Tutorial 5 - Handling Outliers

Introduction

This tutorial teaches you how to handle outliers in your data. An outlier is a value that is significantly outside the range of the other values. Start Change-Point Analyzer and open the file **Trade Deficit.cpa**. This file was copied to your hard drive when you installed the software. It can be found in the directory **C:\Program Files\Taylor Enterprises\Change-Point Analyzer**. It is the same trade deficit data that was analyzed in Tutorial 1. To create an outlier, change the value for June 1987 to 30.

Analyzing the Data

Start by performing a fast analysis on the data in column B. Column A has already been selected as labels. Highlight column B and press the Fast Analysis button. The analysis starts by verifying the assumptions. The program will detect the outlier and display the dialog box shown in Figure 1 to notify you of the problem. The dialog box advises you to check the Assumptions tab once the analysis is completed. Click the OK button and the analysis will continue.
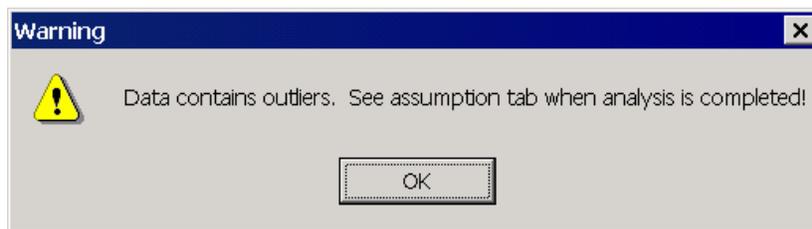

Figure 1: Dialog Box Warning of Outlier

Once the analysis is completed, the Analysis Results window will appear. Immediately click on the Assumptions tab (Figure 2). You will see a message that outliers have been detected along with advice for handling the outliers. This message indicates that outliers make it more difficult to detect changes. It recommends analyzing the ranks using the Custom Analysis dialog box.
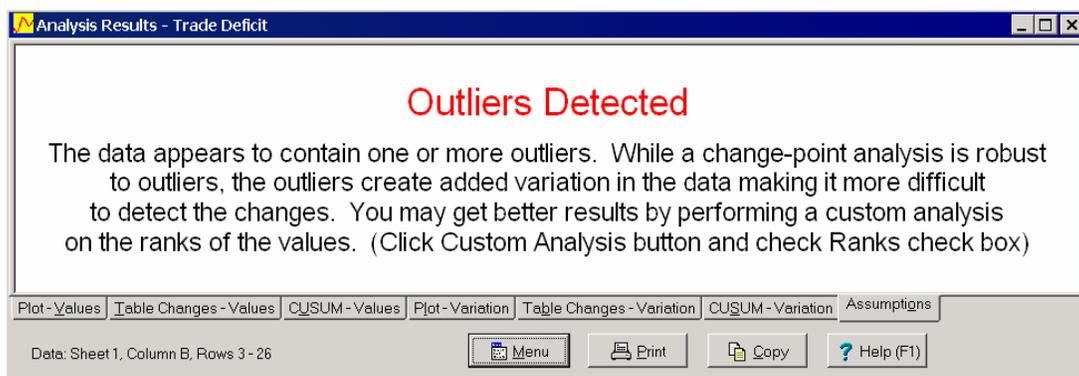

Figure 2: Assumption Tab

Now look at the results for the analysis just performed (Figure 3). The outlier is clearly evident. The blue shading indicates one change was found around November 1987. When this same data, without the outlier, was analyzed in Tutorial 1, two changes were detected. The outlier has created additional variation or noise in the data causing the second change to be missed.
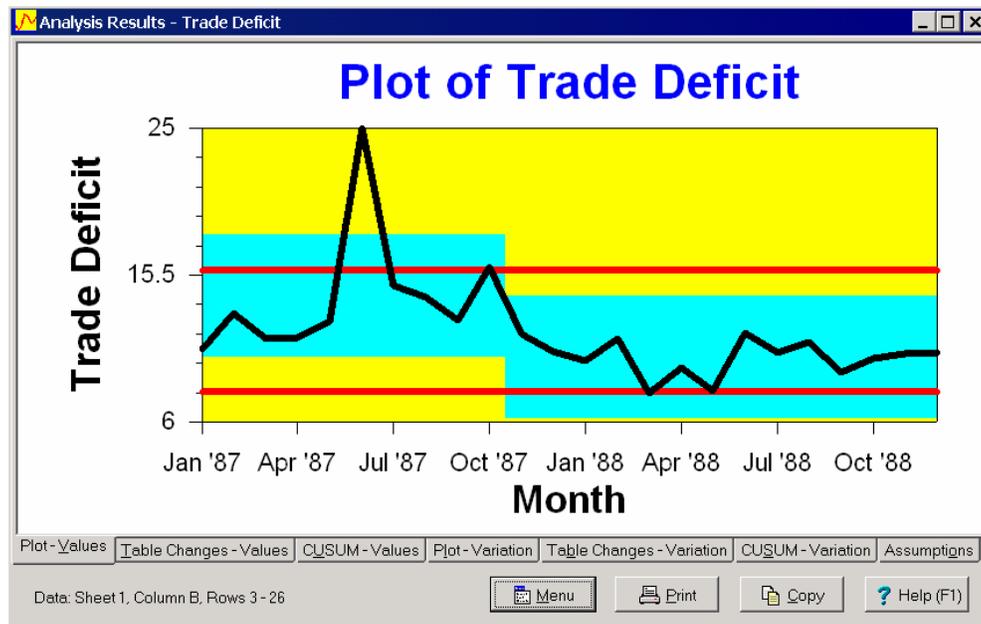

Figure 3: Tab 1 of Analysis Results Window - Analyzed Values

Analyzing Ranks

To analyze ranks, a custom analysis must be performed. Click the Custom Analysis button and select the Rank menu item. Alternatively, display the Custom Analysis dialog box by selecting the last menu item.

Figure 4: Custom Analysis Popup Menu

Once the Custom Analysis dialog box is displayed, select the Handling Violation of Assumptions tab. Then check the Ranks check box as shown in Figure 5.
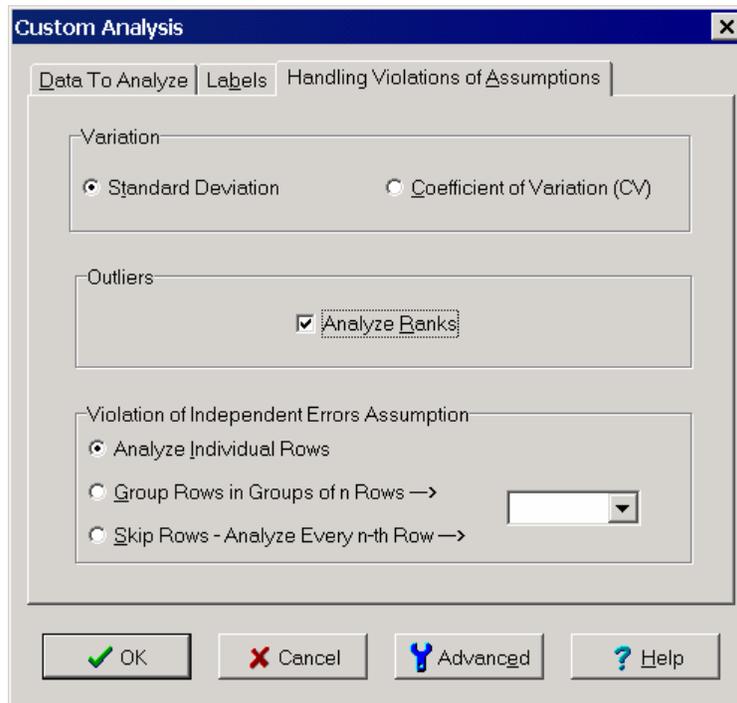
Figure 5: Handling Violation of Assumptions Tab of Custom Analysis Dialog Box

Ranks are determined by sorting the data into order and replacing each value by its relative position in the order. For example, suppose the original data consists of the values:

      1.9   2.7   2.1   3.1   2.5

In sorted order, the values are:

      1.9   2.1   2.5   2.7   3.1

These values are replaced by their relative positions:

      1   2   3   4   5

The largest value is given a rank of 5 and the smallest value is given a rank of 1. This results in larger ranks being associated with larger values and smaller ranks being associated with smaller values. This makes the interpretation of the CUSUM plots simpler. The end result is the following ranked data set:

      1   4   2   5   3

This is the data actually analyzed when the analysis is based on ranked data.

The results of analyzing the ranks is shown in Figures 5 and 6. Two changes are now detected. The results are consistent with those obtained when analyzing the original data in Tutorial 1.
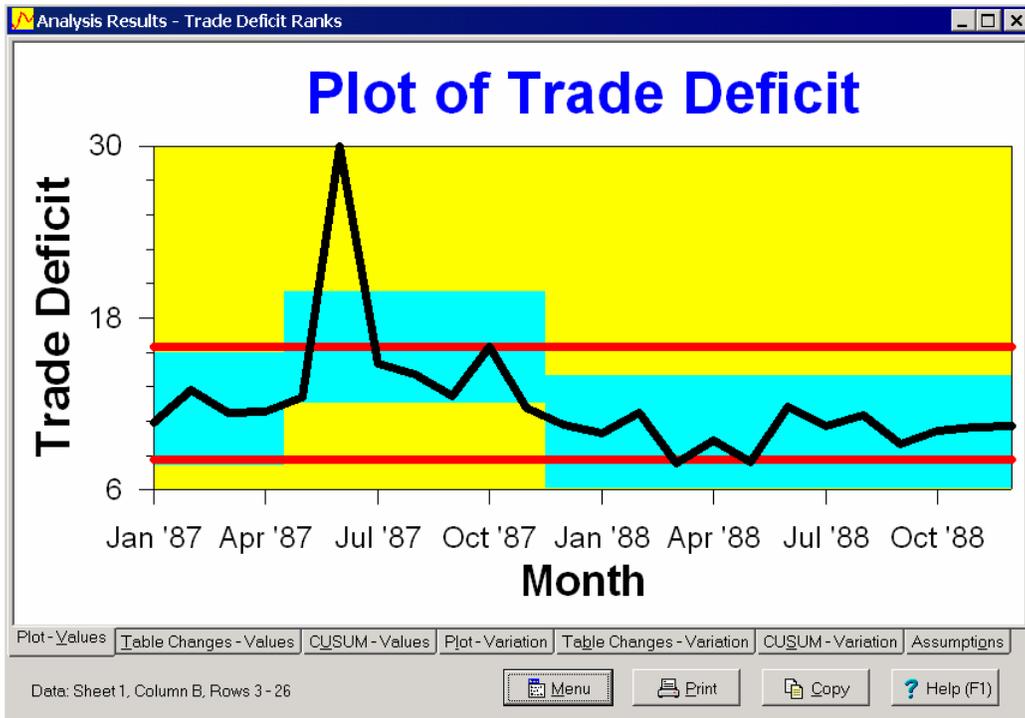
Figure 6: Tab 1 of Analysis Results Window - Analyzed Ranks

Figure 7: Tab 2 of Analysis Results Window - Analyzed Ranks

Conclusion

By analyzing ranks, data containing outliers can be successfully analyzed. However, you don't have to wait until the program detects outliers to analyze the ranks. Other data sets, such as those with heavy tails, can also benefit from this feature.