Grubbs' Test for Outliers

The Grubbs' Test for outliers is one of the new techniques in the just release SPC for Excel version 7. To view what is new in version 7, please select [this link](). To see all the statistical tools in SPC for Excel, please select [this link]().

Outliers are extreme values in a dataset. These values can have a large impact on your statistical analysis. For example, suppose you work in a heat treating facility and you take random samples of incoming steel from a supplier and measure the hardness. You want to find a confidence interval that contains the average hardness to compare to what the supplier claims the average hardness is.

What happens to your calculations if there are outliers present in the data? They could easily become skewed because of the outliers, and you could draw the wrong conclusions. There are several ways to detect outliers in a data set. One method is called the Grubbs' test. This test is introduced in this publication.

In this publication:

- [Introduction]()
- [Example Data]()
- [Grubbs' Test Calculations: Two-Sided]()
- [Visual Conformation]()
- [Grubbs' Test Calculations: One-Sided]()
- [Summary]()
- [Quick Links]()

Introduction

The Grubbs' test is a hypothesis test. The null (H_0) hypothesis and alternate (H_1) hypothesis are given below.

H_0: The sample results are all from the same population

H_1: One of the sample results are not from the same population

Note that the alternate hypothesis is that just one sample is not from the same population. To perform the Grubbs' test, it is assumed that the data comes from a normal distribution. You can create a normal probability plot and calculate the Anderson-Darling statistic to confirm that your data are normal. For more information on this, please see our SPC Knowledge Base article [The Anderson-Darling Test for Normality]().

The Grubbs' test is fairly straight forward. You collect your random samples and measure each sample. You should always look at your data to see if any data point looks like it is a potential outlier. If you have any doubts, you need to check to see if the point is an outlier.

If you are using the Grubbs' test you will be calculating a "G" test statistic. Yes "G" for Grubbs. But before you do that calculation, you must decide on which of three cases you want to test:

- Test whether the smallest value is an outlier
- Test whether the largest value is an outlier

- Test the two sided case (either smallest or largest value) for the outlier

The formula for G is different depending on which test you select.

The other thing you must decide on is the value of alpha. Alpha (α) is called the significance level. Typical values of alpha are 0.05 and 0.10. The significance level gives the risk of concluding that an outlier exists when there is no outlier. So, a value of 0.05 for alpha means that is a 5% risk of concluding that an outlier exists when one does not exist.

You will be calculating a p-value. The p-value is the probability of obtaining the calculated G test statistic if the null hypothesis is true. If the p-value is small (less than the alpha you selected), then you conclude that the probability of getting the calculated G test statistic is small and you reject the null hypothesis. If the p-value is large (greater than the alpha value you selected), you then conclude that the probability of getting the calculated G test statistic is large. You accept the null hypothesis and conclude that there are no outliers.

Let's see how this works with some actual data.

Example Data

Suppose you have collected 30 samples and measured each sample for a quality characteristic, X. The data are shown in Table 1.

Table 1: Example Data

Sample	X	Sample	X
1	97.1	16	91.7
2	94.8	17	110.2
3	86.0	18	90.9
4	99.4	19	113.4
5	94.5	20	122.9
6	87.9	21	99.7
7	96.9	22	93.5
8	102.0	23	91.9
9	93.2	24	86.7
10	93.0	25	98.6
11	147.9	26	97.1
12	106.2	27	109.3
13	93.4	28	93.1
14	104.8	29	105.8
15	114.6	30	109.6

We will use this data to demonstrate how the Grubbs' test works.

Grubbs' Test Calculations: Two-Sided

The first step is to decide on the Grubb's test that we will use. We will use the two-sided test so the outlier can be either the maximum value or minimum value in the data set. The second step is to select alpha. We will use 0.05 in this example. The equation for the two-side G test statistic is given by:

$$G = Max\left(\frac{\bar{y} - y_{min}}{s}, \frac{y_{max} - \bar{y}}{s}\right)$$

where $\bar{y}$ is the sample average, s is the sample deviation, y_{min} is the minimum value, and y_{max} is the maximum value.

You can perform the basic calculations for the data in Table 1:

$$\bar{y} = 100.87$$

$$s = 12.62$$

$$y_{min} = 86.0$$

$$y_{max} = 147.9$$

The G test statistic can then be calculated:

$$G = Max\left(\frac{100.87 - 86.0}{12.62}, \frac{147.9 - 100.87}{12.62}\right) = Max(1.18, 3.73) = 3.73$$

What is the G test statistic measuring for the two sided test? It is comparing the number of standard deviations the minimum value is from the average to the number of standard deviations the maximum value is from the average. The G test statistic is the larger of those two – the one most likely to be an outlier.

In this example, the maximum is 3.73 standard deviations from the average.

Now, we must address the question of how likely we are to get a value of 3.73 if there are no outliers. A p-value needs to be calculated. The calculations for the p-value for the two-sided G test statistic is given below.

$$p\ value = 2NP\left(T > \sqrt{\frac{N(N-2)G^2}{(N-1)^2 - NG^2}}\right)$$

where N is the number of values (30 in our example), and P is the probability of T being greater than the terms under the square root. T has N-2 degrees of freedom.

$$p\ value = 2NP\left(T > \sqrt{\frac{N(N-2)G^2}{(N-1)^2 - NG^2}}\right) = 2(30)P\left(T > \sqrt{\frac{30(30-2)3.73^2}{(30-1)^2 - 30(3.73)^2}}\right)$$

$$p\ value = 2(30)P(T > 5.25)$$

To calculate the probability that T is greater than 5.25, you can use the TDIST function in Excel.

$$P(T > 5.25) = TDIST(5.25, 28, 1) = 0.000007$$

This gives a p-value:

$$\text{p-value} = 2(30)(0.000007) = 0.0004$$

Since the p-value is small, we conclude that the probability of getting a G test statistic equal to 3.73 is very small, and there is an outlier present. The maximum value is an outlier.

Visual Confirmation

You can confirm that the maximum value is an outlier by plotting the points on a run chart as shown in Figure 1.

Figure 1: Run Chart
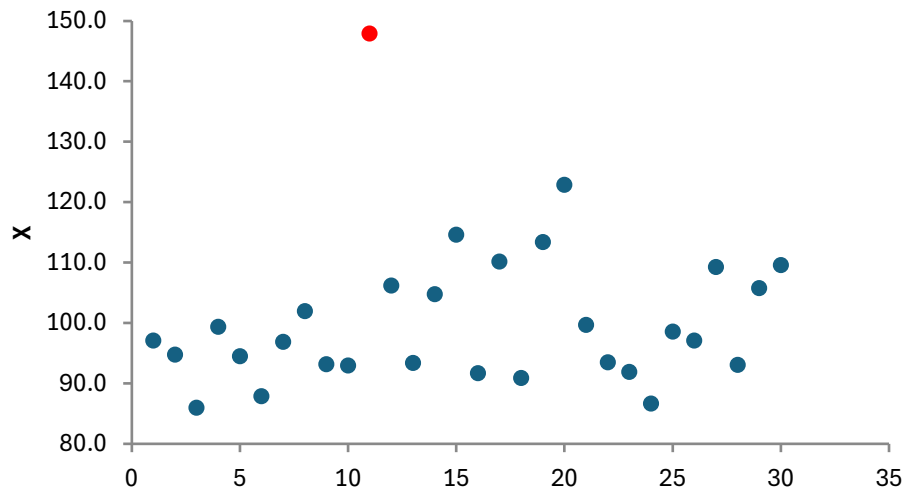
It is easy to see that the maximum point lies well beyond the other data. You will usually plot the data first and then use the Grubbs' test to confirm that an outlier exists.

Grubbs' Test Calculations – One Sided

The example above was the two-sided Grubbs test – looking at both minimum and the maximum. You can also do one-sided tests if you are sure that the outlier may be a minimum or may be the maximum. The equations for these tests are given below.

<u>Test whether the smallest value is an outlier</u>

The G test statistic for testing if the minimum value is an outlier is given below:

$$G = \frac{\bar{y} - y_{min}}{s}$$

The p-value is given by:

$$p - value = NP\left(T > \sqrt{\frac{N(N-2)G^2}{(N-1)^2 - NG^2}}\right)$$

<u>Test whether the largest value is an outlier</u>

The G test statistic for testing if the largest value is an outlier is given by:

$$G = \frac{y_{max} - \bar{y}}{s}$$

The p-value is given by the same equation for testing if the minimum value is the outlier.

Summary

This publication has introduced Grubbs' test for an outlier. The data must be normally distributed. The G test statistic is calculated for either the one-sided or two-sided test. The value of alpha is set so that it can be compared to the p-value. If the p-value is less than alpha, then you reject the null hypothesis and assume that there is an outlier present. If the p-value is greater than alpha, then you accept the null hypothesis and assume there are no outliers present. This test only determines if there is one outlier.

Quick Links

[Visit our home page](#)

[SPC for Excel Software](#)

[Download SPC for Excel Demo](#)

[SPC Training](#)

[SPC Consulting](#)

[SPC Knowledge Base](#)

[Ordering Information](#)

Thanks so much for reading our publication. We hope you find it informative and useful. Happy charting and may the data always support your position.

Sincerely,

Dr. Bill McNeese