References

1. "CRITERION FOR THE REJECTION OF DOUBTFUL OBSERVATIONS," Peirce, Benjamin; The Astronomical Journal 45, Vol II, No 21, July 24, 1852.
2. "Comparison of Outliers Elimination Algorithms," P. Mostarac, R. Malarić, H. Hegedušić; Proceedings of the 7th International Conference, Smolenice, Slovakia, 2009.

Introduction

One of the problems encountered when making experimental measurements is determining whether a particular measurement is not valid; many times a parameter is much larger or smaller than it should be, and the experimentalist would like to "discard" that datum and base his conclusions on only the "valid" data. It turns out that a rigorous statistical method for disqualifying one or more points in a data set was established by Benjamin Peirce (Ref 1). That method was not amenable to easy calculation however and simpler less rigorous methods were employed. Reference 2 discussed both Peirce's method and the easier to employ Chauvenet's method. With the advent of computers and numerical solvers, users of Mathcad can use Ref 2 equations to easily determine when data is invalid. By setting boundaries for outlier elimination, strictness of criterion is defined and arbitration over suspicious data is done.

Criterion for eliminating outliers can be defined by the amount of allowed deviation comparing to standard deviation σ. For N measurements of x, with standard deviation σ, mean μ and suspicious data *x* defined as:

$$n = max\left(\frac{|x-\mu|}{\sigma}\right) \qquad [1]$$

It can be defined [1] that probability in N measurements must be greater or equal to *p* so it can be kept:

$$p \geq N \cdot (1 - P(n \cdot \sigma)) \qquad [2]$$

P is a function defined like integral of probability density function (*pdf*) on interval ±nσ.

$$n \leq \frac{\sqrt{2}}{erf\left(\frac{p}{N} - 1\right)} \qquad [3, 4]$$

For Chauvenet's criterion *p*=0,5. And *p* can be chosen such that our criterion be less or more rigorous.

Peirce's criterion

Peirce's criterion is a more rigorous than Chauvenet's criterion. Peirce's criterion is also able to remove several suspicious data. It is an iterative method based on theory of probability. In [2] it is described and also required conditions for rejection are presented. We will explain the crucial parts of Peirce's criterion. The principle is that the *k* suspicious data *should be rejected when the probability* P, with all data including the suspicious data, *of the system of the errors is less then* a probability without suspicious data multiplied by *the probability of making* those suspicious observations P1.

Let:
k be the number of suspicious data
m be the number of unknown quantities contained in the observations.
Eq [5] and [6] are the same inequality, expressed in different variables.

$$P < P1 \qquad (5)$$

$$\lambda^{N-k} \cdot R^k < Q^N \qquad (6)$$

$$\lambda^2 = \frac{N-m-k \cdot n^2}{N-m-k} \qquad (7)$$

$$R = e^{\frac{n^2-1}{2}} \cdot erf\left(\frac{n}{\sqrt{2}}\right) \qquad (8)$$

$$Q^N = \frac{k^k \cdot (N-k)^{N-k}}{N^N} \qquad (9)$$

Define Functions:

$$\lambda(N, m, k, n) := \sqrt{\frac{N-m-k \cdot n^2}{N-m-k}}$$

$$R(n) := e^{\frac{n^2-1}{2}} \cdot erf\left(\frac{n}{\sqrt{2}}\right)$$

$$QN(N, k) := \frac{k^k \cdot (N-k)^{N-k}}{N^N}$$

Solve for limiting value of n:

$$FN(N, m, k, n) := \lambda(N, m, k, n)^{N-k} \cdot R(n)^k - QN(N, k)$$

$$n_a(N, m, k) := \mathbf{root}(FN(N, m, k, n), n, 1, 20)$$

Case Studies:

One kmown outlier:

Open the area to see the development

$N := \text{rows}(x) = 30$ $\mu := mean(x) = 0.061$ $\sigma := stdev(x) = 0.319$

$k := 1$ $m := 1$ $n := \max\left(\overrightarrow{\frac{|x-\mu|}{\sigma}}\right) = 4.505$

$FN(N,m,k,n) = -0.012$ $\lambda(N,m,k,n)^{N-k} = 4.368 \cdot 10^{-8}$

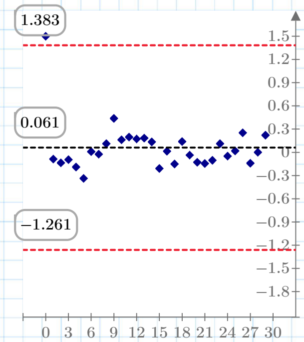
$R(n)^{k} = 15510.423$ $QN(N,k) = 0.012$

$n_a(N,m,k) = 4.14$ $ii := 0 .. \text{rows}(x) - 1$

$Ans(N,m,k,n) := \text{if}\left(n > n_a(N,m,k), \text{“outlier”}, \text{“valid”}\right)$

$Ans(N,m,k,n) = \text{“outlier”}$

Data, mean of all data, limits



Peirce's example 1:

$N := \text{rows}(x) = 30$ $\mu := mean(x) = -0.002$ $\sigma := stdev(x) = 0.398$

$k := 1$ $m := 1$ $n := \max\left(\overrightarrow{\frac{|x-\mu|}{\sigma}}\right) = 2.089$ $k = 1$

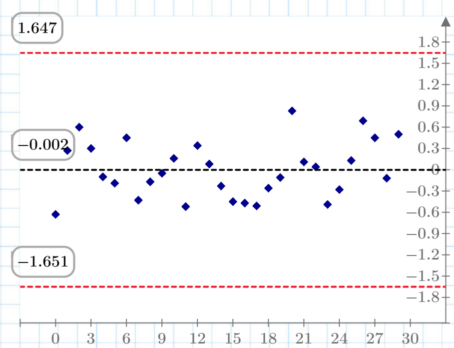
$FN(N,m,k,n) = 0.797$ $\lambda(N,m,k,n)^{N-k} = 1.563 \cdot 10^{-1}$

$R(n)^{k} = 5.18$ $QN(N,k) = 0.012$

$n_a(N,m,k) = 4.14$

$Ans(N,m,k,n) = \text{“valid”}$ Peirce: "neither of them is to be rejected"

Data, mean of all data, limits



No outlier:

$dst := \text{rnorm}(30, 0, 0.01)$ $\max(dst) = 0.021$ $min(dst) = -0.017$ $stdev(dst) = 0.009$

$ol := 1$ $mean(dst) = 0$

$x := dst$

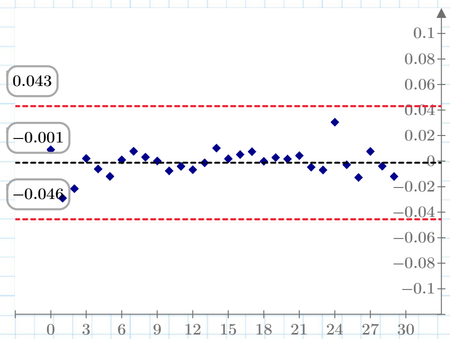
$N := \text{rows}(x) = 30$ $\mu := mean(x) = 0$ $\sigma := stdev(x) = 0.009$

$k := 1$ $m := 1$ $n := \max\left(\frac{\overrightarrow{|x-\mu|}}{\sigma}\right) = 2.329$

$n_a(N, m, k) = 4.14$ $R(n_a(N, m, k)) = 3193.292$ $\lambda(N, m, k, n)^{N-k} = 8.829 \cdot 10^{-3}$

$Ans(N, m, k, n) = \text{"valid"}$

Data, mean of all data, limits



Data fits a straight line

Find the least squares straight line fit $a_i := 1$ $A := \text{augment}(a, z)$ $Ap := A^{T} \cdot A$ $Ab := A^{T} \cdot xx$

$cc := Ap^{-1} \cdot Ab$

$fn(x) := \sum_{i=0}^{\text{rows}(cc)-1} cc_i \cdot x^i$ $cc^{T} = [-0.8653 \quad 2.0181]$

$N := \text{rows}(xx) = 30$ $m := 2$ $k := 1$

standard deviation of fit: $\sigma := \sqrt{\left(\frac{1}{\text{rows}(xx)-1} \cdot \sum_{i=0}^{\text{rows}(x)-1} \left(xx_i - fn(z_i)\right)^2\right)}$ $\sigma = 2.747$

$er := \overrightarrow{xx - fn(z)}$ $n := \max\left(\frac{er}{\mu}\right) = 3346.749$ $n_a(N, m, k) = 4.022$ $Ans(N, m, k, n) = \text{"outlier"}$

Data, mean of all data, limits