Chapter 203

Grubbs' Outlier Test

Introduction

It is well known that outliers (extreme points) often distort the results of an analysis. Because of this, every analysis should begin with either a graphical or statistical check about the possibility of outliers. This procedure computes Grubbs' test (1950) for detecting outliers in normal populations. It also computes Rosner's (2011) test for many outliers. We also recommend Barnett and Lewis (1994) for many more outlier tests.

Deleting Outliers

Once outliers have been detected, a decision has to be made about what to do with them. Most would agree that at the minimum, these values should be investigated for input errors. If the outlier appears to be valid, then what? Some would recommend adopting a technique that performs well when outliers are present such as robust or nonparametric techniques. A pragmatic approach is to run the analysis with and without the outliers and watching whether the fundamental conclusions change.

Most statisticians would agree that outliers should not be removed automatically. They should be carefully studied. One question that must be asked is will more outliers likely occur? If so, they must be dealt with.

Technical Details

This section provides the technical details of this test. We follow the presentation of Rosner (2011).

Grubbs' Test for a Single Outlier

Grubbs' (1950) procedure tests the hypothesis that the value that is the furthest from the sample mean is an outlier. Suppose you have a sample of *n* observations, labelled X_1 to X_n, that are assumed to follow the normal distribution. To test the null hypothesis that no outlier is present versus the alternative hypothesis that a single outlier is present, Grubbs proposed that the quantity ESD (extreme studentize deviate) be used where

$$ESD = max_{i=1,\ldots,n} \frac{|X_i - \bar{X}|}{s}$$

The significance level of ESD can be determined using

$$ESD = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2}{n-2+t^2}}$$

where t is short for $t_{n-2,p}$ and $p = 1 - \alpha/(2n)$. Grubbs' tabulated this result, but we have solved this so that a p-value can be calculated.

The above is for two-sided tests. For one-sided tests, substitute $\alpha/(n)$ for $\alpha/(2n)$.

Rosner's ESD Many-Outlier Procedure

You might be tempted to apply Grubbs' test to the maximum ESD, then removing it and recalculating Grubbs' test on the reduced sample, and so on until a non-significant result is found. However, because of features of multiply outliers called swamping and masking, this procedure will not find all of the outliers. However, Rosner (2011) presents a procedure that will identify a block of outliers, assuming that *n* is greater than 20. His procedure is as follows.

1. Determine a value *k* that is a few larger than the anticipated number of outliers. This may be a number or a percentage of the sample size.
2. Compute the mean, standard deviation, and $ESD_{(n)}$ for the sample of *n* value.
3. Calculate the significance level of $ESD_{(n)}$. Call it $P_{ESD(n)}$.
4. Remove the value from value corresponding to ESD(n).
5. Repeat steps 2, 3, and 4 on each successive sample after removing value corresponding to ESD at each step.
6. Consider the *k* values: $P_{ESD(n)}$ $P_{ESD(n-1)}$, $P_{ESD(n-2)}$, …, $P_{ESD(n-k+1)}$.
7. Beginning with $P_{ESD(n-k+1)}$, look for the first value that is less than a preset significance level, say 0.05.
8. Designate this value as an outlier and all values before it also as outliers, whether their individual values of P_{ESD} were significant or not. For example, suppose n is 50 and k is 5, and the values of P_{ESD} beginning with $P_{ESD(50)}$ are: $P_{ESD(50)} = 0.107$, $P_{ESD(49)} = 0.033$, $P_{ESD(48)} = 0.0482$, $P_{ESD(47)} = 0.203$, and $P_{ESD(46)} = 0.428$. If α were 0.05, we would design the observation corresponding to the first 3 values of ESD as outliers, even though the $P_{ESD(50)}$ by itself was not significant according to Grubbs' test.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This option specifies the variables that will be used in the analysis.

Data Input Type

In this procedure, there are four ways to enter the data on the spreadsheet for analysis. Specify which method you want to use.

- **One Response Variable**

  A single report is calculated for a single column.

  Example Dataset

  Grade
  3.4
  3.7
  2.6
  3.2
  3.3
  2.8

- **One Response Variable**

  A single report is calculated for a single column.

  Example Dataset

Grade
3.4
3.7
2.6
3.2
3.3
2.8

- **One Response Variable and One Group Variable**

  The spreadsheet includes a numeric response variable and grouping variable. A separate analysis is performed for each unique value of the grouping variable.

  Example Dataset

Grp	Response
1	3.4
1	4.7
2	5.1
2	6.3

- **Multiple Response Variables, Analyzed Individually**

  Select the columns you want to analyze. A separate, independent analysis is performed for each column.

  Example Dataset

Y1	Y2
3.2	3.2
4.7	2.3
2.8	4.8
3.5	2.6
4.4	2.4
2.3	4.2

- **Multiple Response Variables, Analyzed Together**

  Select all columns you want to analyze. The data in these columns is joined together and a single analysis is performed.

  Example Dataset

Y1	Y2
3.2	3.2
4.7	2.3
2.8	4.0
3.5	2.6
4.0	2.4
2.3	4.2

Variable(s)

Response Variable (Data Input Type = One Response Variable)

Select a column that contains numeric values to be analyzed. A single analysis is performed on the data in this column.

Response Variable (Data Input Type = One Response Variable and One Group Variable)

Select a column that constains numeric values to be analyzed. These data will be separated into groups according to the values in the Group Variable. A separate analysis is performed for each group.

Example Dataset

Grp	Response
A	3.1
A	4.3
A	5.5
A	2.8
B	2.9
B	5.0
A	4.1
A	6.6
B	3.7
B	5.6
A	5.4
B	3.2

Grouping Variable (Data Input Type = One Response Variable and One Group Variable)

Select a column whose values are used to separate the response values into groups or batches. Each group is analyzed separately and individual reports are output.

Response Variables (Data Input Type = Multiple Response Variables, Analyzed Individually)

Enter one or more columns containing numeric values. A separate analysis is performed for each column.

You can enter the variables names or numbers directly, or double-click in the box to display a Column Selection window that will let you select the variables from a list.

Example Dataset

Y1	Y2
3.2	3.2
4.7	2.3
2.8	4.0
3.5	2.6
4.0	2.4
2.3	4.2

Response Variables (Data Input Type = Multiple Response Variables, Analyzed Together)

Enter one or more columns containing numeric values. All numeric data in these columns is combined into a single variable. A single analysis is performed on this combined variable as if it were a single column of data.

You can enter the variables names or numbers directly, or double-click in the box to display a Column Selection window that will let you select the variables from a list.

Example Dataset

Y1	Y2
3.2	3.2
4.7	2.3
2.8	4.0
3.5	2.6
4.0	2.4
2.3	4.2

Hypothesis Type

Specify whether the alternative hypothesis is two-sided or one-sided. If it is one-sided, specify whether you want to only consider values below the mean or above the mean.

Note that you should make this choice before seeing your data.

- **Two-Sided**

  This is the recommended choice unless you have a good reason to choose otherwise. Values that are way above the mean and way below the mean will be considered as possible outliers.

- **One-Sided Minimum**

  Only values below the mean will be considered as outliers.

- **One-Sided Maximum**

  Only values above the mean will be considered as outliers.

Maximum Number of Possible Outliers: Use the Minimum of

Percent of Data (%)

Rosner's ESD Many Outlier Procedure requires you to specify an upper bound, *k*, for the number of outliers. This option specifies *k* as a percentage of *n*, the total number of values.

For example, if there were 67 values and you specified 10% here, *k* would be 6.

A value of 10% is often used.

Note: the final value of *k* is the minimum of this value and the Count specified below.

Count

Rosner's ESD Many Outlier Procedure requires you to specify an upper bound, *k*, for the number of outliers. This option specifies *k* directly. A value between 5 and 10 is often used.

Note: the final value of *k* is the minimum of this value and the Percent of Data specified above.

Reports Tab

The options on this panel specify which reports will be included in the output.

Select Reports

Run Summary

Check the box to output this report.

Grubbs' Single-Outlier and Rosner's Many-Outlier Report

Check the box to output this report.

Outlier Tests Significance Level

Alpha

The alpha level used to determine is an observation is an outlier. If its probability level is less than this value, it and all observations more extreme are designated as outliers. The recommended value is 0.05. The range of possible values is 0.001 to 0.200.

Report Options Tab

The options on this panel control the label and decimal options of the report.

Report Options

Variable Names

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

Decimal Places

Data and Means to Prob Levels.

These options specify the number of decimal places used in the reports. If one of the Auto options is used, the ending zero digits are not shown. For example, if 'Significant Digits (Up to 7)' is chosen, 0.0500 is displayed as 0.05 and 1.314583689 is displayed as 1.314584.

The output formatting system is not designed to accommodate (Up to 13), and if chosen, this will likely lead to lines that run on to a second line. This option is included, however, for the rare case when a very large number of decimals is needed.

Plots Tab

The options on this panel control the inclusion and appearance of the plots. The plot format used depends on the setting of the Data Input Type setting

Select Plots

Probability Plot

Check the box to display a normal probability plot. Click the plot format button to change the plot settings.

Example 1 – Search for Outliers in Grubbs Dataset

This section presents an example of how to check a variable for outliers. A test score was calculated on each of 100 subjects. These scores were stored in the first column of the Grubbs database. The researchers wish to check whether there are any outliers in the data.

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 Grubbs' Outlier Test window.

1 Open the Grubbs dataset.

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

2 Open the Grubbs' Outlier Test window.

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

3 Specify the variables.

- Select the **Variables tab**. (This is the default.)
- Set the **Data Input Type** to "One Response Variable"
- Double-click in the **Response Variable** box. This will bring up the variable selection window.
- Select **Score** from the list of variables and then click **Ok**. "Score" will appear in this box.

4 Run the procedure.

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

The following reports and charts will be displayed in the Output window.

Descriptive Statistics

```
Parameter                Value
Count                    100
Mean                     50.1268
Standard Deviation       14.1729
```

This report provides the basic descriptive statistics of the variable.

Grubbs’ Single Outlier Test and Rosner’s ESD Many-Outliers Test

```
Grubbs' Single-Outlier Test and Rosner's ESD Many-Outliers Test for Score
Alternative Hypothesis: Two-Sided

                                 Grubbs'
                                 Single-     Conclude
Value                            Outlier      Outlier
of                                  Test           by
Possible                 ESD        Prob     Rosner’s
Outlier                  |Z|       Level   Procedure?
1.2345                3.4497      0.0381          Yes
98.3123               3.5718      0.0223          Yes
4.1234                3.6787      0.0137          Yes
81.1439               2.6205      0.7519           No
78.8369               2.5302      0.9820           No
```

The report provides the results of Grubbs’ and Rosner’s outlier tests.

Value of Possible Outlier

These are the data values whose outlier test results are shown. The number of values displayed here is control by the ‘Maximum Number of Possible Outliers’ settings. The values that are the most extreme from the sample mean are shown.

ESD |Z|

These are the ESD values corresponding to the data values shown to the left. The first value, 3.4497, is calculated from all the data. The second value, 3.5718, is calculated from the 99 values that remain after the first value, 1.2345, is omitted. The third value, 3.6787, is calculated from the 98 values that remain after both 1.2345 and 98.3123 are omitted. And so on.

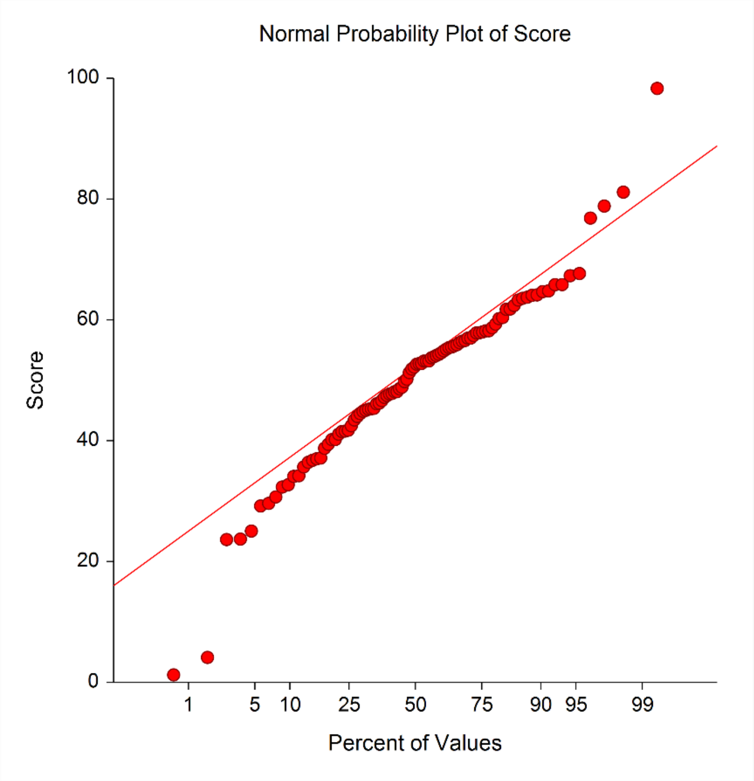
Grubbs’ Single-Outlier Test Prob Level

These are the p-values corresponding to the ESD values to the left. It is possible for the calculation to yield a number greater than one. When this occurs, the p-value is reset to 1.0.

Conclude Outlier by Rosner’s Procedure?

Rosner’s test is conducted by proceeding up the Grubbs Prob Level from the bottom. When a value is found that is less than alpha (e.g. 0.05), that row and all above it are concluded to be statistically significant. In the example, the top three rows. Note that they are concluded to be outliers *even if their individual p-values are greater than 0.05*.

Normal Probability Plot



If the observations fall along a straight line, this indicates the data follow a normal distribution. Outliers are shown at either end of the trend line.