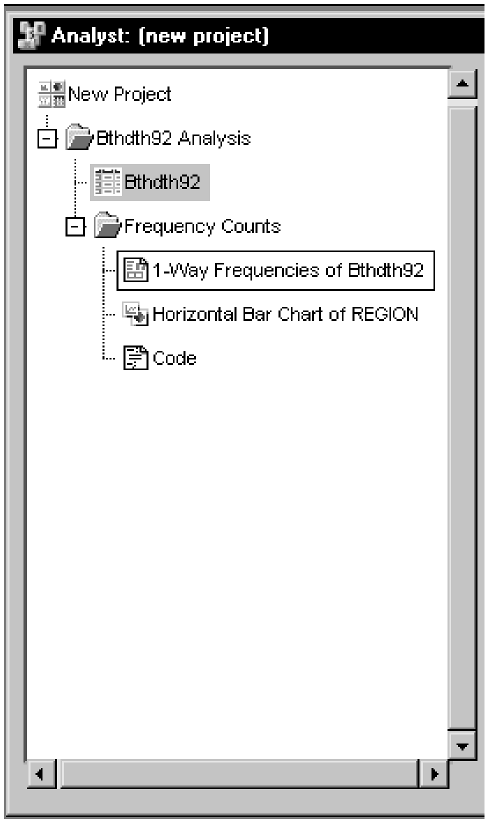
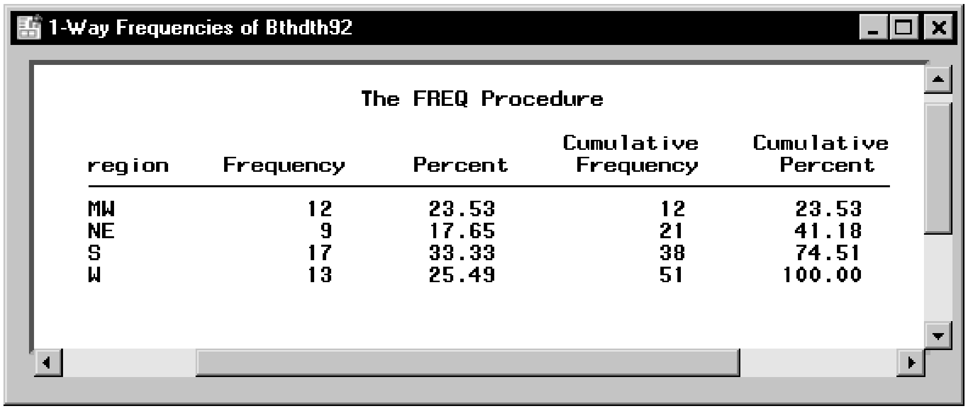


Figure 7.4. Frequency Counts: Project Tree

You can double-click on any node in the project tree to view the contents in a separate window. Note that the first output generated is displayed by default.

Figure 7.5 displays the table of frequency counts for the variable region.



1-Way Frequencies of Bthdth92

The FREQ Procedure

region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
MW	12	23.53	12	23.53
NE	9	17.65	21	41.18
S	17	33.33	38	74.51
W	13	25.49	51	100.00

Figure 7.5. Frequency Counts: One-Way Frequencies of the Variable region

The table shows that about 33% of the observations in the data set are located in the southern region, and roughly 25% of the observations are located in the western and midwestern regions, respectively. Approximately 18% of the observations are located in the northeastern region.

To display the bar chart of the frequency counts, double-click the node labeled **Horizontal Bar Chart of REGION** (Figure 7.6).

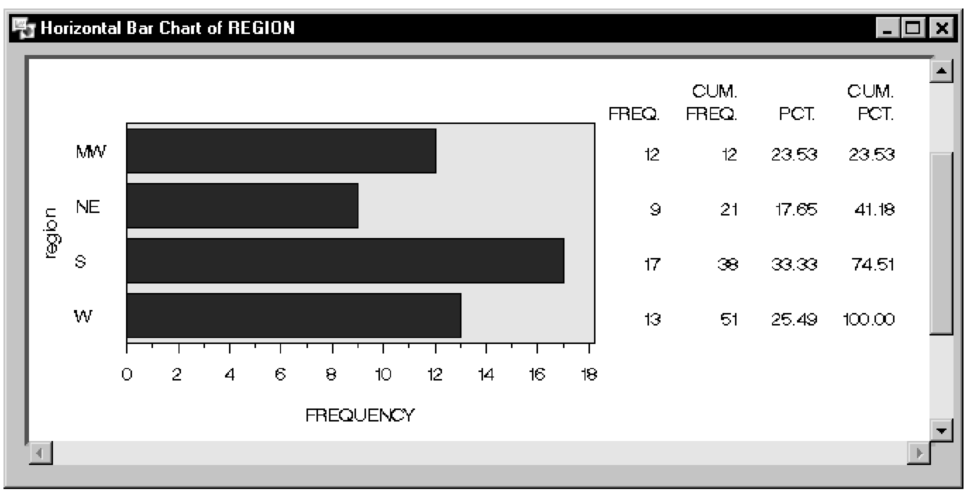


Figure 7.6. Frequency Counts: Horizontal Bar Chart by Region

Computing Summary Statistics

In this task, summary statistics (such as the mean, standard deviation, and minimum and maximum values) are desired for the birth and infant mortality rates for each region. In addition, box-and-whisker plots are requested.

Request Summary Statistics

To request the Summary Statistics task, follow these steps:

1. Select **Statistics** → **Descriptive** → **Summary Statistics**. . .
2. Select the analysis variables birth and death from the candidate list.

You can specify a classification variable to define groups within your data. When you specify a classification variable, the Analyst Application produces summary statistics for the analysis variables at each level of the classification variable.

3. Select region as the classification variable.

Figure 7.7 displays the Summary Statistics main dialog with birth and death specified as the analysis variables and region specified as the classification variable.

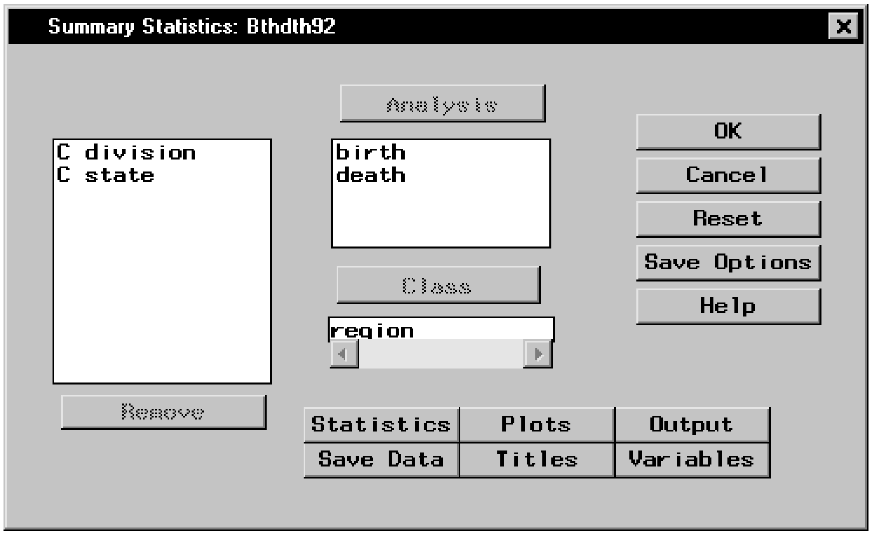


Figure 7.7. Summary Statistics Dialog

Request Box-and-Whisker Plots

To request box-and-whisker plots, follow these steps:

1. Click on the **Plots** button.
2. Select **Box-&-whisker plot**.
3. Click **OK**.

Figure 7.8 displays the Plots dialog with **Box-&-whisker plot** selected.

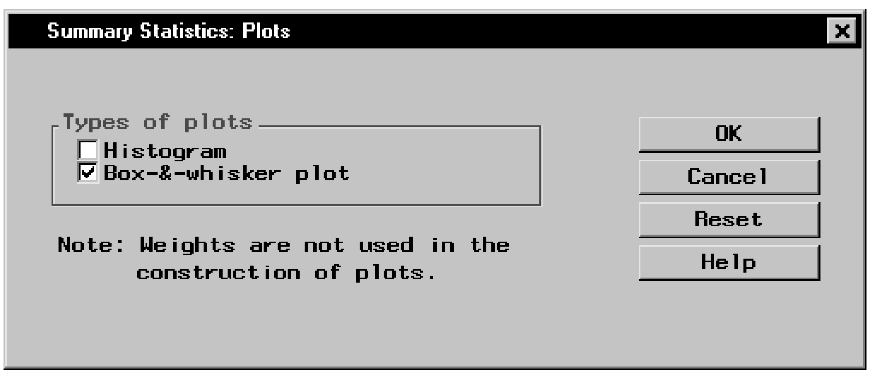


Figure 7.8. Summary Statistics: Plots Dialog

To perform the analysis, click **OK** in the main dialog.

Review the Results

The results are presented in the project tree under the **Summary Statistics** folder, as displayed in Figure 7.9. The four icons represent the summary statistics output, the box-and-whisker plots for each analysis variable, and the SAS programming statements (labeled **Code**) that generate the output.

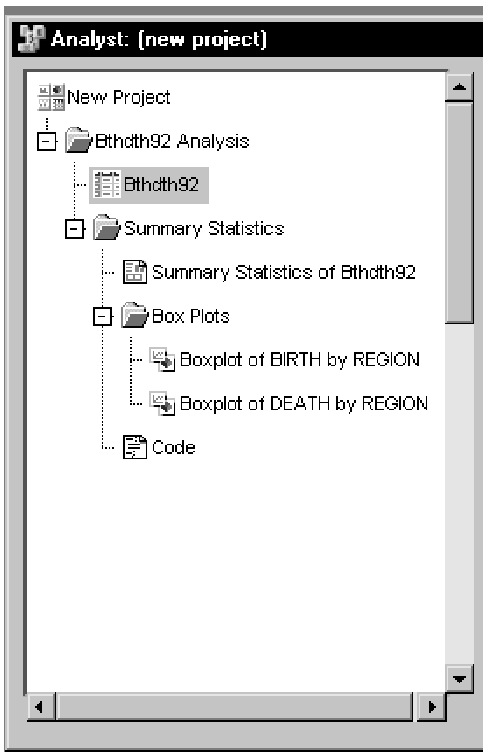


Figure 7.9. Summary Statistics: Project Tree

Double-click on any of the icons to display the corresponding information in a separate window.

Figure 7.10 displays, for each value of the classification variable region, the number of observations, the mean, the standard deviation, and the minimum and maximum values of each analysis variable.

The western region has the highest birth rate (16.89) and the southern region has the highest death rate (10.15).

Summary Statistics of Bthdth92

The MEANS Procedure

region	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
MW	12	birth	12	14.8250000	0.7581377	13.7000000	16.5000000
		death	12	8.5916667	1.0974833	7.1000000	10.2000000
NE	9	birth	9	14.3666667	0.8930286	13.0000000	15.9000000
		death	9	7.3777778	1.2194033	5.6000000	9.0000000
S	17	birth	17	15.4647059	1.4924565	12.3000000	18.7000000
		death	17	10.1529412	2.6241946	7.8000000	19.6000000
W	13	birth	13	16.8923077	2.1864970	14.0000000	20.5000000
		death	13	7.4769231	0.9670866	5.9000000	8.9000000

Figure 7.10. Summary Statistics: Statistics for birth and death

Figure 7.11 displays the box-and-whisker plot for the variable birth for each level of the region variable.

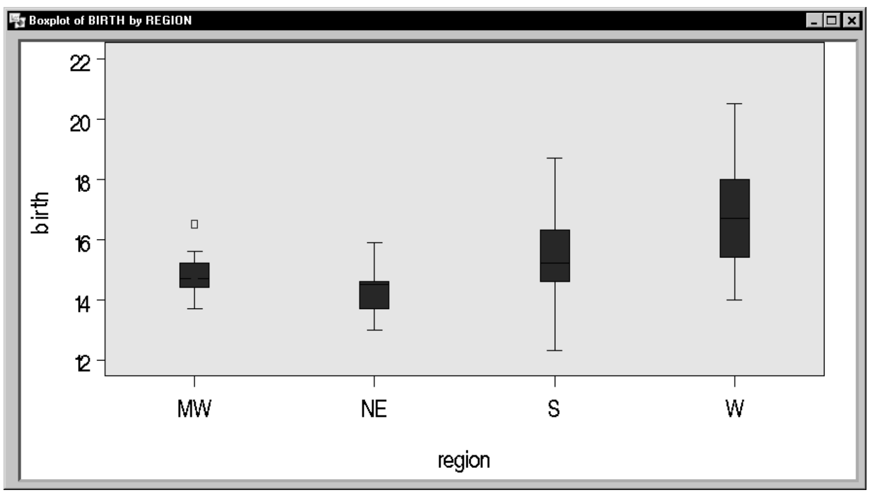


Figure 7.11. Summary Statistics: Box-and-Whisker Plot for Birth Rate by Region

This plot reveals a possible outlier in the birth rate for the midwestern region (region='MW'). The western region (region='W') is noticeable as the region with the highest birth rate.

Examining the Distribution

You can examine the distributional properties of your data with the Distributions task. This task enables you to produce descriptive statistics for the variables, test the fit of several distributions to your data, and examine displays such as histograms and probability plots. In this task, interest lies in examining the birth and infant mortality rates for each region.

Request a Distributions Analysis

To request the Distributions task, follow these steps:

1. Select **Statistics**→ **Descriptive** → **Distributions** . . .
2. Select birth and death as the analysis variables.
3. Select region as the classification variable.

Figure 7.12 displays the Distributions main dialog with the preceding variable specifications.

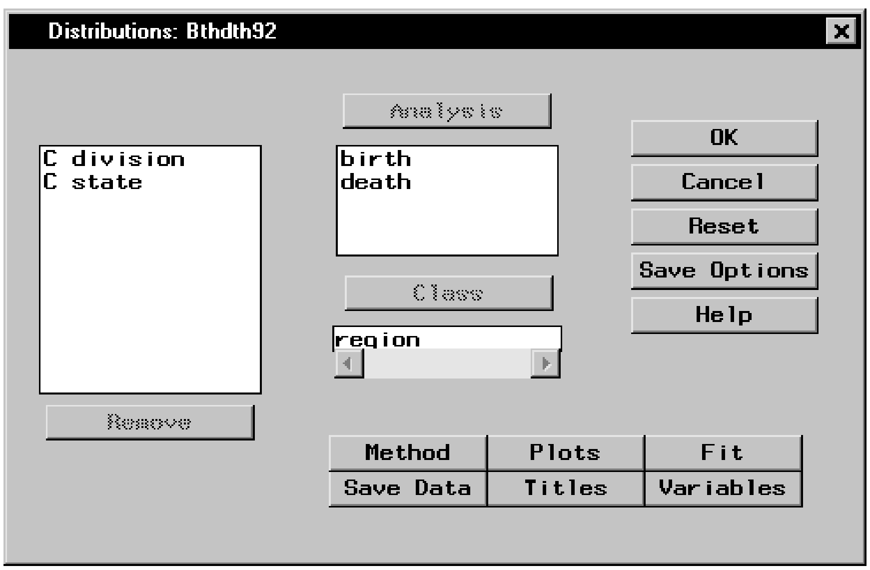


Figure 7.12. Distributions Dialog

The default analysis provides moments, quartiles, and measures of variability.

Request Plots

To request box-and-whisker plots and histograms, follow these steps:

1. Click on the **Plots** button.
2. Select **Box-&-whisker plot**.
3. Select **Histogram**.
4. Click **OK**.

Figure 7.13 displays the Plots dialog.

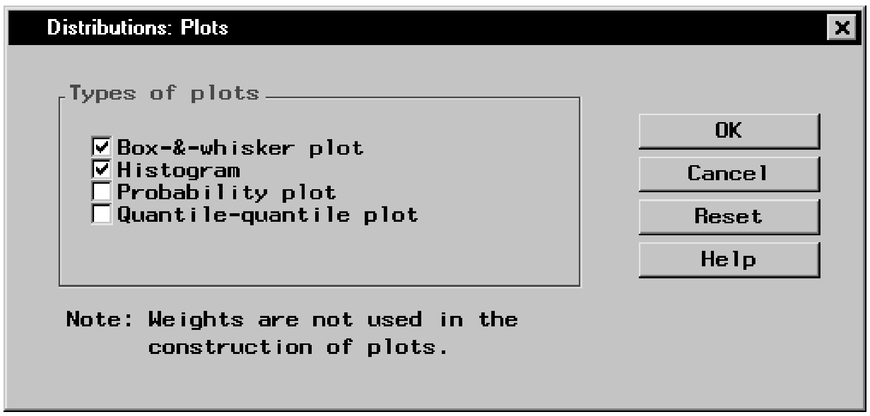


Figure 7.13. Distributions: Plots Dialog

Request Fitted Distribution

To fit a normal distribution to these data, follow these steps:

1. Click on the **Fit** button in the main dialog.
2. Select **Normal**.

By default, parameter values are calculated from the data when you fit the normal distribution. If you want to enter specific parameter values, click on the down arrow (displayed in Figure 7.14) and select **Enter values**. For the lognormal, exponential, and Weibull distributions, you can specify that parameters be calculated by maximum likelihood estimation (MLE), or you can enter specific parameter values.

3. Click **OK**.

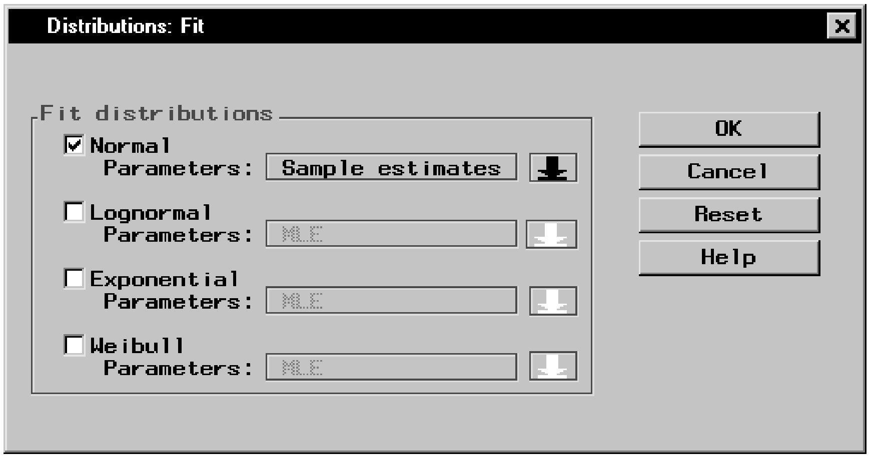


Figure 7.14. Distributions: Fit Dialog

When you have completed your selections, click **OK** in the main dialog to perform the analysis. The results are presented in the project tree displayed in Figure 7.15.

Review the Results

Double-click on any of the resulting eight icons to display the corresponding output in a separate window.

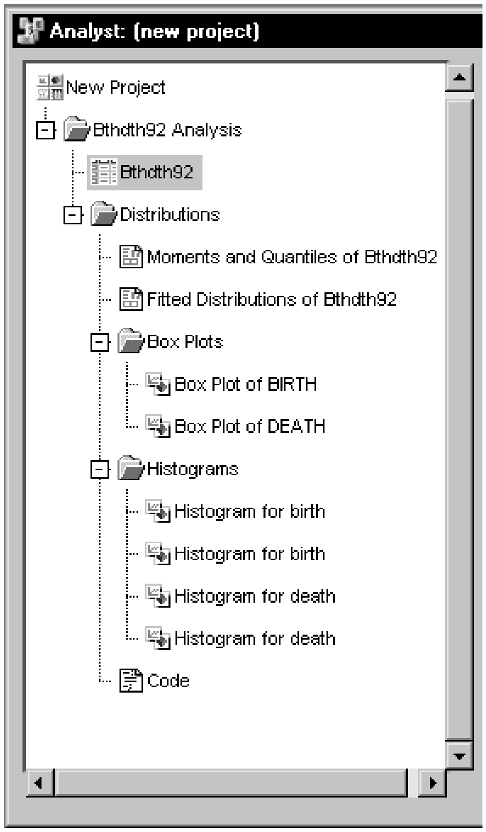
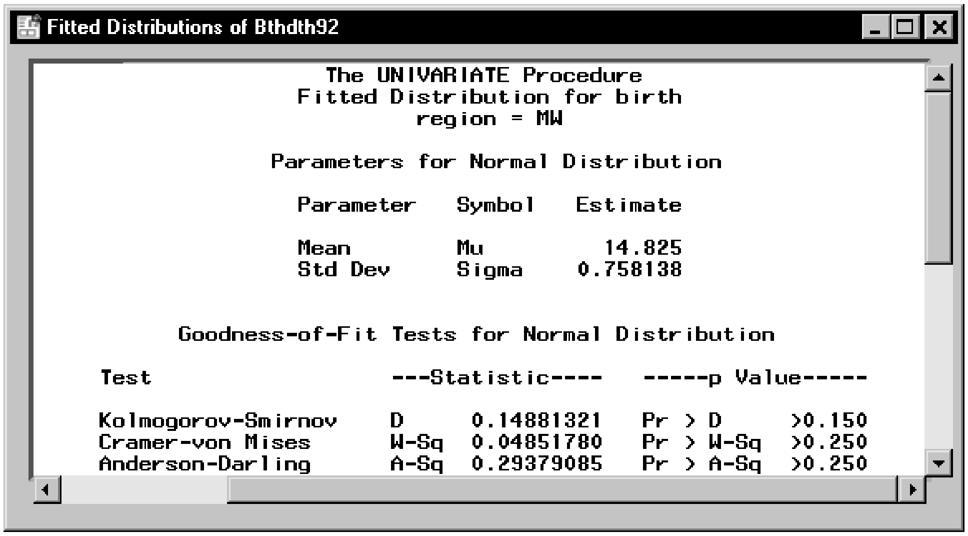


Figure 7.15. Distributions: Project Tree

The Moments and Quantiles output provides summary information for each variable. Figure 7.16 displays the output labeled Fitted Distributions of Bthdth92, which summarizes how closely the normal distribution fits each variable, by region.



Fitted Distributions of Bthdth92

```
                 The UNIVARIATE Procedure
               Fitted Distribution for birth
                       region = MW

              Parameters for Normal Distribution

               Parameter    Symbol    Estimate

               Mean         Mu          14.825
               Std Dev      Sigma     0.758138


          Goodness-of-Fit Tests for Normal Distribution

Test                    ---Statistic----    -----p Value-----

Kolmogorov-Smirnov      D      0.14881321   Pr > D      >0.150
Cramer-von Mises        W-Sq   0.04851780   Pr > W-Sq   >0.250
Anderson-Darling        A-Sq   0.29379085   Pr > A-Sq   >0.250
```

Figure 7.16. Distributions: Fitted Distributions Results

Based on the test results displayed in Figure 7.16, the null hypothesis that the variable birth is normally distributed cannot be rejected at the $\alpha = 0.05$ level of significance (*p*-values for all tests are greater than 0.15). The same is true for the variable death except for the southern region (region='S'). The hypothesis is rejected at the $\alpha = 0.05$ level of significance for the death rate in the southern region.

Two sets of box plots and four sets of histograms are also produced. A single box-and-whisker plot is created for each of the two variables. The box-and-whisker plot for the variable birth is displayed when you double-click **Box Plot of BIRTH** in the project tree.

Two histograms are created for each variable. Each graphic contains a histogram for two levels of the classification variable region. The first histogram contains the information for the midwestern and northeastern regions (region='MW' and region='NE'), as displayed in Figure 7.17. The second histogram (not shown) contains the information for the southern and western regions (region='S' and region='W').

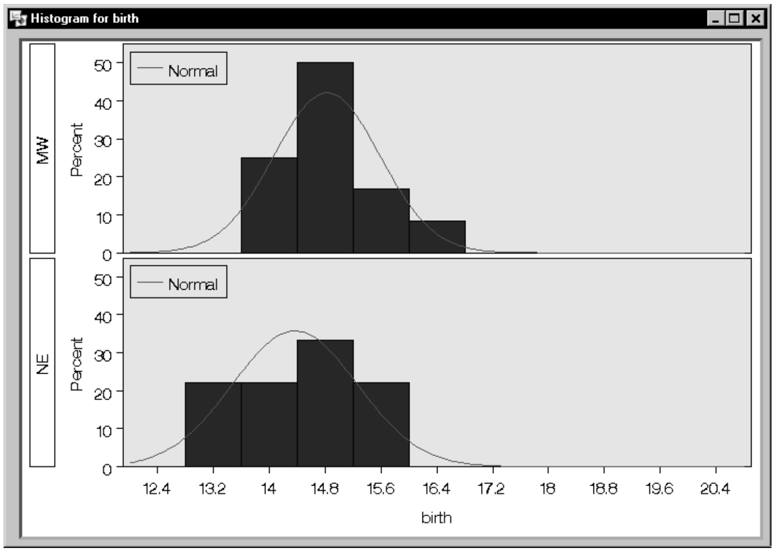


Figure 7.17. Distributions: Histogram for birth

The normal curve overlaid on the histogram displayed in Figure 7.17 is the result of requesting a normal distribution fit in the Fit dialog (Figure 7.14). The statistical details of the fit are located in the output labeled Fitted Distributions of Bthdth92, which also includes the details of the fit for the variable death.

Computing Correlations

You can use the Correlations task to compute pairwise correlation coefficients for the variables in your data set. The correlation is a measure of the strength of the linear relationship between two variables. This task can compute the standard Pearson product-moment correlations, nonparametric measures of association, partial correlations, and Cronbach's coefficient alpha. The task also can produce scatter plots with confidence ellipses.

The following example computes correlation coefficients for four variables in the Fitness data set. This data set contains measurements made on groups of men taking a physical fitness course at North Carolina State University. The variables are as follows:

age	age, in years
weight	weight, in kilograms
oxygen	oxygen intake rate, in milliliters per kilogram of body weight per minute
runtime	time taken to run 1.5 miles, in minutes
rstpulse	heart rate while resting
runpulse	heart rate while running
maxpulse	maximum heart rate recorded while running
group	group number

This example includes looking at correlations between the variables runtime, runpulse, maxpulse, and oxygen and also producing the corresponding scatter plots with confidence ellipses.

Open the Fitness Data Set

To open the Fitness data set, follow these steps:

1. Select **Tools** → **Sample Data** . . .
2. Select Fitness.
3. Click **OK** to create the sample data set in your Sasuser directory.
4. Select **File** → **Open By SAS Name** . . .
5. Select Sasuser from the list of **Libraries**.
6. Select Fitness from the list of members.
7. Click **OK** to bring the Fitness data set into the data table.

Request Correlations

To compute correlations for variables in the Fitness data set, follow these steps:

1. Select **Statistics** → **Descriptive** → **Correlations** . . .
2. Select the variables runtime, runpulse, maxpulse, and oxygen to correlate.

Figure 7.18 displays the resulting Correlations dialog.

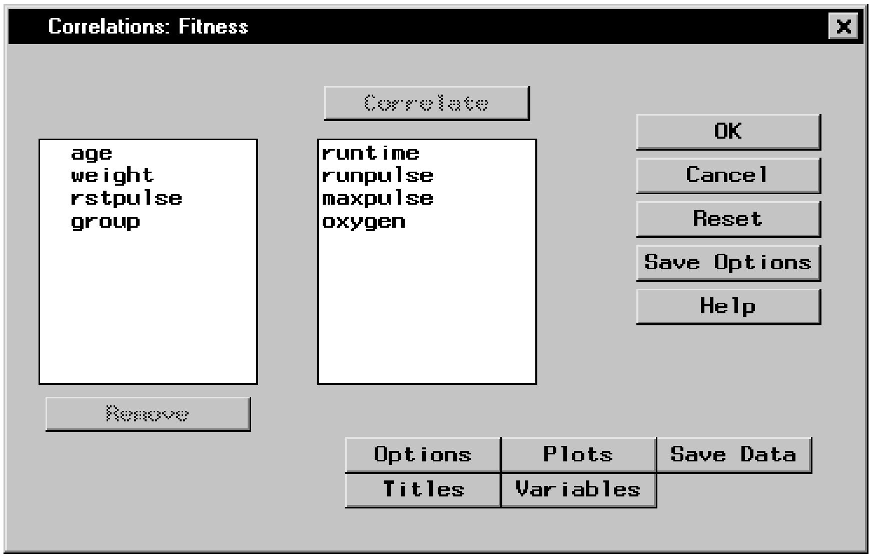


Figure 7.18. Correlations Dialog

If you click **OK** in the Correlations main dialog, the default output, which includes Pearson correlations, is produced. Or, you can request specific types of correlations by using the Options dialog.

Request a Scatter Plot

To request a scatter plot with a confidence ellipse, follow these steps:

1. Click on the **Plots** button.
2. Select **Scatter plots**.
3. Select **Add confidence ellipses**.

The confidence level used in calculating the confidence ellipse is 0.95. To use a different level, type that value in the **Probability value:** field, as displayed in Figure 7.19.

4. Click **OK**.

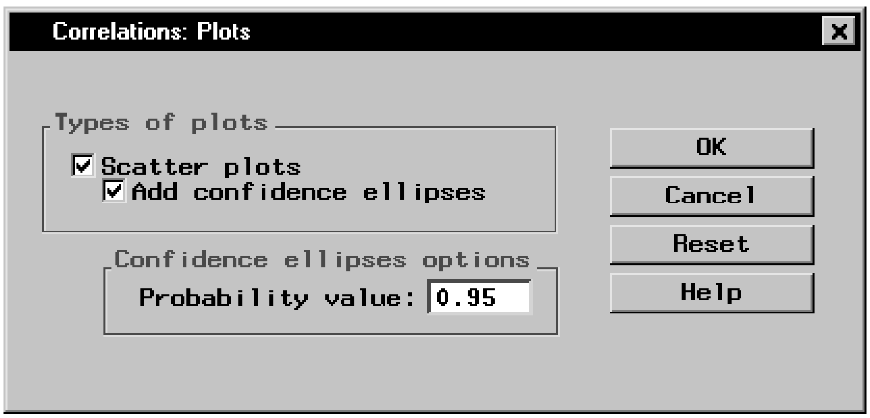


Figure 7.19. Correlations: Plots Dialog

Click **OK** in the main dialog to perform the analysis.

Review the Results

The results are presented in the project tree, as displayed in Figure 7.20.

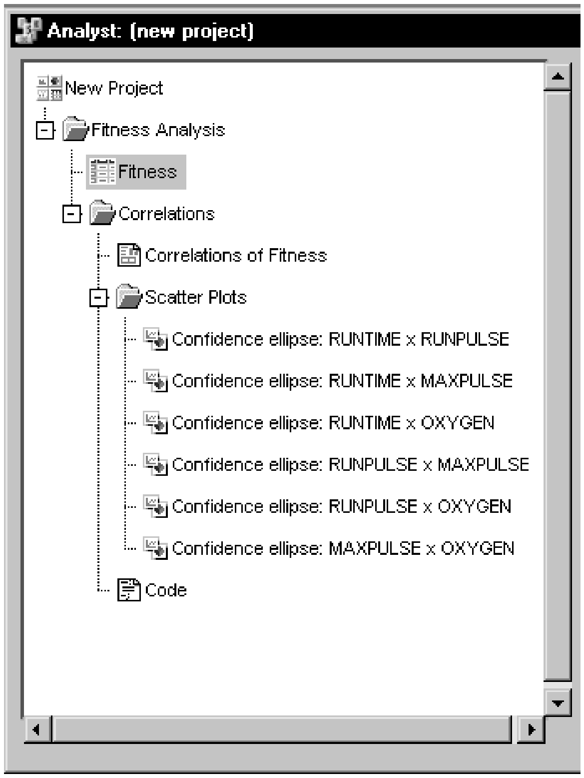


Figure 7.20. Correlations: Project Tree

You can double-click on any of the resulting nodes in the project tree to view the information in a separate window.

Figure 7.21 displays univariate statistics for each of the analysis variables. The table provides the number of observations, the mean, the standard deviation, the sum, and the minimum and maximum values for each variable.

Correlations of Fitness

The CORR Procedure

4 Variables: runtime runpulse maxpulse oxygen

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
runtime	31	10.58613	1.38741	328.17000	8.17000	14.03000
runpulse	31	169.64516	10.25199	5259	146.00000	186.00000
maxpulse	31	173.77419	9.16410	5387	155.00000	192.00000
oxygen	31	47.37581	5.32723	1469	37.38800	60.05500

Simple Statistics

Variable	Label
runtime	Min. to run 1.5 miles
runpulse	Heart rate while running
maxpulse	Maximum heart rate
oxygen	Oxygen consumption

Figure 7.21. Correlations: Univariate Statistics

Figure 7.22 displays the table of correlations. The *p*-value, which is the significance probability of the correlation, is displayed under each of the correlation coefficients. For example, the correlation between the variables maxpulse and runtime is 0.22610, with an associated *p*-value of 0.2213, and the correlation between the variables oxygen and runpulse is −0.39797, with an associated *p*-value of 0.0266.

Correlations of Fitness

Pearson Correlation Coefficients, N = 31
Prob > |r| under H0: Rho=0

	runtime	runpulse	maxpulse	oxygen
runtime Min. to run 1.5 miles	1.00000	0.31365 0.0858	0.22610 0.2213	-0.86219 <.0001
runpulse Heart rate while running	0.31365 0.0858	1.00000	0.92975 <.0001	-0.39797 0.0266
maxpulse Maximum heart rate	0.22610 0.2213	0.92975 <.0001	1.00000	-0.23674 0.1997
oxygen Oxygen consumption	-0.86219 <.0001	-0.39797 0.0266	-0.23674 0.1997	1.00000

Figure 7.22. Correlations: Table of Correlations

Six scatter plots, each of which includes a 95% confidence ellipse, are produced in this analysis. Each plot displays the relationship between one pair of the analysis variables. The scatter plot of runtime versus oxygen is displayed in Figure 7.23.

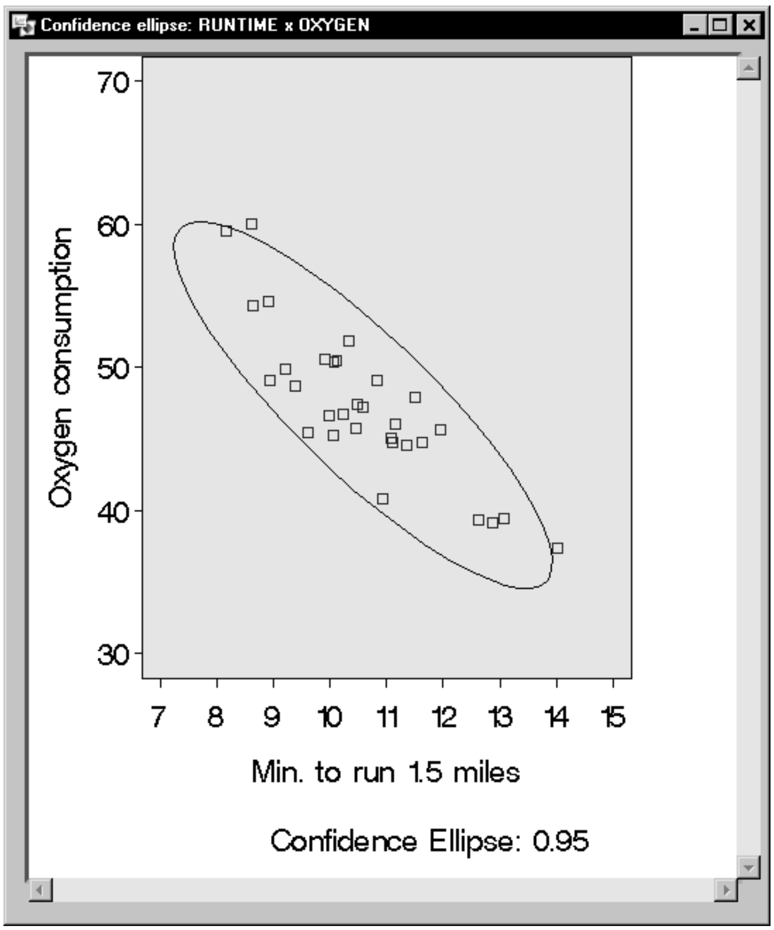


Figure 7.23. Correlations: Scatter Plot with Confidence Ellipse

Confidence ellipses are used as a graphical indicator of correlation. When two variables are uncorrelated, the confidence ellipse is circular in shape. The ellipse becomes more elongated the stronger the correlation is between two variables.

References

SAS Institute Inc. (1999), *SAS Procedures Guide, Version 7-1*, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1999), *SAS/STAT User's Guide, Version 7-1*, Cary, NC: SAS Institute Inc.

Schlotzhauer, Sandra D. and Littell, Ramon C. (1991), *SAS System for Elementary Statistical Analysis, Second Edition*, Cary, NC: SAS Institute Inc.

U.S. Bureau of the Census (1995), *Statistical Abstract of the United States*, Washington, D.C.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *The Analyst Application, First Edition*, Cary, NC: SAS Institute Inc., 1999. 476 pp.

The Analyst Application, First Edition

ISBN 1–58025–446–2

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

1st printing, October 1999