Robust Measures of Scale

Or

“Why isn't sigma-cutting working well?”

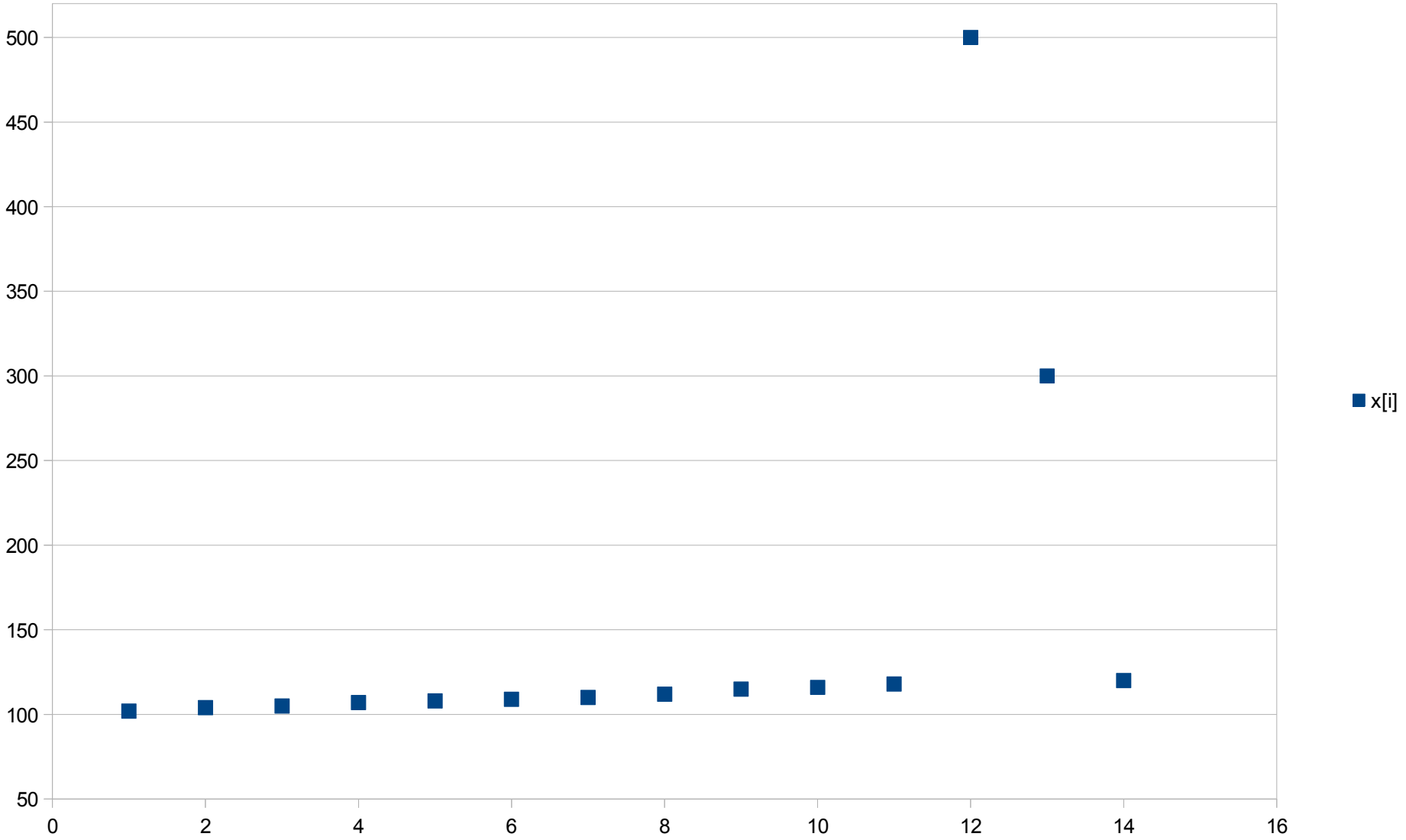
by Guilherme Teixeira

Programmer's Club 2015-06-25

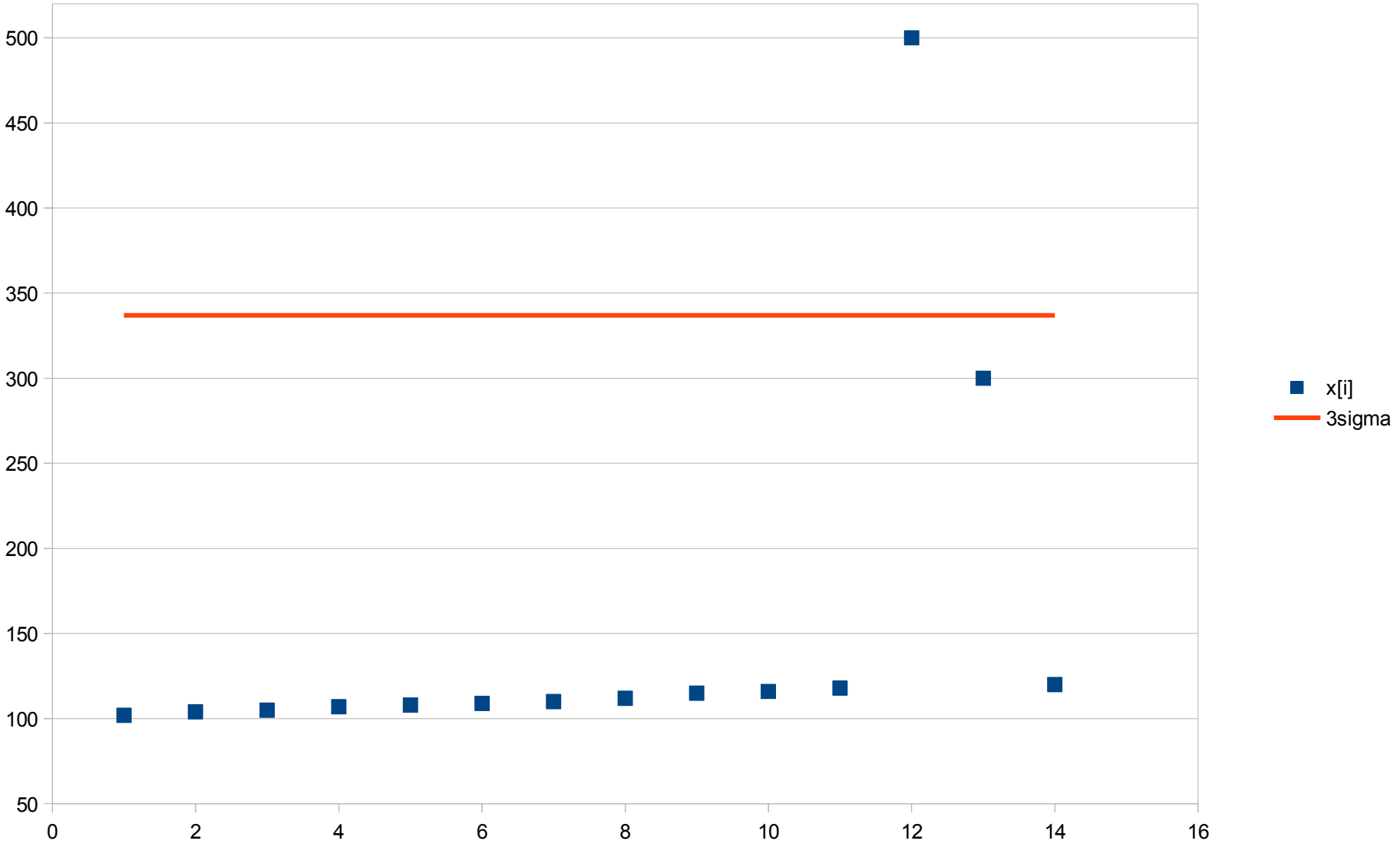
So, what's the issue?

- *When analyzing huge data sets we try to fit functions but we have to deal with Outliers*
- *One of the most used Outlier-characterizing indicator is the standard deviation of the data distribution*
  - *Removing data more than 3-sigma away*
  - *Considering signals with more than 3-sigma as significant*
- *Standard deviation is sensitive to extreme outliers!*

Seriously, how bad can it be?



Seriously, how bad can it be?



But isn't standard deviation the only way to do it?

- *There are several alternatives to standard deviation => Robust statistical indicators*
- *Median Absolute Deviation*

$$MAD = median_i\left(\left|X_i - median_j\left(X_j\right)\right|\right)$$

- *Interquartile Range*

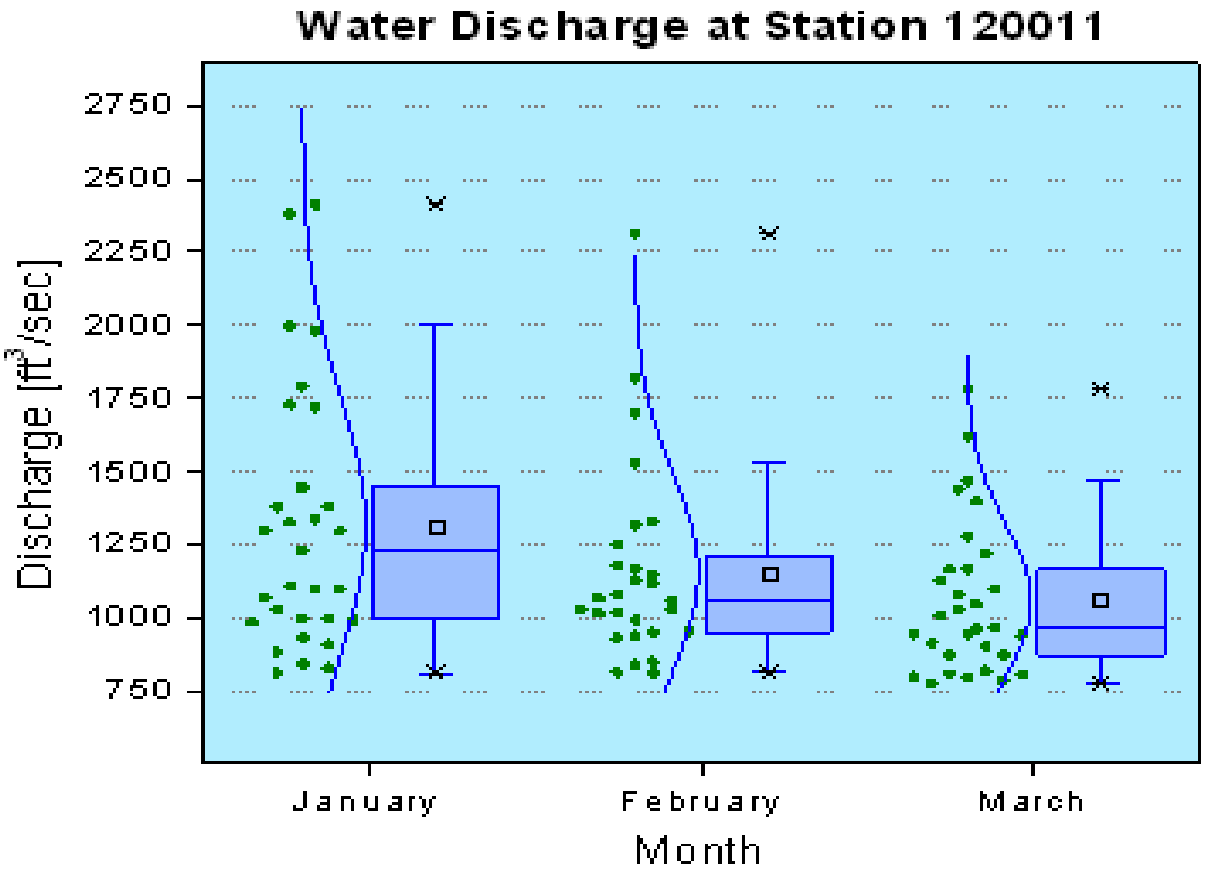
$$IQR = Q_3 - Q_1$$

MAD

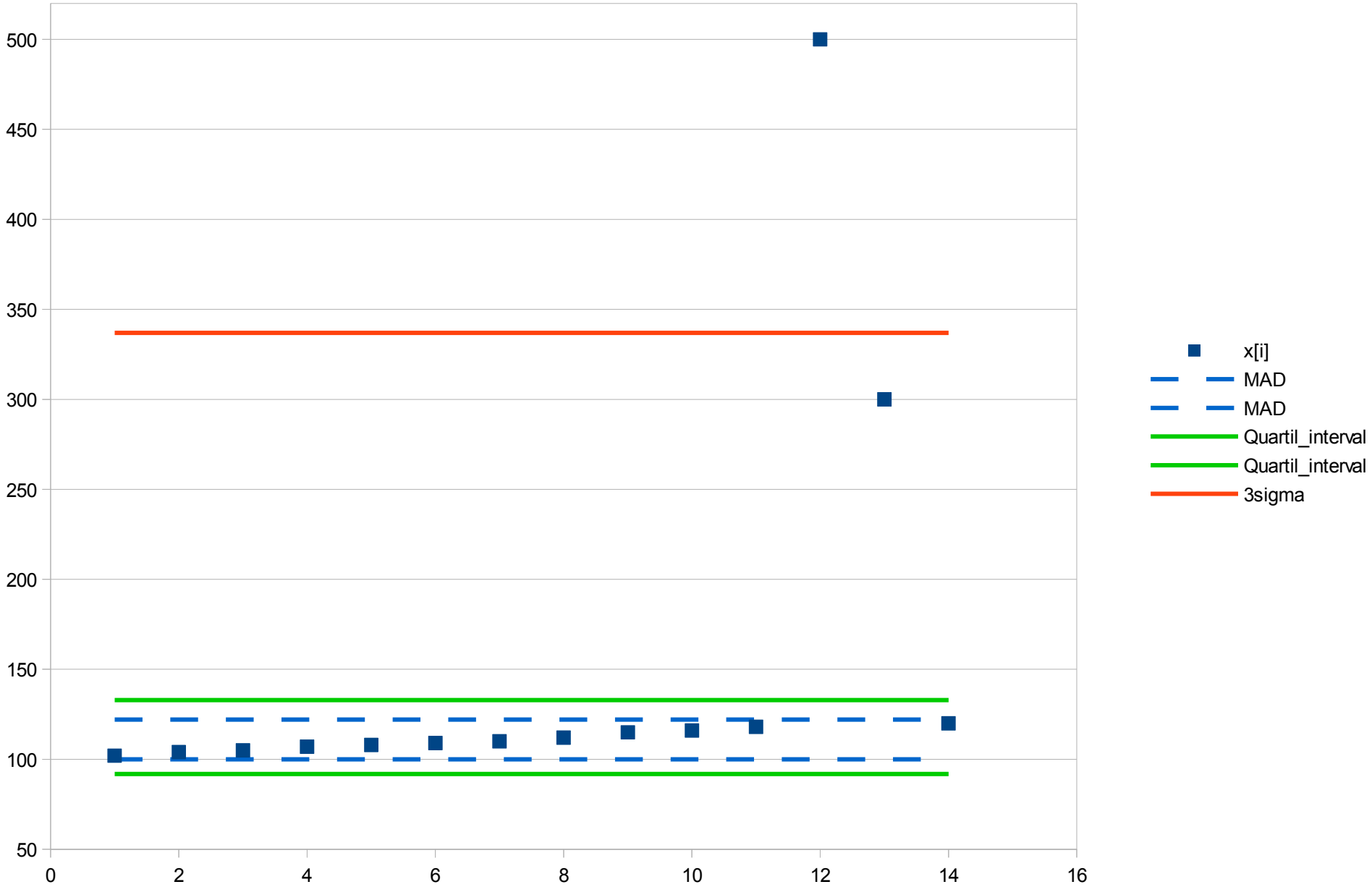
- *Less sensitive to outliers than standard deviation*
- *Implication that the distribution of values is simmetric*
- *The scale factor depends on the distribution*

IQR

- *Difference between the 3rd and 1st quartile*
- *Actually used for box-plot representations of distributions*
- *Robust*



Seriously, how bad can it be?



Ok, ok. But aren't all these formulas hard to apply?

- *Actually both MAD and IQR are easily computed and applied*
- *MAD relies on the power of the Median of a distribution. A good outlier cut would be:*

$$[Median - 2 * MAD, Median + 2 * MAD]$$

- *IQR relies on the power of the 1st and 3rd Quartil of a distribution. A good outlier cut:*

$$[Q1 - 1.5 * IQR, Q3 + 1.5 * IQR]$$

Sure, but what if I don't know how to program them?

- *Luckily there are already coded alternatives:*
  - *MAD:*

    http://statsmodels.sourceforge.net/

    R-stats

    Matlab:Statistics and Machine Learning toolbox

    NAG Fortran Library
  - *IQR:*
    Python – use numpy percentile function to get quartils

    NAG Fortran Library

    Matlab:Statistics and Machine Learning toolbox

    R-Stats

So, should we stop using standard deviations altogether?

- *Definite no!*
- *MAD relies on an implied simmetry of the distribution*
- *Both MAD and IQR don't give as much information as standard deviation.*
- *MAD or IQR should be used as an initial cut, removing extreme outliers, and then move on to a standard deviation analysis.*

What's the takeway?

- *Always improve. Always adapt. Don't get set in a way of doing things.*

  "You must be shapeless, formless, like water. When you pour water in a cup, it becomes the cup. When you pour water in a bottle, it becomes the bottle. When you pour water in a teapot, it becomes the teapot. Water can drip and it can crash. Become like water my friend."- Bruce Lee

Thanks

For

Listening!