Common test statistics for suspected results and outliers –

In our routine analytical works, particularly during a new method development process, we often come across analytical datasets which contain suspected values that seem inconsistent with the majority. This observation is quite common in a set of repeated analytical results which broadly resemble a normal distribution. To such data, we often suspect that they are the outcome of a large uncontrollable variation (i.e. a mistake) in procedure, and they can have a large influence on classical statistics, especially standard deviations and variances.

Data given below from a series of determination of total aflatoxins (μg/kg) in a moldy corn sample and also shown in Figure 1 can be taken as a typical example:

15.2, 24.9, 26.2, 27.2, 28.1, 30.7

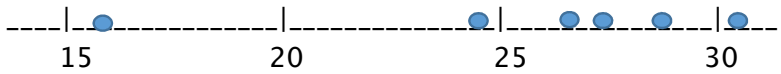

Figure 1: Results of a determination repeated by six analysts. The result of 15.2 μg/kg looks suspiciously like an outlier

It is quite difficult to identify suspected values visually as outliers. Very often, an outlier statistic test is to be performed before further action such as deletion of such data or further testing, as we do not wish to delete such data without sound statistical justification. This is also because such discrepant data may actually be part of the random error system in the course of repeated analyses in the test method concerned.

Statistical tests for outliers abound, including PaHTa's, Chauvenet's, Dixon's, Grubbs, Cochran's, Bartlett's, Hartley's, Levene's, Thompson's and Brown-Forsythe's, etc., but they somehow tend to suffer from some defects. Let us discuss the pros and cons of a few common tests.

The simplest outlier's test is the **PaHTa's rule** which states that if, given a series of test values, $x_1, x_2, \ldots, x_n$ with a mean value $\bar{x}$ and standard deviation s, the difference between the extreme value x_i and the mean value is greater than $3s$, then that value is taken to be a significant outlier. Such conclusion however can only be sound when there is a large number of data for consideration.

The **Chauvenet's** test seems to treat identification of outlier more seriously as it states that if we have a series of test values, $x_1, x_2, \ldots, x_n$ with a mean value $\bar{x}$ and standard deviation s, and if the difference between the extreme value x_i and the mean value is greater than $\omega.s$, where ω value is referred to the Chauvenet's table based on the number of data n, then that particular value is taken to be an outlier. By Chauvenet's accounts, the PaHTa is to be valid only when the total number of data is close to n = 200. The table is shown below:

Table 2: Chauvenet's ω_i values against number of data, n

n	***ω***	***n***	***ω***	***n***	***ω***
3	1.38	**13**	2.07	**23**	2.3
4	1.53	**14**	2.10	**24**	2.3
5	1.65	**15**	2.13	**25**	2.3
6	1.73	**16**	2.15	**30**	2.4
7	1.80	**17**	2.17	**40**	2.5
8	1.86	**18**	2.20	**50**	2.6
9	1.92	**19**	2.22	**75**	2.7
10	1.96	**20**	2.24	**100**	2.8
11	2.00	**21**	2.26	**200**	3.0
12	2.03	**22**	2.28	**500**	3.2

Another simple statistic test, **Dixon's** Q test, requires the dataset to be re-arranged in ascending order and calculate a ratio Q:
Q = | *suspected value – nearest value* | / *(largest value – smallest value)*
with the total number of values noted. If the Q ratio is larger than the critical value corresponding to the same number of data in the Dixon's table, then that suspected value is an outlier with either 95% or 99% confidence. If the calculated Q value is found to be between the critical values of 95% and 99% confidence table, the value is consider as a straggler and the test recommends to have another few repeated analyses to be done for another round of evaluation.

Dixon's test has the following rules to be followed:

1. For values $X_1, X_2, \ldots, X_{n-1}, Xn$ where Xn is suspected to be extremely high:
   - For datasets of 3 through 7 values:
     $Q = (X_n - X_{n-1}) / (X_n - X_1)$

- For datasets of 8 through 12 values:

  $Q = (X_n - X_{n-1}) / (X_n - X_2)$
- For sets of 13 through 40 values:

  $Q = (X_n - X_{n-2}) / (X_n - X_3)$

2. For values $X_1, X_2, ..., X_{n-1}, Xn$ where X_i is suspected to be extremely low:
   - For sets of 3 through 7 values:

     $Q = (X_2 - X_1) / (X_n - X_1)$
   - For sets of 8 through 12 values:

     $Q = (X_2 - X_1) / (X_{n-1} - X_1)$
   - For sets of 13 through 40 values:

     $Q = (X_3 - X_1) / (X_{n-2} - X_1)$

The Dixon's table is appended below in Table 2.

Table 2 : Critical Values for the Dixon Test

Test Criteria	n	95%	99%
	3	0.970	0.994
$D(3...7) = [x_2 - x_1] / [x_n - x_1]$	4	0.829	0.926
Or	5	0.710	0.821
$D(3...7) = [x_n - x_{n-1}]) / [x_n - x_1]$	6	0.628	0.740
(*Whichever is the greater*)	7	0.569	0.680
	8	0.608	0.717
$D(8...12) = [x_2 - x_1] / [x_{n-1} - x_1]$	9	0.564	0.672
Or	10	0.530	0.635
$D(8...12) = [x_n - x_{n-1}]) / [x_n - x_2]$	11	0.502	0.605
(*Whichever is the greater)*	12	0.479	0.579

	13	0.611	0.697
	14	0.586	0.670
	15	0.565	0.647
	16	0.546	0.633
	17	0.529	0.610
	18	0.514	0.594
	19	0.501	0.580
	20	0.489	0.567
	21	0.478	0.555
	22	0.468	0.544
	23	0.459	0.535
$D(13\ldots40) = [x_3 - x_1] / [x_{n-2} - x_1]$	24	0.451	0.526
Or	25	0.443	0.517
$D(13\ldots40) = [x_n - x_{n-2}] / [x_n - x_3]$	26	0.436	0.510
(*Whichever is the greater*)	27	0.429	0.502
	28	0.423	0.495
	29	0.417	0.489
	30	0.412	0.483
	31	0.407	0.477
	32	0.402	0.472
	33	0.397	0.467
	34	0.393	0.462
	35	0.388	0.458
	36	0.384	0.454
	37	0.381	0.450
	38	0.377	0.446
	39	0.374	0.442
	40	0.371	0.438

A problem with this simple test is that it may be foiled by the presence of a second outlier at either end of the value range. Indeed Dixon has tried to consider the second or third largest or smallest values in the *Q* ratio calculated when the dataset grows in number.

The **Grubbs test** is a more sophisticated test for outliers than Dixon's. It is used to detect outliers in a dataset by testing for one outlier at a time. Any outlier which is detected is deleted from the data and the test is repeated until no outliers are detected. However, multiple iterations may change the

probabilities of detection, and the test should not be used for small sample sizes of six or less because it frequently tags most of the points as outliers. The basic assumption underlying the Grubbs test is that, outliers aside, the data are normally distributed. The null hypothesis is that there are no outliers in the dataset.

The test statistic G is calculated for each result x_i from the sample mean $\bar{x}$ and standard deviation s as

$$G = \max | x_i - \bar{x} | / s$$

This statistic calculates the value with the largest absolute deviation from the sample mean in units of the sample standard deviation. This form of the Grubbs test is therefore a two-tailed test. The steps taken are as follow:

Step 1: to quantify how far the outlying figure is from the other data.

Calculate the ratio G as the difference between the suspected value and the mean divided by the standard deviation s, i.e.

$$G = \frac{| x_i - \bar{x} |}{s}$$

The standard deviation s comes from the whole set of data, including the consideration of the outlying data. Hence, one can expect the presence of an outlier increases the calculated standard deviation s but since the presence of an outlier increases both the numerator (difference between mean and the value) and denominator (standard deviation s of all values), the ratio G does not get very large. In fact, no matter how the data are distributed, the G cannot get larger than:

$$\frac{(n-1)}{\sqrt{n}}$$

where n is the number of values under consideration.

Step 2 : Compare the G value against the critical Grubbs G value at the designated confidence level, normally at P = 0.05 with 95% confidence .

If the calculated value of G is greater than the critical value in the Grubbs table (Table 3), then the *P*-value is less than 0.05, i.e. the null hypothesis does not hold and there is more than 95% chance that we would encounter an outlier so far from the others in either direction by chance alone, *assuming* all the data were really sampled from a single Gaussian or normal distribution.

We recall that the Grubbs method only works for testing the most extreme value in the sample. Normally, we would calculate *G* for all values but only estimate a *P*–value for the Grubbs test from the most extreme value of G, as follows:

First we calculate:

$$t = \sqrt{\frac{n(n-2)G^2}{(n-1)^2 - nG^2}}$$

where

n is the number of values in the sample and,
G is calculated for the suspected outlier as shown above.

Now, let us look up the 2–tailed *P*–value for the Student's *t*–distribution table with the calculated value of *t* at (*n-2*) degrees of freedom, and multiply the *P*–value obtained in above by *n*. The result is an approximate *P*–value for the outlier test.

This *P*–value is the chance of observing one point so far from the others if the data were all sampled from a Gaussian distribution. If *G* is large, this *P*–value can be very accurate. For Example:

Trial #	Value	*G* - value	*t* - value	2-Tailed *P*-Value calcd	Estimated P-Value
1	56.5	0.152			
2	56.2	0.419			
3	56.8	0.724			
4	56.5	0.152			
5	56.3	0.229			
6	57.0	1.105			
7	56.4	0.038			
8	57.2	1.486			
9	56.1	0.610			
10	* 55.2	2.324	4.001	0.00395	0.0395
Mean =	56.4				
s =	0.525				
n =	10				
Critical *G* =	2.29				

Note: The 2–tailed *P*–value was calculated using Excel's "=TDIST(*t*,*df*,*2*)" where "2" is for the 2–tailed P value.

Hence it is concluded that the value 55.2 was an outlier with 95% confidence by Grubbs test with *P*–value being less than 0.05.

Table 3 : Critical G Values for Grubbs test at 95% confidence

Critical values for G. *Calculate G as shown above. Look up the critical value of G in the table below, where n is the number of values in the group. If your value of G is higher than the tabulated value, the P-value is less than 0.05.*

n	**Critical G**		***n***	**Critical G**
3	1.15		27	2.86
4	1.48		28	2.88
5	1.71		29	2.89
6	1.89		30	2.91
7	2.02		31	2.92
8	2.13		32	2.94
9	2.21		33	2.95
10	2.29		34	2.97
11	2.34		35	2.98
12	2.41		36	2.99
13	2.46		37	3.00
14	2.51		38	3.01
15	2.55		39	3.03
16	2.59		40	3.04
17	2.62		50	3.13
18	2.65		60	3.20
19	2.68		70	3.26
20	2.71		80	3.31
21	2.73		90	3.35
22	2.76		100	3.38
23	2.78		110	3.42
24	2.80		120	3.44
25	2.82		130	3.47
26	2.84		140	3.49