Robust Statistics – MAD Method

In the course of repeated chemical analysis which is similar to normal (roughly symmetrical and unimodal) , we often encounter a few apparent outliers which of course can be statistically identified and deleted.   The statistical methods for outliers are Dixon's, Grubb's and many other test statistics.

The use of robust statistics however enables us to circumvent this sometimes contentious issue of the outlier deletion and provides perhaps the best method of identifying them. In fact, robust methods reduce the influence of outlying results and heavy tails in distributions, thus providing statistics that describe the distribution of the central or "good" part of the data collected.

Let's see how the simple MAD method which calculates robust mean and standard deviation without requiring any decisions about rejecting outliers works.

MAD stands for "Median of Absolute Deviations", and is a more robust estimator than sample variance and standard deviation.   Recall that the median is the middle value in an ordered sequence of data. Thus, the median is *unaffected* by extreme values in a set of data.

Suppose we have replicated results as follows:

145	157	183	151	143	147	153	163	130	148

Putting these figures in ascending order, we have:

130	143	145	147	**148**	**151**	153	157	163	183

Therefore the median of these results is the mean of 148 and 151, namely 149.5. We call this median $\hat{\mu} = 149.5$ as the robust estimate of the mean which is unaffected by making the extreme values even more extremes.

Now let's subtract the median of the data from each individual result and ignore the sign of the deviation, giving the absolute deviations:

19.5	6.5	4.5	2.5	1.5	1.5	3.5	7.5	13.5	33.5

If we sort these absolute deviations into ascending order, we have:

1.5	1.5	2.5	3.5	**4.5**	**6.5**	7.5	13.5	19.5	33.5

Hence, the median of these results (the median of absolute deviation MAD) is the mean of 4.5 and 6.5, namely 5.5. Again, we see that this median is not affected by the magnitude of the extreme differences.

In general, we can express MAD as below:

$$\text{MAD} = \text{median}_i\ (\ |\ X_i - \text{median}_j\ (X_j)\ |\ ) \qquad [1]$$

In order to use MAD as the robust estimator of standard deviation, we take:

$$\hat{\sigma} = K.MAD \qquad [2]$$

where K is a constant factor, depending on the types of probability distribution. Use the K factor of 1.4826 which is derived from the properties of the normal distribution. Why?

Since MAD is the middle value of a sequence of results, the probability of finding a value X on either side of mean $\hat{\mu}$ is 50%. We may express this as below:

$$P(|\ X - \hat{\mu}\ | \le MAD) = 0.5$$

By dividing the above equation by standard deviation σ, we have:

$$P\left(\left|\frac{X - \hat{\mu}}{\hat{\sigma}}\right| \le \frac{MAD}{\hat{\sigma}}\right) = 0.5$$

Since $Z = \dfrac{X - \hat{\mu}}{\hat{\sigma}}$, we now have:

$$P\left(|Z| \le \frac{MAD}{\hat{\sigma}}\right) = 0.5$$

If ϕ is the cumulative normal distribution function, we must have that:

$$\phi\left(\frac{MAD}{\hat{\sigma}}\right)-\phi\left(-\frac{MAD}{\hat{\sigma}}\right)=0.5$$

but
$$\phi\left(-\frac{MAD}{\hat{\sigma}}\right)=1-\phi\left(\frac{MAD}{\hat{\sigma}}\right)$$

therefore:
$$\phi\left(\frac{MAD}{\hat{\sigma}}\right)-\phi\left(-\frac{MAD}{\hat{\sigma}}\right)=\phi\left(\frac{MAD}{\hat{\sigma}}\right)-1+\phi\left(\frac{MAD}{\hat{\sigma}}\right)=0.5$$

and hence,
$$2\phi\left(\frac{MAD}{\hat{\sigma}}\right)=1+0.5=1.5$$

or,
$$\phi\left(\frac{MAD}{\hat{\sigma}}\right)=0.75 \qquad [3]$$

or,
$$\frac{MAD}{\hat{\sigma}}=\phi^{-1}(0.75)=0.6745$$

Note: In Excel, the cumulative distribution of $\phi^{-1}(0.75) = 0.6745$ is given by function “=NORM.INV(0.75,0,1)”.

Therefore, Equation [2] gives the constant $K = \{1/\ \phi^{-1}(0.75)\} = 1.4826$, the reciprocal of $\phi^{-1}(0.75)$.

Returning to the example, the robust estimate of the standard deviation, is hence $\hat{\sigma}=5.5\times1.4826=8.2$ (to 2 significant figures). The robust estimates are thus
$$\hat{\mu}=149.5;\hat{\sigma}=8.2$$

In conclusion, the MAD method is quick and simple and has a negligible deleterious effect on the statistics if the dataset does include outliers.