Stats Tutorial - Dealing with Outliers:

When you take a set of linear measurements, it is all-too common for one or more data points to lie "far away" from the expected value. This sort of data is called an *outlier* and, if it is due to an erroneous measurement, can easily skew your regression line. However, an outlier could also reveal information about an incomplete regression, or the requirement for more a complex regression model.

This is why each calibration point should ideally be obtained at least in triplicate; you can then see if any of the replicates is an outlier by applying the Q-test directly to the replicate data. Alternatively, you can use the mean and standard deviation for each calibration solution in a technique called weighted least squares analysis. A third approach is to use robust regression methods. Both weighted and robust regression are beyond the scope of this tutorial; those interested should consult one of the chemometrics books suggested in the bibliography, or consult the article by del Rio *et al* in *Analyst*, 2001, **126**, 1113-1117.

It is not always practical or possible to take replicates, usually because time is limited. This leaves the problem of how to identify and address potential outliers within your calibration data.

Statistics provides a few tools for dealing with outliers. Some of these methods are only valid for small sample sizes, and none of them are overly reliable for regression analysis. The key is to be careful - if you have too many outliers in your data, it may be an indication that you should redo your experiment, or choose an alternative experimental method to collect your data.

The method we cover here is the *Q*-test. The Q-test is a good first try, but it is not designed for regression analysis, since they require the *y*-values to be independent. It is still acceptable to use a *Q*-test in regression analysis, but be aware that it is not intended for this purpose and care should be taken.

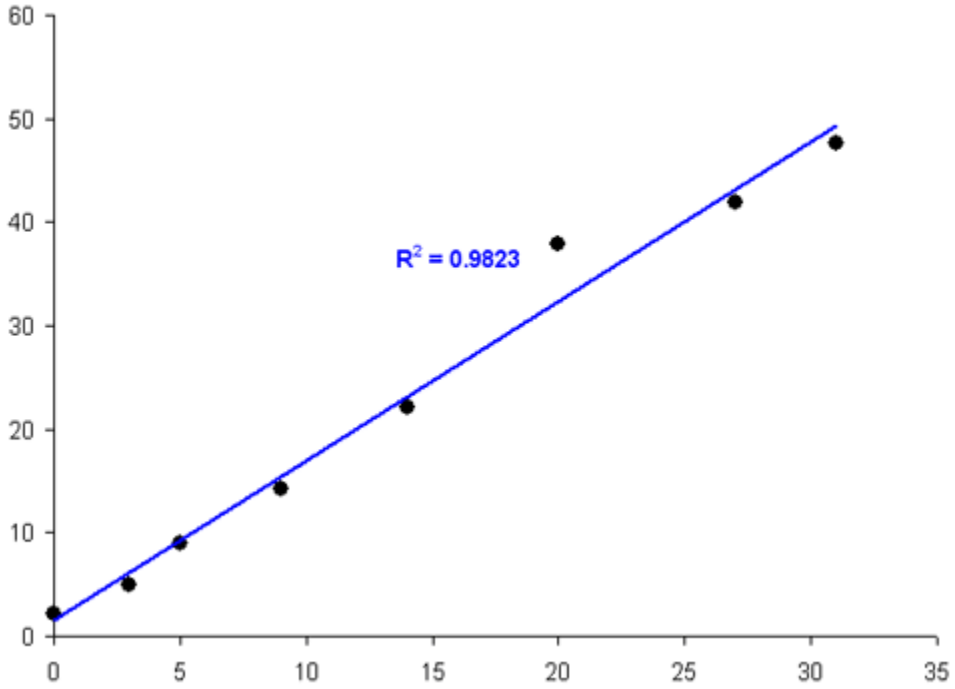
Q-Test

The *Q*-statistic is calculated with the following formula

$$Q = \frac{|y_{suspect} - y_{nearest}|}{|y_{highest} - y_{lowest}|}$$

This formula is not well-suited for regression data, since in a regression analysis, the *y*-values data points change for each consecutive value. For this reason, we need to normalize the data by using the regression residual $(y_i - \hat{y}_i)$, where $\hat{y}_i$ is the *y*-value obtained from the calibration curve for x_i. You can then determine the *Q*-statistic for the residual values, and compare the result to a tabulated value, which are available in any statistics textbooks for 95% confidence intervals. When comparing the calculated value to the tabulated value, your reject the outlier when $\boldsymbol{Q_{calc} > Q_{n,95\%}}$.

Consider the graph shown below to illustrate how to deal with outliers. The value at x_i=20 is possibly an outlier and skewing the regression line. Can we discard it? The residuals are shown beside the graph.



$(y_i - \hat{y}_i)$
0.6
-1.1
-0.2
-1.1
-0.9
5.6
-1.2
-1.7

In this case, $y_{nearest}$ is 0.6 and $y_{smallest}$ is -1.7. Applying the *Q*-test, we find Q_{calc}=0.684. Referring to a table of *Q*-values for n=8 and a 95% confidence interval, we find $Q_{8,95\%}$=0.526. Since $Q_{calc} > Q_{8,95\%}$, we can reject the outlier. Bear in mind that this is for the 95% confidence interval, so there is still 1 chance in 20 that the data point is a real value and should not have been rejected.

This concludes the section on linear regression and calibration curves. From this lesson, you should have all the statistical tools you need to create linear regression calibration curves and analyze the errors associated with determining unknown sample concentrations from a measured signal.

The following and last section covers more advanced statistics used for comparing sets of data based on mean and variance, as well as a more detailed look at some of the statistical concepts discussed in earlier sections.

Last updated: September 26th, 2006

Q test From Wikipedia, the free encyclopedia

Retrieved on 2/4/2008 from http://en.wikipedia.org/wiki/Q_test

In statistics, the ***Q* test** is used for identification and rejection of outliers. This test should be used sparingly and never more than once in a data set. To apply a *Q* test for bad data, arrange the data in order of increasing values and calculate *Q* as defined:

$Q = Q_{gap}/Q_{range}$

Where Q_{gap} is the absolute difference between the outlier in question and the closest number to it.
If $Q_{calculated} > Q_{table}$ then reject the questionable point.

Table

Number of values:	3	4	5	6	7	8	9	10
$Q_{90\%}$:	0.941	0.765	0.642	0.560	0.507	0.468	0.437	0.412
$Q_{95\%}$:	0.970	0.829	0.710	0.625	0.568	0.526	0.493	0.466

Example

For the data:

0.189,0.169,0.187,0.183,0.186,0.182,0.181,0.184,0.181,0.177

Arranged in increasing order:

0.169,0.177,0.181,0.181,0.182,0.183,0.184,0.186,0.187,0.189

Outlier is 0.169. Calculate *Q*:

$$Q = \frac{\text{gap}}{\text{range}} = \frac{(0.177 - 0.169)}{(0.189 - 0.169)} = 0.400.$$

With 10 observations at 90% confidence, $Q_{calculated} < Q_{table}$. Therefore keep 0.169 at 90% confidence.