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.

Example 1 – Screening Data

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

Setup

To run this example, complete the following steps:

1 Open the PCA2 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **PCA2** and click **OK**.

2 Specify the Data Screening procedure options

- Find and open the **Data Screening** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab

Variables to Screen **X1-Normal**
Missing Value Estimation.................................**Multivariate Normal**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Descriptive Statistics

Descriptive Statistics

		Number of Values					Standard
Data Type	**Variable**	**Data**	**Missing**	**Minimum**	**Maximum**	**Mean**	**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.

Number of Values: Data

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

Number of Values: Missing

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

Normality Tests

Normality Tests

	Skewness Test			Kurtosis Test			Omnibus Test		Variable Normal
Variable	**Value**	**Z**	**P-Value**	**Value**	**Z**	**P-Value**	**K²**	**P-Value**	**at α = 0.05?**
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 depend 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. If in doubt, you should supplement these tests with additional tests and graphs.

Pairwise Missing Data Counts

```
Pairwise Missing Data Counts
_____________________________________________________________________________

Section 1
          X1    X2    X3    X4    X5    X6    Q1    Q2    Q3    Q4    Q5    Q6

X1         0     0     0     0     0     0     0     0     1     0     3     1
X2         0     0     0     0     0     0     0     0     1     0     3     1
X3         0     0     0     0     0     0     0     0     1     0     3     1
X4         0     0     0     0     0     0     0     0     1     0     3     1
X5         0     0     0     0     0     0     0     0     1     0     3     1
X6         0     0     0     0     0     0     0     0     1     0     3     1
Q1         0     0     0     0     0     0     0     0     1     0     3     1
Q2         0     0     0     0     0     0     0     0     1     0     3     1
Q3         1     1     1     1     1     1     1     1     1     1     4     2
Q4         0     0     0     0     0     0     0     0     1     0     3     1
Q5         3     3     3     3     3     3     3     3     4     3     3     4
Q6         1     1     1     1     1     1     1     1     2     1     4     1
Q7         0     0     0     0     0     0     0     0     1     0     3     1
Q8         2     2     2     2     2     2     2     2     3     2     5     3
Q9         0     0     0     0     0     0     0     0     1     0     3     1
Normal     0     0     0     0     0     0     0     0     1     0     3     1

Section 2
          Q7    Q8    Q9    Normal

X1         0     2     0         0
X2         0     2     0         0
X3         0     2     0         0
X4         0     2     0         0
X5         0     2     0         0
X6         0     2     0         0
Q1         0     2     0         0
Q2         0     2     0         0
Q3         1     3     1         1
Q4         0     2     0         0
Q5         3     5     3         3
Q6         1     3     1         1
Q7         0     2     0         0
Q8         2     2     2         2
Q9         0     2     0         0
Normal     0     2     0         0
_____________________________________________________________________________
```

This report provides a pairwise breakdown 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.

Pairwise Missing Data Percentages

Pairwise Missing Data Percentages

Section 1

	X1	X2	X3	X4	X5	X6	Q1	Q2
X1	0.0%	0.0%	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%	0.0%	0.0%
X3	0.0%	0.0%	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%	0.0%	0.0%
X5	0.0%	0.0%	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%	0.0%	0.0%
Q1	0.0%	0.0%	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%	0.0%	0.0%
Q3	3.3%	3.3%	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%	0.0%	0.0%
Q5	10.0%	10.0%	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%	3.3%	3.3%
Q7	0.0%	0.0%	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%	6.7%	6.7%
Q9	0.0%	0.0%	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%	0.0%	0.0%

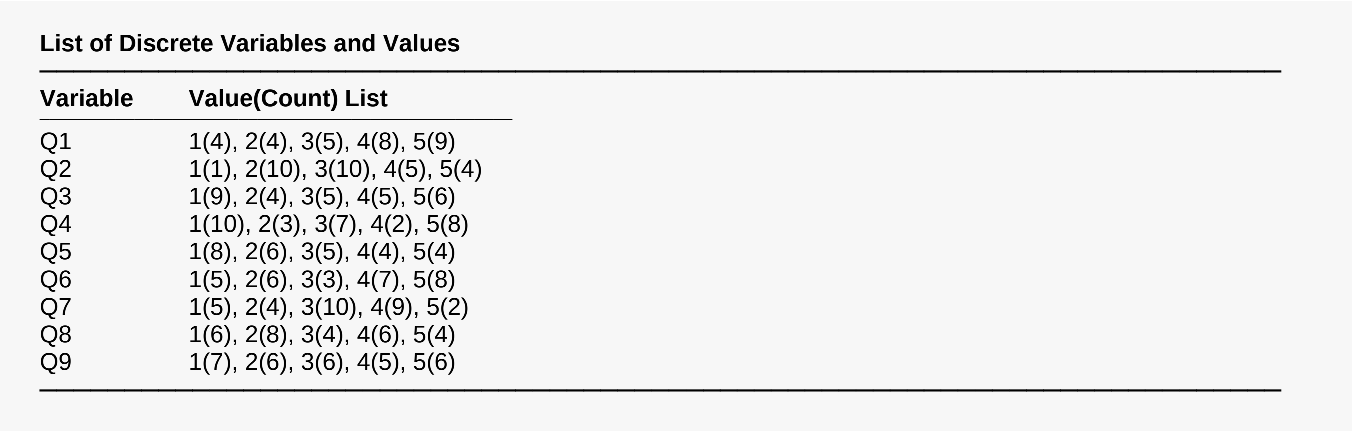
Section 2

	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Normal
X1	3.3%	0.0%	10.0%	3.3%	0.0%	6.7%	0.0%	0.0%
X2	3.3%	0.0%	10.0%	3.3%	0.0%	6.7%	0.0%	0.0%
X3	3.3%	0.0%	10.0%	3.3%	0.0%	6.7%	0.0%	0.0%
X4	3.3%	0.0%	10.0%	3.3%	0.0%	6.7%	0.0%	0.0%
X5	3.3%	0.0%	10.0%	3.3%	0.0%	6.7%	0.0%	0.0%
X6	3.3%	0.0%	10.0%	3.3%	0.0%	6.7%	0.0%	0.0%
Q1	3.3%	0.0%	10.0%	3.3%	0.0%	6.7%	0.0%	0.0%
Q2	3.3%	0.0%	10.0%	3.3%	0.0%	6.7%	0.0%	0.0%
Q3	3.3%	3.3%	13.3%	6.7%	3.3%	10.0%	3.3%	3.3%
Q4	3.3%	0.0%	10.0%	3.3%	0.0%	6.7%	0.0%	0.0%
Q5	13.3%	10.0%	10.0%	13.3%	10.0%	16.7%	10.0%	10.0%
Q6	6.7%	3.3%	13.3%	3.3%	3.3%	10.0%	3.3%	3.3%
Q7	3.3%	0.0%	10.0%	3.3%	0.0%	6.7%	0.0%	0.0%
Q8	10.0%	6.7%	16.7%	10.0%	6.7%	6.7%	6.7%	6.7%
Q9	3.3%	0.0%	10.0%	3.3%	0.0%	6.7%	0.0%	0.0%
Normal	3.3%	0.0%	10.0%	3.3%	0.0%	6.7%	0.0%	0.0%

This report provides a pairwise breakdown 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

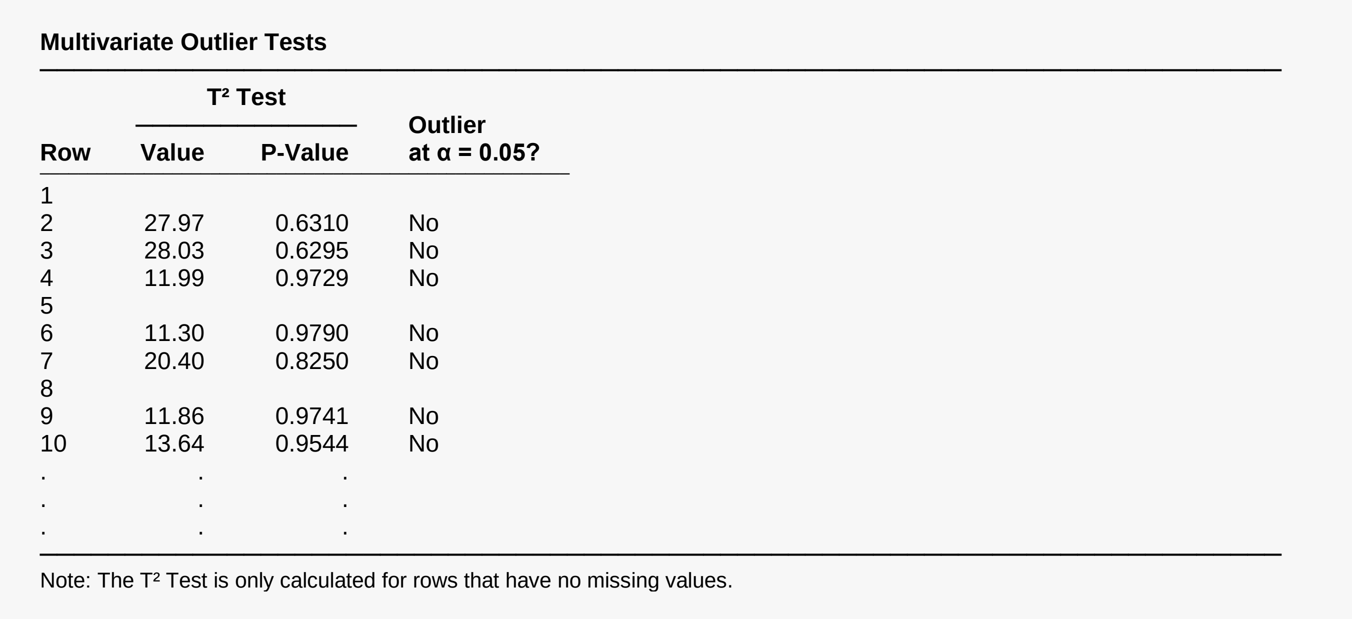


```
List of Discrete Variables and Values
─────────────────────────────────────────────────────────────────────────────
Variable      Value(Count) List
              ──────────────────────────
Q1            1(4), 2(4), 3(5), 4(8), 5(9)
Q2            1(1), 2(10), 3(10), 4(5), 5(4)
Q3            1(9), 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 particularly useful in helping you find out-of-range values in discrete data.

Multivariate Outlier Section



```
Multivariate Outlier Tests
─────────────────────────────────────────────────────────────────────────────
              T² Test
         ──────────────────        Outlier
Row      Value      P-Value        at α = 0.05?
────────────────────────────────────────────────
1
2        27.97      0.6310         No
3        28.03      0.6295         No
4        11.99      0.9729         No
5
6        11.30      0.9790         No
7        20.40      0.8250         No
8
9        11.86      0.9741         No
10       13.64      0.9544         No
.            .           .
.            .           .
.            .           .
─────────────────────────────────────────────────────────────────────────────
Note: The T² Test is only calculated for rows that have no missing values.
```

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:

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

The following mathematical relationship between the T^2 and the F-distribution is used to calculate the *p*-values:

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

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 *p*-value is less than the value indicated in the T^2 Test Alpha box, the observation is indicated as an outlier.

Rows with Missing Values

```
Rows with Missing Values
─────────────────────────────────────────────────────
Row      Pattern of Missing Values
1        |.||
5        .|||
8        |||.
12       |||.
14       ||.|
16       |.||
19       |.||
─────────────────────────────────────────────────────
Key: | = Data, . = Missing


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 Value Estimation Iterations

```
Missing Value Estimation Iteration Details
───────────────────────────────────────────────────────────────────────
                      Covariance     Percent
Iteration     Count   Matrix Trace   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.