



Missing Values, Outliers, Robust Statistics & Non-parametric Methods

Shaun Burke, RHM Technology Ltd, High Wycombe, Buckinghamshire, UK.

This article, the fourth and final part of our statistics refresher series, looks at how to deal with 'messy' data that contain transcription errors or extreme and skewed results.

This is the last article in a series of short papers introducing basic statistical methods of use in analytical science. In the three previous papers (1–3) we have assumed the data has been 'tidy'; that is, normally distributed with no anomalous and/or missing results. In the real world, however, we often need to deal with 'messy' data, for example data sets that contain transcription errors, unexpected extreme results or are skewed. How we deal with this type of data is the subject of this article.

Transcription errors

Transcription errors can normally be corrected by implementing good quality control procedures before statistical analysis is carried out. For example, the data can be independently checked or, more rarely, the data can be entered, again independently, into two separate files and the files compared electronically to highlight any discrepancies. There are also a number of outlier tests that can be used to highlight anomalous values before other statistics are calculated. These tests do not remove the need for good quality assurance; rather they should be seen as an additional quality check.

Missing data

No matter how well our experiments are planned there will always be times when something goes wrong, resulting in gaps in the data. Some statistical procedures will not work as well, or at all, with some data missing. The best recourse is always to repeat the experiment to generate the complete data set. Sometimes, however, this is not feasible, particularly where

readings are taken at set times or the cost of retesting is prohibitive, so alternative ways of addressing this problem are needed. Current statistical software packages typically deal with missing data by one of three methods:

Casewise deletion excludes all examples (cases) that have missing data in at least one of the selected variables. For example, in ICP–AAS (inductively coupled plasma–atomic absorption spectroscopy) calibrated with a number of standard solutions containing several metal ions at different concentrations, if the aluminium value were missing for a particular test portion, all the results for that test portion would be disregarded (See Table 1). This is the usual way of dealing with missing data, but it does not guarantee correct answers. This is particularly so, in complex (multivariate) data sets where it is possible to end up deleting the majority of your data if the missing data are randomly distributed across cases and variables.

Pairwise deletion can be used as an alternative to casewise deletion in situations where parameters (correlation coefficients, for example) are calculated on successive pairs of variables (e.g., in a recovery experiment we may be interested in the correlations between material recovered and extraction time, temperature, particle size, polarity, etc. With pairwise deletion, if one solvent polarity measurement was missing only this single pair would be deleted from the correlation and the correlations for recovery versus extraction time and particle size would be unaffected) (see Table 2).

Pairwise deletion can, however, lead to serious problems. For example, if there is a 'hidden' systematic distribution of missing points then a bias may result when calculating a correlation matrix (i.e., different correlation coefficients in the matrix can be based on different subsets of cases).

Mean substitution replaces all missing data in a variable by the mean value for that variable. Though this looks as if the

	Al	B	Fe	Ni
Solution 1		94.5	578	23.1
Solution 2	567	72.1	673	7.6
Solution 3		34.0	674	44.7
Solution 4	234	97.4	429	82.9

Casewise deletion. Statistical analysis only carried out on the reduced data set.

	Al	B	Fe	Ni
Solution 2	567	72.1	673	7.6
Solution 4	234	97.4	429	82.9

table 1 Casewise deletion.

data set is now complete, mean substitution has its own disadvantages. The variability in the data set is artificially decreased in direct proportion to the number of missing data points, leading to underestimates of dispersion (the spread of the data). Mean substitution may also considerably change the values of some other statistics, such as linear regression statistics (3), particularly where correlations are strong (See Table 3).

Examples of these three approaches are illustrated in Figure 1, for the calculation of a correlation matrix, where the correlation coefficient (r) (3) is determined for each paired combination of the five variables, A to E. Note, how the r value can increase, diminish or even reverse sign depending on which method is chosen to handle the missing data (i.e., the A, B correlation coefficients).

Extreme values, stragglers and outliers

Extreme values are defined as observations in a sample, so far separated in value from the remainder as to suggest that they may be from a different population, or the result of an error in measurement (6). Extreme values can also be subdivided into stragglers, extreme values detected between the 95% and 99% confidence levels; and outliers, extreme values at > 99% confidence level.

It is tempting to remove extreme values automatically from a data set, because they can alter the calculated statistics, e.g., increase the estimate of variance (a measure of spread), or possibly introduce a bias in the calculated mean. There is one golden rule however: no value should be removed from a data set on statistical grounds alone. 'Statistical grounds' include outlier testing.

Outlier tests tell you, on the basis of some simple assumptions, where you are most likely to have a technical error; they do not tell you that the point is 'wrong'. No matter how extreme a value is in a set of data, the suspect value could nonetheless be a correct piece of information (1). Only with experience or the identification of a particular cause can data be declared 'wrong' and removed.

So, given that we understand that the tests only tell us where to look, how do we test for outliers? If we have good grounds for believing our data is normally distributed then a number of 'outlier tests' (sometimes called Q-tests) are available that identify extreme values in an objective

way (7,8). Good grounds for believing the data is normal are

- past experience of similar data
- passing normality tests, for example, Kolmogorov–Smirnov–Lilliefors test,

Shapiro–Wilk's test, skewness test, kurtosis test (7,9) etc.

- plots of the data, e.g., frequency histogram normal probability plots (1,7). Note that the tests used to check

	Recovery %	Extraction time (mins)	Particle Size (μm)	Solvent Polarity (pKa)
Sample 1	93	20	90	
Sample 2	105	120	150	1.8
Sample 3	99	180	50	1.0
Sample 4	73	10	500	1.5

Pairwise deletion. Statistical analysis unaffected except for when one of a pair of data points are missing.

	Recovery vs Extraction time	Recovery vs Particle Size	Recovery vs Solvent Polarity
r (number of data points in the correlation)	0.728886 (4)	-0.87495 (4)	0.033942 (3)

table 2 Pairwise deletion.

	Al	B	Fe	Ni
Solution 1		94.5	578	23.1
Solution 2	567	72.1	673	7.6
Solution 3		34.0	674	44.7
Solution 4	234	97.4	429	82.9

Mean substitution. Statistical analysis carried out on pseudo completed data with no allowance made for errors in estimated values.

	Al	B	Fe	Ni
Solution 1	400.5	94.5	578	23.1
Solution 2	567	72.1	673	7.6
Solution 3	400.5	34.0	674	44.7
Solution 4	234	97.4	429	82.9

table 3 Mean substitution.

Box 1: Imputation (4,5) is yet another method that is increasingly being used to handle missing data. It is, however, not yet widely available in statistical software packages. In its simplest ad hoc form an imputed value is substituted for the missing value (e.g., mean substitution already discussed above is a form of imputation). In its more general/systematic form, however, the imputed missing values are predicted from patterns in the real (non-missing) data. A total of m possible imputed values are calculated for each missing value (using a suitable statistical model derived from the patterns in the data) and then m possible complete data sets are analysed in turn by the selected statistical method. The m intermediate results are then pooled to yield the final result (statistic) and an estimate of its uncertainty. This method works well providing that the missing data is randomly distributed and the model used to predict the imputed values is sensible.

normality usually require a significant amount of data (a minimum of 10–15 results are recommended depending on the normality test applied). For this reason there will be many examples in analytical science where either it will be impractical to carry out such tests, or the tests will not tell us anything meaningful.

If we are not sure the data set is normally distributed then robust statistics and/or non-parametric (distribution independent) tests can be applied to the data. These three approaches (outlier tests, robust estimates and non-parametric methods) are examined in more detail below.

Outlier tests

In analytical chemistry it is rare that we have large numbers of replicate data, and small data sets often show fortuitous grouping and consequent apparent outliers. Outlier tests should, therefore, be used with care and, of course, identified data points should only be removed if a technical reason can be found for their aberrant behaviour.

Most outlier tests look at some measure of the relative distance of a suspect point from the mean value. This measure is then assessed to see if the extreme value could reasonably be expected to have arisen by chance. Most of the tests look for single extreme values (Figure 2(a)), but sometimes it is possible for several ‘outliers’ to be present in the same data set. These can be identified in one of two ways:

- by iteratively applying the outlier test
- by using tests that look for pairs of extreme values, i.e., outliers that are masking each other (see Figure 2(b) and 2(c)).

Note, as a rule of thumb, if more than 20% of the data are identified as outlying you should start to question your assumption about the data distribution and/or the quality of the data collected.

The appropriate outlier tests for the three situations described in Figure 2 are: 2(a) Grubbs 1, Dixon or Nalimov; 2(b) Grubbs 2 and 2(c) Grubbs 3.

We will concentrate on the three Grubbs’ tests (7). The test values are calculated using the formulae below, after the data are arranged in ascending order.

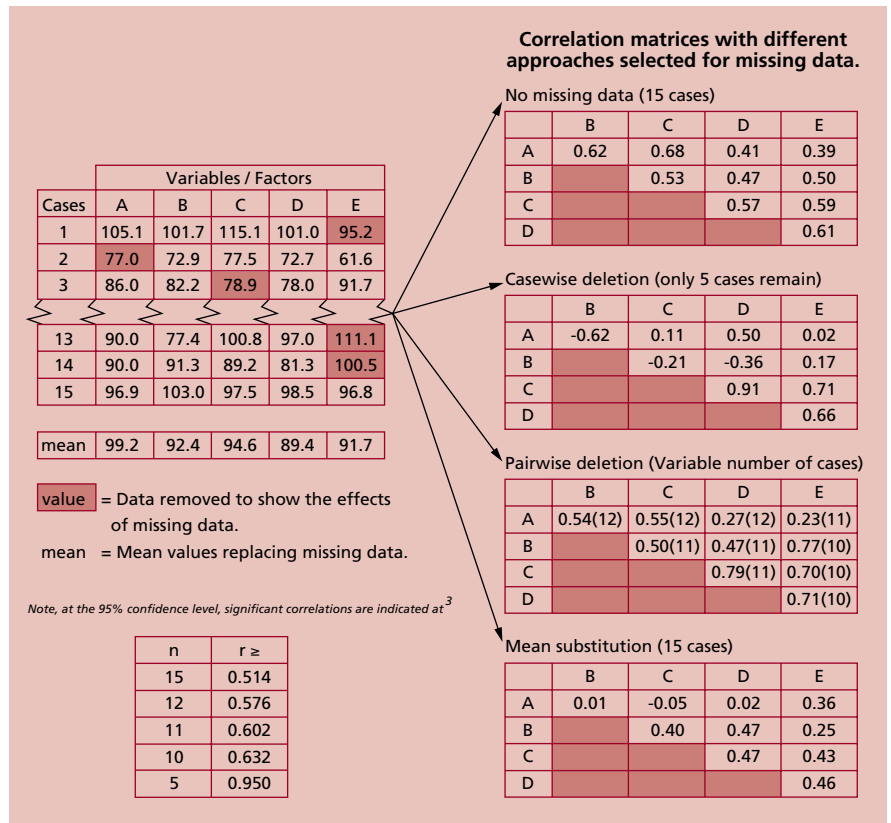


figure 1 Effect of missing data on a correlation matrix.

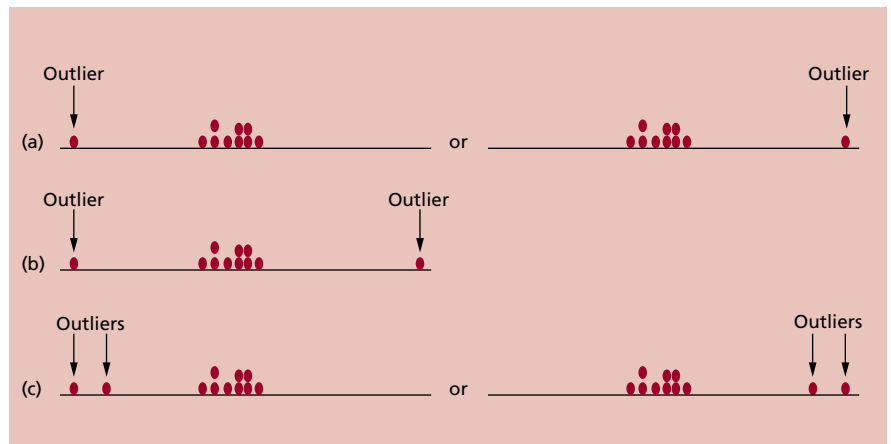


figure 2 Outliers and masking.

$$G_1 = \frac{|\bar{x} - x_j|}{s}$$

$$G_2 = \frac{x_n - x_1}{s}$$

$$G_3 = 1 - \left(\frac{(n-3) \times s_{n-2}^2}{(n-1) \times s^2} \right)$$

where, *s* is the standard deviation for the whole data set, *x_j* is the suspected single outlier, i.e., the value furthest away from the mean, | | is the modulus — the value of a calculation ignoring the sign of the result, \bar{x} is the mean, *n* is the number of data points, *x_n* and *x₁* are the most extreme values, *s_{n-2}* is the standard deviation for the data set

excluding the suspected pair of outlier values, i.e., the pair of values furthest away from the mean.

If the test values (G_1 , G_2 , G_3) are greater than the critical value obtained from tables (see Table 4) then the extreme value(s) are unlikely to have occurred by chance at the stated confidence level (see Box 2).

Pitfalls of outlier tests

Figure 3 shows three situations where outlier tests can misleadingly identify an extreme value.

Figure 3(a) shows a situation common in chemical analysis. Because of limited measurement precision (rounding errors) it is possible to end up comparing a result which, no matter how close it is to the other values, is an infinite number of standard deviations away from the mean of the remaining results. This value will therefore always be flagged as an outlier.

In Figure 3(b) there is a genuine long tail on the distribution that may cause successive outlying points to be identified. This type of distribution is surprisingly common in some types of chemical analysis, e.g., pesticide residues.

If there is very little data (Figure 3(c)) an outlier can be identified by chance. In this situation it is possible that the identified point is closer to the 'true value' and it is the other values that are the outliers. This occurs more often than we would like to admit; how many times do your procedures state 'average the best two out of three determinations'?

Outliers by variance

When the data are from different groups (for example when comparing test methods via interlaboratory comparison) it

is not only possible for individual points within a group to be outlying but also for the group means to have outliers with respect to each other. Another type of 'outlier' that can occur is when the spread of data within one particular group is unusually small or large when compared with the spread of the other groups (see Figure 4).

- The same Grubbs' tests that are used to determine the presence of within group outlying replicates may also be used to test for suspected outlying means.
- The Cochran's test can be used to test for the third case, that of a suspected outlying variance.

To carry out the Cochran's test, the suspect variance is compared with the sum of all group variances. (The variance is a measure of spread and is simply the square of the standard deviation (1).)

$$C_{\bar{n}} = \frac{\text{suspected}(s^2)}{\sum_{i=1}^g S_i^2} \quad \text{where } g \text{ is the number of groups and } \bar{n} = \frac{\sum_{i=1}^g n_i}{g}$$

If this calculated ratio, $C_{\bar{n}}$, exceeds the critical value obtained from statistical tables (7) then the suspect group spread is extreme. The choice of \bar{n} is the average number of all sample results produced by all groups.

The Cochran's test assumes the number of replicates within the groups are the same or at least similar (± 1). It also assumes that none of the data have been rounded and there are sufficient numbers of replicates to get a reasonable estimate of the variance. The Cochran's test should not be used iteratively as this could lead to a large percentage of data being removed (See Box 3).

Robust statistics

Robust statistics include methods that are largely unaffected by the presence of extreme values. The most commonly used of these statistics are as follows:

Median: The median is a measure of central tendency¹ and can be used instead of the mean. To calculate the median (\tilde{x}) the data are arranged in order of magnitude and the median is then the central member of the series (or the mean of the two central members when there is an even number of data, i.e., there are equal numbers of observations smaller and greater than the median). For a symmetrical distribution the mean and median have the same value.

$$\tilde{x} = \begin{cases} x_m & \text{when } n \text{ is odd } 1, 3, 5, \dots \\ \frac{x_m + x_{m+1}}{2} & \text{when } n \text{ is even } 2, 4, 6, \dots \end{cases} \quad \text{where } m = \text{round up} \left(\frac{n}{2} \right)$$

Median Absolute Deviation (MAD): The MAD value is an estimate of the spread in the data similar to the standard deviation.

Box 2: Grubbs' tests (worked example).

13 replicates are ordered in ascending order.

$$x_1 \quad 47.876 \quad 47.997 \quad 48.065 \quad 48.118 \quad 48.151 \quad 48.211 \quad 48.251 \quad 48.559 \quad 48.634 \quad 48.711 \quad 49.005 \quad 49.166 \quad 49.484 \quad x_n$$

$n = 13$, mean = 48.479, $s = 0.498$, $s_{n-2}^2 = 0.123$

$$G_1 = \frac{49.484 - 48.479}{0.498} = 2.02$$

$$G_2 = \frac{49.484 - 47.876}{0.498} = 3.23$$

$$G_3 = 1 - \left(\frac{10 \times 0.123}{12 \times 0.498^2} \right) = 0.587$$

Grubbs' critical values for 13 values are $G_1 = 2.331$ and 2.607 , $G_2 = 4.00$ and 4.24 , $G_3 = 0.6705$ and 0.7667 for the 95% and 99% confidence levels. Since the test values are less than their respective critical values, in all cases, it can be concluded there are no outlying values.

For n values $MAD = \text{median}(|x_i - \bar{x}|_{i=1, 2, \dots, n})$

If the MAD value is scaled by a factor of 1.483 it becomes comparable with a standard deviation, this is the MAD_E value.

$$MAD_E = 1.483 \times MAD$$

Other robust statistical estimates include trimmed mean and deviations, Winsorized mean and deviation, least median of squares (robust regression), Levene's test (heterogeneity in ANOVA), etc. A discussion of robust statistics in analytical chemistry can be found elsewhere (10, 11).

level n	95% confidence level			99% confidence		
	G(1)	G(2)	G(3)	G(1)	G(2)	G(3)
3	1.153	2.00	---	1.155	2.00	---
4	1.463	2.43	0.9992	1.492	2.44	1.0000
5	1.672	2.75	0.9817	1.749	2.80	0.9965
6	1.822	3.01	0.9436	1.944	3.10	0.9814
7	1.938	3.22	0.8980	2.097	3.34	0.9560
8	2.032	3.40	0.8522	2.221	3.54	0.9250
9	2.110	3.55	0.8091	2.323	3.72	0.8918
10	2.176	3.68	0.7695	2.410	3.88	0.8586
12	2.285	3.91	0.7004	2.550	4.13	0.7957
13	2.331	4.00	0.6705	2.607	4.24	0.7667
15	2.409	4.17	0.6182	2.705	4.43	0.7141
20	2.557	4.49	0.5196	2.884	4.79	0.6091
25	2.663	4.73	0.4505	3.009	5.03	0.5320
30	2.745	4.89	0.3992	3.103	5.19	0.4732
35	2.811	5.026	0.3595	3.178	5.326	0.4270
40	2.866	5.150	0.3276	3.240	5.450	0.3896
50	2.956	5.350	0.2797	3.336	5.650	0.3328
60	3.025	5.500	0.2450	3.411	5.800	0.2914
70	3.082	5.638	0.2187	3.471	5.938	0.2599
80	3.130	5.730	0.1979	3.521	6.030	0.2350
90	3.171	5.820	0.1810	3.563	6.120	0.2147
100	3.207	5.900	0.1671	3.600	6.200	0.1980
110	3.239	5.968	0.1553	3.632	6.268	0.1838
120	3.267	6.030	0.1452	3.662	6.330	0.1716
130	3.294	6.086	0.1364	3.688	6.386	0.1611
140	3.318	6.137	0.1288	3.712	6.437	0.1519

table 4 Grubbs' critical value table (5).

Non-parametric tests

Typical statistical tests incorporate assumptions about the underlying distribution of data (such as normality), and hence rely on distribution parameters. 'Non-parametric' tests are so called because they make few or no assumptions about the distributions, and do not rely on distribution parameters. Their chief advantage is improved reliability when the distribution is unknown. There is at least one non-parametric equivalent for each parametric type of test (see Table 5). In a short article, such as this, it is impossible to describe the methodology for all these tests but more information can be found in other publications (12, 13).

Conclusions

- Always check your data for transcription errors. Outlier tests can help to identify them as part of a quality control check.
- Delete extreme values only when a technical reason for their aberrant behaviour can be found.
- Missing data can result in misinterpretation of the resulting statistics so care should be taken with the method chosen to handle the gaps. If at all possible, further experiments should be carried out to fill in the missing points.

Box 3: Cochran's test (worked example).

An interlaboratory study was carried out by 13 laboratories to determine the amount of cotton in a cotton/polyester fabric, 85 determinations were carried out in total. The standard deviations of the data obtained by each of the 13 laboratories was as follows:

Std. Dev. 0.202 0.402 0.332 0.236 0.318 0.452 0.210 0.074 0.525 0.067 0.609 0.246 0.198

$$\bar{n} = \frac{85}{13} = 6.54 \approx 7 \quad C_{\bar{n}} = \frac{0.609^2}{0.202^2 + 0.402^2 + \dots + 0.246^2 + 0.198^2} = \frac{0.371}{1.474} = 0.252$$

Cochran's critical value for $\bar{n} = 7$ and $g = 13$ is 0.23 at the 95% confidence levels⁷.

As the test value is greater than the critical values it can be concluded that the laboratory with the highest standard deviation (0.609) has an outlying spread of replicates and this laboratory's results therefore need to be investigated further. It is normal practice in inter-laboratory comparisons not to test for low variance outliers, i.e., laboratories reporting unusually precise results.

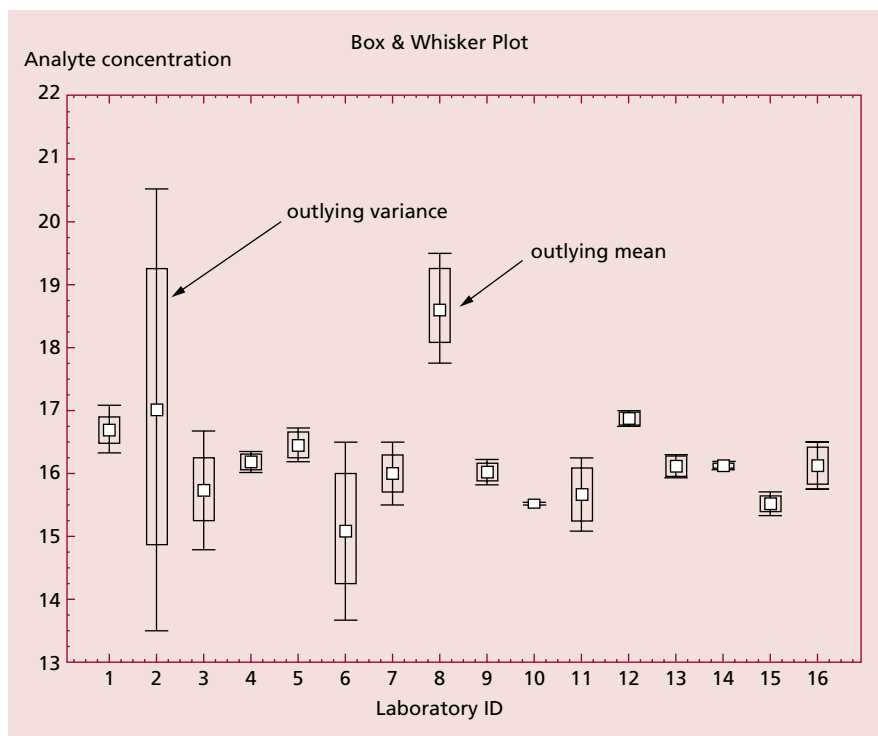


figure 4 Different types of outlier in grouped data.

Types of comparison	Parametric methods	Non-parametric methods (12, 13)
Differences between independent groups of data	t-test for independent groups ²	Wald–Wolfowitz runs test Mann–Whitney U test Kolmogorov–Smirnov two-sample test
	(ANOVA/MANOVA) ²	Kruskal–Wallis analysis of ranks. Median test
Differences between dependent groups of data	t-test for dependent groups ²	Sign test Wilcoxon’s matched pairs test McNemar’s test χ^2 (Chi-square) test
	ANOVA with replication ²	Friedman’s two-way ANOVA Cochran Q test
Relationships between continuous variables	Linear regression ³ Correlation coefficient ³	Spearman R Kendall Tau
Homogeneity of Variance	Bartlett’s test ⁷	Levene’s test, Brown & Forsythe
Relationships between counted variables		coefficient Gamma χ^2 (Chi-square) test Phi coefficient Fisher exact test Kendall coefficient of concordance

table 5 Non-parametric alternatives to parametric statistical tests.

- Outlier tests assume the data distribution is known. This assumption should be checked for validity before these tests are applied.
- Robust statistics avoid the need to use outlier tests by down-weighting the effect of extreme values.
- When knowledge about the underlying data distribution is limited, non-parametric methods should be used.

NB: It should be noted that following a judgement in a US court, the Food and Drug Administration (FDA) in a guide — Guide to inspection of pharmaceutical quality control laboratories — has specifically prohibited the use of outlier tests.

Acknowledgement

The preparation of this paper was supported under a contract with the UK’s Department of Trade and Industry as part of the National Measurement System Valid Analytical Measurement Programme (VAM) (14).

References

- (1) S. Burke, *Scientific Data Management* 1(1), 32–38, 1997.
- (2) S. Burke, *Scientific Data Management* 2(1), 36–41, 1998.
- (3) S. Burke, *Scientific Data Management* 2(2), 32–40, 1998.
- (4) J.L. Schafer, *Monographs on Statistics and Applied Probability* 72 — *Analysis of Incomplete Multivariate Data*, Chapman & Hall (1997) ISBN 0-412-04061-1.
- (5) R.J.A. Little & D.B. Rubin, *Statistical Analysis With Missing Data*, John Wiley & Sons (1987), ISBN 0-471-80243-9.
- (6) ISO 3534. Statistics — Vocabulary and Symbols. Part 1: Probability and general statistical terms, section 2.64. Geneva 1993.
- (7) T.J. Farrant, *Practical statistics for the analytical scientist: A bench guide*, Royal Society of Chemistry 1997. (ISBN 0 85404 442 6).
- (8) V. Barret & T. Lewis, *Outliers in Statistical Data*, 3rd Edition, John Wiley (1994).
- (9) William H. Kruskal & Judith M. Tanur, *International Encyclopaedia of Statistics*, Collier Macmillan Publishers, 1978. ISBN 0-02-917960-2.
- (10) Analytical Methods Committee, Robust Statistics — How Not to Reject Outliers Part 2. *Analyst* 1989 114, 1693–7.
- (11) D.C. Hoaglin, F. Mosteller & J.W. Tukey, *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons (1983), ISBN 0-471-09777-2.
- (12) M. Hollander & D.A. Wolf, *Non-parametric statistical methods*, Wiley & Sons, New York 1973.
- (13) W.W. Daniel, *Applied non-parametric statistics*, Houghton Mifflin, Boston 1978.
- (14) M. Sargent, *VAM Bulletin*, Issue 13, 4–5, Autumn. Laboratory of the Government Chemist, 1995.