

Chapter 118

Data Screening

Introduction

This procedure performs a screening of data in a database, reporting on the:

1. Type of data (discrete or continuous)
2. Normality of each variable
3. Missing-value patterns
4. Presence of outliers

When you have missing values in your database, this program estimates the missing values using either a simple average or a more elaborate multiple regression technique.

Data screening should be carried out prior to any statistical procedure. Often data screening procedures are so tedious that they are skipped. Then, after an analysis produces unanticipated results, the data are scrutinized. This program automates the whole data screening process. When used in conjunction with histograms and scatter plots, you will be able to verify most of your data assumptions before beginning the actual analysis.

Data Structure

The data are entered in one or more variables. Only numeric values are allowed. Missing values are represented by blanks. Text values are treated as missing values.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables to be analyzed.

Data Variables

Variables to Screen

Specify the variables to be screened. Only numeric values are analyzed.

Options

Max Discrete Levels

The maximum number of unique values that a variable can have and still be designated as *discrete* rather than *continuous*.

Data Screening

Missing Value Estimation

Specify the type of missing value estimation (imputation) if any.

- **None**

No missing value estimation is carried out.

- **Average**

Estimate the missing value using the average of the variable.

Although this method is fast and simple, it does have disadvantages. First, the variance of the variable will be understated since adding values near the overall mean has little impact on the variance. Second, correlations with other variables may be incorrect since they involve the variances of the two variables.

- **Multivariate Normal**

Estimate missing values using the multivariate normal procedure. This method takes extra time but results in much more reliability estimates.

A regression analysis is conducted using the variable containing the missing value as the dependent variable and all variables with nonmissing data in this row as independent variables. The values of these nonmissing variables from the row containing the missing value are used in the regression equation to compute a predicted value for the missing value. Finally, if you are estimating a discrete value, the predicted value is rounded to the nearest possible discrete value. This process is iterated by using the imputed missing values from one run during the estimation phase of the next.

This procedure provides reasonable estimates of the missing values. It does have a few disadvantages. First, it assumes a multivariate normal distribution which may not be accurate. Second, it tends to provide estimates that understate the size of the variance of the variable. Third, it relies on the correlations between the variable with the missing value and the other variables in the database. If these correlations are all small, the resulting regression equation may not be very reliable.

Number of Iterations

This option specifies the number of iterations used during the estimation of missing values. Usually, only three or four iterations are necessary.

Zero

Specify the value used as zero by the numerical routines. Because of round-off problems, values less than this amount (in absolute value) are changed to zero during the calculations.

Treatment of Blanks at the End

This option specifies how to treat blanks at the end of each variable. Since the number of observations in each variable (column) can vary, an option is needed to control how the blanks at the end of each column are treated.

Let $MAXN$ represent the row number of the largest, non-blank row in the database. Let N_i represent the largest, non-blank row number in column i . Two options are available:

- **Spreadsheet**

With this option, blanks at the end (bottom) of each variable are ignored. That is, the N_i of each variable is determined separately.

- **Database**

With this option, blanks at the end (bottom) of each variable are included. That is, each N_i is set equal to $MAXN$.

Data Screening

Store Estimated Values

Checking this option will cause missing values in the database to be replaced by their estimates. Remember, these new values are not stored permanently until you manually save the database.

Reports Tab

The following options control the format of the reports that are displayed.

Select Reports

Descriptive Statistics ... Iteration Report

Indicate whether to display the indicated reports.

Report Options

T2 Alpha

This is the probability value used to identify outliers. Observations with a T2 probability less than this are designated as outliers.

Normality Test Alpha

This is the probability value used to identify variables that are not normally distributed. Variables with a normal test probability less than this are designated as being not normal.

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy.

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Storage Tab

Specify columns to store result values.

Select Variables for Data Storage

T2 Value and T2 Prob Level

Designate the name of the variables on the spreadsheet to contain the stored values.

Example 1 – Screening Data

This section presents an example of how to screen the data in the PCA2 dataset.

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the Data Screening window.

1 Open the PCA2 dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **PCA2.NCSS**.
- Click **Open**.

2 Open the Data Screening window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Data Screening** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Data Screening window, select the **Variables tab**.
- Double-click in the **Variables to Screen** box. This will bring up the variable selection window.
- Select **X1** to **Normal** from the list of variables and then click **Ok**.
- Enter **Multivariate Normal** in the **Missing Value Estimation** box.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Descriptive Statistics Section

Descriptive Statistics Section							
Data Type	Variable	Value Count	Missing Count	Minimum	Maximum	Mean	Standard Deviation
Continuous	X1	30	0	3	85	44.2	24.66241
Continuous	X2	30	0	1	102	51.53333	30.57803
Continuous	X3	30	0	2	105	54.93333	29.05753
Continuous	X4	30	0	5	91	41.7	25.3175
Continuous	X5	30	0	6	91	43.66667	26.65143
Continuous	X6	30	0	0	102	47.63334	34.18962
Discrete	Q1	30	0	1	5	3.466667	1.407696
Discrete	Q2	30	0	1	5	3.033333	1.098065
Discrete	Q3	29	1	1	5	2.827586	1.559967
Discrete	Q4	30	0	1	5	2.833333	1.620628
Discrete	Q5	27	3	1	5	2.62963	1.445102
Discrete	Q6	29	1	1	5	3.241379	1.50369
Discrete	Q7	30	0	1	5	2.966667	1.188547
Discrete	Q8	28	2	1	5	2.785714	1.397276
Discrete	Q9	30	0	1	5	2.9	1.470398
Continuous	Normal	30	0	79.98992	119.8588	100.6198	10.17271

This report gives descriptive statistics and counts for each variable. Note that using the missing value imputation option will not influence the values in this report. Most of these statistics have been defined in the Descriptive Statistics chapter.

Data Type

The type of data contained in each variable. If the number of unique values is less than the cutoff value given in the Max Discrete Levels option, the variable will be categorized as *Discrete*. Otherwise, it is categorized as *Continuous*. It is important to know the data type of each variable early in an analysis.

Data Screening

Value Count

This is the number of rows for which there were valid numeric values.

Missing Count

This is the number of rows for which there were missing values.

Normality Tests Section

Normality Tests Section									
Variable	----- Skewness Test -----			----- Kurtosis Test -----			- Omnibus Test -		Variable Normal?
	Value	Z	Prob	Value	Z	Prob	K2	Prob	
X1	0.11	0.30	0.7662	1.93	-1.80	0.0715	3.34	0.1885	Yes
X2	0.11	0.30	0.7664	1.88	-2.01	0.0446	4.12	0.1274	No
X3	-0.02	-0.06	0.9552	2.14	-1.14	0.2534	1.31	0.5201	Yes
X4	0.41	1.05	0.2918	2.31	-0.72	0.4712	1.63	0.4425	Yes
X5	0.39	0.99	0.3221	2.02	-1.50	0.1327	3.24	0.1978	Yes
X6	0.29	0.75	0.4535	1.63	-3.18	0.0015	10.65	0.0049	No
Q1	-0.50	-1.26	0.2072	1.97	-1.65	0.0980	4.33	0.1148	Yes
Q2	0.41	1.05	0.2925	2.27	-0.79	0.4269	1.74	0.4191	Yes
Q3	0.12	0.31	0.7595	1.53	-3.67	0.0002	13.57	0.0011	No
Q4	0.17	0.46	0.6483	1.52	-3.85	0.0001	15.03	0.0005	No
Q5	0.36	0.89	0.3732	1.81	-2.09	0.0368	5.15	0.0761	No
Q6	-0.23	-0.58	0.5597	1.59	-3.33	0.0009	11.41	0.0033	No
Q7	-0.31	-0.80	0.4208	2.19	-1.00	0.3186	1.64	0.4398	Yes
Q8	0.22	0.57	0.5713	1.75	-2.43	0.0153	6.20	0.0450	No
Q9	0.11	0.28	0.7765	1.67	-2.93	0.0034	8.64	0.0133	No
Normal	-0.28	-0.72	0.4715	2.44	-0.43	0.6701	0.70	0.7047	Yes

This report shows the results of the three D'Agostino normality tests. These tests are described in detail in the Descriptive Statistics chapter. If any of the three probability values are less than the user supplied Normality Test Alpha, the variable is designated as *not normal* (Variable Normal = No). Otherwise, the variable is designated as *normal* (Variable Normal = Yes).

We should remind you that the results of these tests depends heavily on sample size. If you have a small sample size (less than 50), these tests may fail to reject normality because the sample size is too small—not because the data are actually normal. Likewise, if your sample size is very large (greater than 1000), these tests may reject normality even though the data are nearly normal. When in doubt, you should supplement these tests with additional tests and graphs.

Pair-wise Missing Data Counts Section

Pair-wise Missing Data Counts Section						
	X1	X2	X3	X4	X5	X6
X1	0	0	0	0	0	0
X2	0	0	0	0	0	0
X3	0	0	0	0	0	0
X4	0	0	0	0	0	0
X5	0	0	0	0	0	0
X6	0	0	0	0	0	0
Q1	0	0	0	0	0	0
Q2	0	0	0	0	0	0
Q3	1	1	1	1	1	1
Q4	0	0	0	0	0	0
Q5	3	3	3	3	3	3
Q6	1	1	1	1	1	1
Q7	0	0	0	0	0	0
Q8	2	2	2	2	2	2
Q9	0	0	0	0	0	0
Normal	0	0	0	0	0	0

(report continues)

Data Screening

This report provides a pair-wise break down of the number of rows with missing values in at least one of each pair of variables. This is the number of observations that would be omitted from the calculation of the correlation coefficient between these two variables.

An understanding of the distribution of missing values is extremely important when conducting an analysis that is based on correlations such as factor analysis or multiple regression. You may determine that much more data would be used if you omit two or three variables that have high counts of missing values.

Pair-wise Missing Data Percentages Section

Pair-wise Missing Data Percentages Section						
	X1	X2	X3	X4	X5	X6
X1	0.0	0.0	0.0	0.0	0.0	0.0
X2	0.0	0.0	0.0	0.0	0.0	0.0
X3	0.0	0.0	0.0	0.0	0.0	0.0
X4	0.0	0.0	0.0	0.0	0.0	0.0
X5	0.0	0.0	0.0	0.0	0.0	0.0
X6	0.0	0.0	0.0	0.0	0.0	0.0
Q1	0.0	0.0	0.0	0.0	0.0	0.0
Q2	0.0	0.0	0.0	0.0	0.0	0.0
Q3	3.3	3.3	3.3	3.3	3.3	3.3
Q4	0.0	0.0	0.0	0.0	0.0	0.0
Q5	10.0	10.0	10.0	10.0	10.0	10.0
Q6	3.3	3.3	3.3	3.3	3.3	3.3
Q7	0.0	0.0	0.0	0.0	0.0	0.0
Q8	6.7	6.7	6.7	6.7	6.7	6.7
Q9	0.0	0.0	0.0	0.0	0.0	0.0
Normal	0.0	0.0	0.0	0.0	0.0	0.0

(report continues)

This report provides a pair-wise break down of the percentage of rows with missing values in at least one of each pair of variables. This is the percentage of observations that would be omitted from the calculation of the correlation coefficient between these two variables.

An understanding of the distribution of missing values is extremely important when conducting an analysis that is based on correlations such as factor analysis or multiple regression. You may determine that much more data would be used if you omit two or three variables that have high counts of missing values.

List of Discrete Variables and Values Section

List of Discrete Variables and Values Section	
Variable	Value1(Count1) Value2(Count2) etc.
Q1	1(4) 2(4) 3(5) 4(8) 5(9)
Q2	1(1) 2(10) 3(10) 4(5) 5(4)
Q3	1(9) 2(4) 3(5) 4(5) 5(6)
Q4	1(10) 2(3) 3(7) 4(2) 5(8)
Q5	1(8) 2(6) 3(5) 4(4) 5(4)
Q6	1(5) 2(6) 3(3) 4(7) 5(8)
Q7	1(5) 2(4) 3(10) 4(9) 5(2)
Q8	1(6) 2(8) 3(4) 4(6) 5(4)
Q9	1(7) 2(6) 3(6) 4(5) 5(6)

This report lists each of the discrete variables (as defined by Max Discrete Levels) followed by a list of the discrete values and corresponding counts of those values. For example, the first entry of 1(4) means that four 1's occurred in this variable.

This report is particular useful in helping you find out-of-range values in discrete data.

Multivariate Outlier Section

Multivariate Outlier Section			
Row	T2 Value	T2 Prob	Outlier?
1			
2	27.97	0.6310	
3	28.03	0.6295	
4	11.99	0.9729	
5			
6	11.30	0.9790	
7	20.40	0.8250	
8			
9	11.86	0.9741	
10	13.64	0.9544	
.	.	.	
.	.	.	
.	.	.	

This report tests each observation to determine if it is a multivariate outlier. The program uses a T^2 test based on the Mahalanobis distance of each point from the variable means. The formula for T^2 is:

$$T_i^2 = (n-1)(X_i - \bar{X})' \left[(X - \bar{X})' (X - \bar{X}) \right]^{-1} (X_i - \bar{X})$$

The following mathematical relationship between the T^2 and the F-distribution is used to calculate the probability levels:

$$T_{p,n,\alpha}^2 = \frac{p(n-1)}{n-p} F_{p,n-p,\alpha}$$

Note that as the number of variables, p , approaches the sample size, n , the denominator degrees of freedom approaches zero. As $n-p$ approaches zero, the power of the test also approaches zero.

This test is only calculated for rows that have no missing values. To test rows with missing values, you will need to store imputed values on the database and rerun the analysis.

When the probability level is less than the value indicated in the T2 Alpha box, the observation is starred.

Rows With Missing Values Section

Rows With Missing Values Section	
Row	Pattern of Missing Values (= data, . = missing)
1	.
5	..
8	. .
12	. .
14	.
16	.
19	.

Variables With Missing Values
 Q3
 Q5
 Q6
 Q8

This report presents a list of only those variables and rows that had missing values. It lets you consider the pattern of missing values more closely.

For each row, missing values are represented by a period and valid values are represented by a vertical bar. These symbols were selected because they have about the same width in most fonts.

Missing Values Estimation Iteration Section

Missing Values Estimation Iteration Section

Iteration No.	Count	Covariance Matrix Trace	Percent Change
0	23	4815.0995	0.00
1	30	5029.5552	4.45
2	30	5029.4391	0.00
3	30	5029.3897	0.00

This report shows the percent change in the trace of the variance-covariance matrix as you progress from one iteration to the next during the estimation of missing values. You would use the report to determine if enough iterations have been run during the estimation of missing values. Once the percent change is less than four percent after the first two iterations, you could terminate the procedure. If the last two iterations show very different values, you should rerun the analysis with a higher number of iterations.