The Q-test for the Outlying Data Point

Katherine Dorfman
UMass Biology Department 2019

Sometimes you collect data by repeated measurements that seem to cluster around a certain value, except for one odd data point. You want to be honest; you really did get that measurement, and you don't want to cheat. However, it does seem to be from another universe. For example, suppose these are the readings from your experiment:

25.1 21.2 27.5 22.7 23.8 26.3 40.6 22.9

That 40.6 not only looks out of place, but it has an effect on the mean and standard deviation of your data set:

	mean	standard deviation
including maximum value	26.1	6.3
excluding maximum value	24.0	2.1

It may also have an effect on statistical significance when you compare this distribution to another one by a t-test.

Is it a mistake? Is it noise? Did your recording device hiccup? What is the likelihood of getting an outlier this far from the next value when you are picking from a population with these characteristics?

Chemists (and other scientists) often test for outliers with the Q-test. This test calculates the ratio between the putative outlier's distance from its nearest neighbor and the range of values:

Notice that as the distance between the potential outlier and its nearest neighbor increases, so does Q.

$$Q = \frac{|potential\ outlier - nearest\ neighbor|}{maximum\ value - minimum\ value}$$

For our hypothetical data listed above,

$$Q = \frac{|potential\ outlier - nearest\ neighbor|}{maximum\ value - minimum\ value} = \frac{40.6 - 27.5}{40.6 - 21.2} = 0.675$$

The significance test consists of comparing your calculated Q to the theoretical Q that is expected to occur 5% of the time if you were sampling from a population with this mean and standard deviation. The null hypothesis is that all your data points come from the same pool. If your Q is as big as or bigger than the critical Q, then you can reject that null hypothesis with 95% confidence, and safely exclude the one odd data point.

Table 1, on page 2, lists critical values of Q at the 5% significance level. Notice that the smaller the number of data points, the larger Q must be for you to reject a data point. For a set of 8 data points, according to the table, Q would be greater than 0.526 just by luck of the draw 5% of the time or less. Since our experimental Q (0.675) is larger than 0.526, we may safely omit the 40.6 value from our data set, because such a large value is unlikely to occur by chance alone. We can conclude that that data point measures something other than what all the others are measuring. To maintain honesty, we can report something like this: "One very large (or small) value was rejected from the data set by the Q-test with 95% confidence."

You may only reject one data point from a data set by this method.

Excel Tricks to help you calculate Q

Sort

Select your data, then choose Sort from the Data menu. This makes visual inspection for outliers easier. ***Note: if there are identifiers adjacent to the data values, be sure to include them in the sort, or you will lose the connection between the labels and the values!***

MAX

From the f_x menu, under statistical. This returns the maximum value from a list of values. The result goes in the cell where you typed the equation. You can specify the data array either by typing the range of cells where your data are found, or by clicking and dragging over them.

=MAX(data array)

MIN

This function works just like MAX, but tells you the minimum value in a series.

=MIN(data array)

COUNT

This tells you how many numerical values there are in an array. (If you use COUNTA, it will return the number of non-empty cells, rather than the number of data points.)

=COUNT(data array).

LARGE

This function tells you the kth largest value in a series. Set k = 2 to find the second largest value. Set k = n-1 to find the next-to-last value in a set of n data points.

=LARGE(data array, k)

SMALL

This function tells you the kth smallest value in a series. Set k = 2 to find the second smallest value.

=SMALL(data array, k)

ABS

This function returns the absolute value of a number. The number may be the result of a calculation, or may refer to a particular cell in which a calculation is done.

=ABS(number)

=ABS(number1+number2)

Table 1. Critical values of Q

at the 95% confidence ($\alpha = 0.05$) level, for data sets up to n = 30

number of data points[1]	$Q_{0.05}$
3	0.970
4	0.829
5	0.710
6	0.625
7	0.568
8	0.526
9	0.493
10	0.466
11	0.444
12	0.425
13	0.410
14	0.396
15	0.384
16	0.374
17	0.365
18	0.356
19	0.349
20	0.342
21	0.337
22	0.331
23	0.326
24	0.321
25	0.317
26	0.312
27	0.308
28	0.305
29	0.301
30	0.298

[1] Table from David B. Rorabacher, 1991. Statistical treatment for rejection of deviant values: critical values of Dixon's "Q" parameter and related subrange ratios at the 95% confidence level. *Anal. Chem., 63* (2): 139–146

http://pubs.acs.org/doi/abs/10.1021/ac00002a010, accessed 9/10/09.