

amc technical brief

Analytical Methods Committee

No.6 Apr 2001

© Royal Society of Chemistry 2001

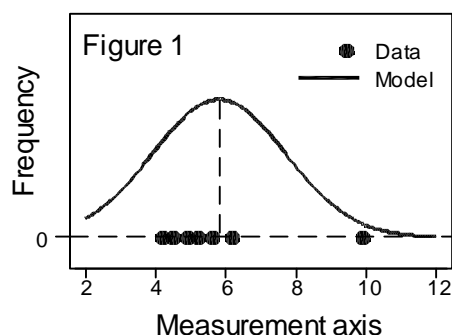
Robust statistics: a method of coping with outliers

Robust statistics is a convenient modern way of summarising results when we suspect that they include a small proportion of outliers. Most estimates of central tendency (e.g., the arithmetic mean) and dispersion (e.g., standard deviation) depend for their interpretation on an implicit assumption that the data comprise a random sample from a normal distribution. But we know that analytical data often depart from that model. They are often heavy tailed (contain a higher than expected proportion of results far from the mean) and sometimes contain outliers.

Let's consider an example data set:

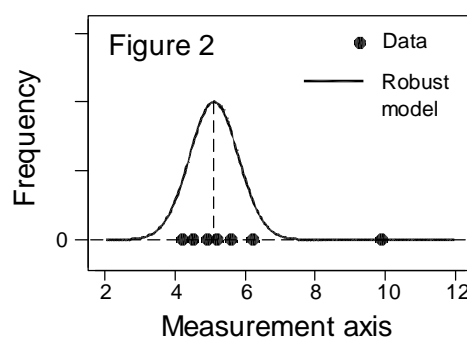
4.5 4.9 5.6 4.2 6.2 5.2 9.9.

The value 9.9 is clearly suspect, even in such a small sample. If we include the suspect value in the calculations, we obtain: $\bar{x} = 5.8$, $s = 1.9$. These statistics, used to define a model based on a normal distribution, describe the data, but not well. The mean seems to have a high bias, while the standard deviation seems too large (Fig 1). Moreover, the numerical values of these estimates, especially that of the standard deviation, are highly dependent on the actual value of the suspect value.



A more reasonable interpretation of these data is that they comprise a random sample from a population with a mean of about 5 and a standard deviation of about 1, with an outlier at 9.9. If we exclude the outlier from the calculations we find $\bar{x} = 5.1$, $s = 0.7$. These statistics provide a plausible normal model for most of the data (Fig

2). Although it provides us with no warning about the possible presence of outliers, this model is often preferable in applications in analytical science.



Outlier tests and robust methods

Typically we handle suspect values by employing tests such as Dixon's test or Grubbs' test to identify them as outliers at particular confidence levels. This procedure is not necessarily straightforward. Firstly, simple versions of the tests may mislead if two or more outliers are present. Secondly, we have to decide whether to exclude the outlier during the calculation of further statistics. This raises the contentious question of when it is justifiable to exclude outliers.

Robust statistics provides an alternative procedure, which provides a model describing the 'good' part of the data, but does not require us to identify specific observations as outliers or exclude them. There are many different robust estimators of mean and standard deviation. First we will look at a very simple method, and then a more sophisticated one.

The median/MAD method

In this method we simply take the central value of the ordered data (the median) as the estimate of the mean.

4.2 4.5 4.9 **5.2** 5.6 6.2 9.9

We notice that the median does not change however much we increase the value of the outlier. The median is a robust estimator of the mean, given by $\hat{m} = 5.2$. (We call

the estimator \hat{m} (pronounced ‘mu-hat’) to distinguish it from \bar{x} , the ordinary arithmetic mean.)

To estimate the standard deviation we first calculate the differences between the values and the median, namely (in the same order):

-1.0 -0.7 -0.3 0.0 0.4 1.0 4.7.

Then we arrange the differences in order of magnitude (*i.e.*, without regard to the sign) and find the median of these values (the median absolute difference, or MAD). This gives:

0.0 0.3 0.4 **0.7** 1.0 1.0 4.7,

and a value of MAD = 0.7. Again we notice that increasing the outlying result has no effect on the value of MAD. We find the robust standard deviation estimate by multiplying the MAD by a factor that happens to have a value close to 1.5. This gives us a robust value (‘sigma-hat’) of $\hat{s} = 1.05$.

If we use this method on data without outliers, it provides estimates that are close to \bar{x} and s , so no harm is done.

Huber’s method

Huber’s method makes more use of the information provided by the data. In this method, we progressively transform the original data by a process called winsorisation.¹ Assume that we have initial estimates called \hat{m}_0, \hat{s}_0 . (These could be evaluated as median-MAD estimates, or simply \bar{x} and s .) If a value x_i falls above $\hat{m}_0 + 1.5\hat{s}_0$ then we change it to $\tilde{x}_i = \hat{m}_0 + 1.5\hat{s}_0$.

Likewise if the value falls below $\hat{m}_0 - 1.5\hat{s}_0$ then we change it to $\tilde{x}_i = \hat{m}_0 - 1.5\hat{s}_0$. Otherwise, we let $\tilde{x}_i = x_i$. We then calculate an improved estimate of mean as $\hat{m}_1 = \text{mean}(\tilde{x}_i)$, and of the standard deviation as

$\hat{s}_1 = 1.134 \times \text{stdev}(\tilde{x}_i)$. (The factor 1.134 is derived from the normal distribution, given a value 1.5 for the multiplier most often used in the winsorisation process.)

Our example data set is somewhat small to subject to winsorisation, but it serves as an illustration of the method. By using $\hat{m}_0 = 5.2, \hat{s}_0 = 1.05$, winsorisation transforms the data set into

4.5 4.9 5.6 4.2 6.2 5.2 **6.775**,

and the improved estimates are $\hat{m}_1 = 5.34, \hat{s}_1 = 1.04$. This procedure is now iterated by using the current improved estimates for the winsorisation at each cycle. Eventually the process converges to an acceptable degree of accuracy, and the resulting values are the robust estimates. For our data we find that $\hat{m}_{Hub} = 5.36, \hat{s}_{Hub} = 1.15$. The

procedure converges slowly, so the method is not suitable for hand calculation. A Minitab implementation of the algorithm is provided in *AMC Software*.

Other robust statistics

More complex types of statistics such as analysis of variance² and regression³ can also be robustified. Robust analysis of variance is particularly useful in analytical science for the interpretation of data from collaborative trials⁴. Robust regression would be useful in calibration, but no analytical studies are yet available.

A cautionary note

Using robust estimates of mean and standard deviation to predict future values from a normal distribution may mislead the unwary because the presence or probability of outliers is not predicted. Employing robust estimates for estimating confidence limits is often useful but the values obtained should be regarded as suggestive only and not for exact interpretation.

When *not* to use robust methods

Robust methods assume that the underlying distribution is roughly normal (and therefore unimodal and symmetrical) but contaminated with outliers and heavy tails. The methods will give misleading results if they are applied to data sets that are markedly skewed or multimodal, or if a large proportion of the data are identical in value.

A final word on outliers

Obtaining a robust statistical model of a data set provides probably the best method for identifying suspect values for further investigation. Taking our example data, we simply transform them by $z = (x - \hat{m}) / \hat{s}$. Using $\hat{m} = 5.36, \hat{s} = 1.15$, we obtain the results:

$z = [-0.7 -0.4 0.2 -1.0 0.7 -0.1 \mathbf{3.9}]$.

Any value greater than about 2.5 could be regarded as suspect, and our candidate outlier is clearly visible.

References

1. AMC, *Analyst*, 1989, **114**, 1489.
2. AMC, *Analyst*, 1989, **114**, 1693.
3. P J Rousseeuw, *J. Chemomet*, 1991, **5**, 1.
4. P J Lowthian, M Thompson, R Wood, *Analyst*, 1998, **123**, 2803.

AMC Technical Briefs are informal but authoritative bulletins on technical matters of interest to the analytical community. Correspondence should be addressed to: The Secretary, The Analytical Methods Committee, The Royal Society of Chemistry, Burlington House, Piccadilly, London W1J 0BA. AMC Technical Briefs may be freely reproduced and distributed in exactly the same form as published here, in print or electronic media, without formal permission from the Royal Society of Chemistry. Copies must not be offered for sale and the copyright notice must not be removed or obscured in any way. Any other reuse of this document, in whole or in part, requires permission in advance from the Royal Society of Chemistry. Other AMC Technical Briefs can be found on: www.rsc.org/lap/rsccom/amc/amc_index.htm