Critical values for 22 discordancy test variants for outliers in normal samples up to sizes 100, and applications in science and engineering

Surendra P. Verma[*] and Alfredo Quiroz-Ruiz

Centro de Investigación en Energía, Universidad Nacional Autónoma de México,
Priv. Xochicalco s/no., Col Centro, Apartado Postal 34, 62580 Temixco, Morelos, Mexico
** spv@cie.unam.mx*

ABSTRACT

In this paper, the modifications of the simulation procedure as well as new, precise, and accurate critical values or percentage points (for the majority of data with four decimal places; respective average standard error of the mean ~0.0001–0.0025) of nine discordancy tests, with 22 test variants, and each with seven significance levels α = 0.30, 0.20, 0.10, 0.05, 0.02, 0.01, and 0.005, for normal samples of sizes n up to 100 are reported. Prior to our work, only less precise critical values were available for most of these tests, viz., with one (for n <20) and three decimal places (for greater n) for test N14; two decimal places for tests N2, N3–k=2,3,4, N6, and N15; and three decimal places for N1, N4–k=3,4, N5, and N8; but all of them with unknown errors. In fact, the critical values were available for n only up to 20 for test N2, up to 30 for test N8, and up to 50 for N4–k=1,3,4, whereas for most other tests, in spite of the availability for n up to 100 (or more), interpolations were required because tabulated values were not reported for all n in the range 3–100. Therefore, the applicability of these discordancy tests is now extended up to 100 observations of a particular parameter in a statistical sample, without any need of interpolations. The new more precise and accurate critical values will result in a more reliable application of these discordancy tests than has so far been possible. Thus, we envision that these new critical values will result in wider applications of these tests in a variety of scientific and engineering fields such as agriculture, astronomy, biology, biomedicine, biotechnology, chemistry, electronics, environmental and pollution research, food science and technology, geochemistry, geochronology, isotope geology, meteorology, nuclear science, paleontology, petroleum research, quality assurance and assessment programs, soil science, structural geology, water research, and zoology. The multiple-test method with new critical values proposed in this work was shown to perform better than the box-and-whisker plot method used by some researchers. Finally, the so-called "two standard deviation" method frequently used for processing inter-laboratory databases was shown to be statistically-erroneous, and should therefore be abandoned. Instead, the multiple-test method with 15 tests and 33 test variants, all of which now readily applicable to sample sizes up to 100, should be used. To process inter-laboratory databases, our present approach of multiple-test method is also shown to perform better than the "two standard deviation" method.

Keywords: outlier methods, normal sample, two standard deviation method, 2s, reference materials, Monte Carlo simulations, critical value tables, Dixon Q-test, skewness, kurtosis, petroleum hydrocarbon.

RESUMEN

En este trabajo se reportan las modificaciones del procedimiento de la simulación así como valores críticos o puntos porcentuales nuevos y más precisos y exactos (para la mayoría de los datos con cuatro puntos decimales; el error estándar de la media ~0.0001–0.0025) para nueve pruebas de discordancia con 22 variantes, y cada una con siete niveles de significancia α = 0.30, 0.20, 0.10, 0.05, 0.02, 0.01 y 0.005, para muestras normales con tamaño n hasta 100. Antes de nuestro trabajo, solamente se disponía de valores críticos menos precisos para la mayoría de estas pruebas, viz., con uno (para n <20) y tres puntos decimales (para n mayores) para la prueba N14, dos puntos decimales para las pruebas N2, N3–k=2,3,4, N6 y N15, y tres puntos decimales para N1, N4–k=3,4, N5 y N8, pero todos ellos con *errores desconocidos. En realidad se disponía de los valores críticos solamente para n hasta 20 para la prueba N2, hasta 30 para la prueba N8, y hasta 50 para N4–k=1,3,4, mientras que para muchas otras pruebas, a pesar de la disponibilidad para hasta n 100 (o más) se requería de interpolaciones, dado que los valores tabulados no fueron reportados para todos los n en el intervalo de 3 a100. Por consecuencia, la aplicabilidad de las pruebas de discordancia es extendida hasta 100 observaciones de un determinado parámetro en una muestra estadística, sin necesidad de realizar las interpolaciones de los valores críticos. Los valores críticos nuevos y más precisos y exactos resultarán en una aplicación más confiable de las pruebas de discordancia que ha sido posible hasta ahora. De esta manera, consideramos que estos nuevos valores críticos resultarán en aplicaciones más amplias de estas pruebas en una variedad de campos de conocimiento científico y de ingenierías, tales como agricultura, astronomía, biología, biomedicina, biotecnología, ciencia del suelo, ciencia nuclear, ciencia y tecnología de los alimentos, contaminación ambiental, electrónica, geocronología, geología estructural, geología isotópica, geoquímica, investigación del agua, investigación del petróleo, meteorología, paleontología, programas de aseguramiento de calidad, química y zoología. El método de pruebas múltiples con nuevos valores críticos propuesto aquí proporciona mejores resultados que el método de la gráfica de "box y whisker" usado por algunos investigadores. Finalmente, se demostró que el así llamado método de "dos desviaciones estándar", frecuentemente usado para procesar las bases de datos interlaboratorios, es erróneo y, por lo tanto, debe ser abandonado. En su lugar debe usarse el método de pruebas múltiples con 15 pruebas y 33 variantes, todas ellas ahora rápidamente aplicables para los tamaños de muestras hasta 100. Nuestro procedimiento de pruebas múltiples parece funcionar mejor que el método de "dos desviaciones estándar" para el procesamiento de datos geoquímicos provenientes de muchos laboratorios.*

Palabras clave: métodos de valores desviados, muestra normal, prueba de dos desviaciones estándar, 2s, materiales de referencia, simulaciones Monte Carlo, tablas de valores críticos, prueba Q de Dixon, sesgo, curtosis, hidrocarburos de petróleo.

INTRODUCTION

In a recent paper (Verma and Quiroz-Ruiz, 2006), we covered the following points: (1) explained the need of new critical values or percentage points of statistical tests for normal univariate samples; (2) developed and reported a highly precise and accurate Monte Carlo type simulation procedure for N(0,1) random normal variates; (3) presented new, precise, and accurate critical values for seven significance levels α = 0.30, 0.20, 0.10, 0.05, 0.02, 0.01, and 0.005, and for sample sizes *n* up to 100 for six Dixon discordancy tests (with 11 test variants) for normal univariate samples; and (4) highlighted the use of these new values in very diverse fields of science and engineering, including the Earth Sciences. We had included all six frequently used discordancy tests (N7 and N9-N13; see pp. 218-236 of Barnett and Lewis, 1994), initially proposed by Dixon (1950, 1951, 1953), for simulating new, precise, and accurate critical values for *n* up to 100 (number of data in a given statistical sample, *n* = 3(1)100 for test N7, *i.e.*, for all values of *n* between 3 and 100; *n* = 4(1)100 for tests N9 and N11; *n* = 5(1)100 for tests N10 and N12; and *n* = 6(1)100 for test N13).

It is pertinent to mention that researchers (*e.g.*, Dybczyński *et al.*, 1979; Dybczyński, 1980; Barnett and Lewis, 1994; Verma, 1997, 1998, 2005; Verma *et al.*, 1998; Velasco *et al.*, 2000; Guevara *et al.*, 2001; Velasco-Tapia *et al.*, 2001) have recommended that most, if not all the available discordancy tests should be applied to a given data set to detect discordant outliers. Because the critical values for the six Dixon tests have now been significantly improved and extended (Verma and Quiroz-Ruiz, 2006), there is still the need for improving the existing critical values and simulating new ones for the remaining tests for normal samples listed by Barnett and Lewis (1994).

It is also important to note that we are dealing with tests for normal univariate samples, *i.e.*, the statistical sample is assumed to be drawn from a normal population. Obviously, different sets of critical values would be required for other types of distributions such as exponential or Poisson distributions (see Barnett and Lewis, 1994, or Zhang, 1998, for more details).

In the present work, for simulating new, precise, and accurate critical values for the same seven significance levels (α = 0.30 to 0.005) and for *n* up to 100, we have included most of the remaining tests for normal univariate samples (nine tests with 22 test variants): N1 (upper or lower version), N2 (two-sided), N3–k=2,3,4 (upper or lower), N4–k=1,2,3,4 (upper or lower), N5–k=2 (upper-lower pair), N6–k=2 (upper-lower pair), N8 (two-sided; also known as Dixon Q-test), N14 (skewness or third moment test), and N15 (kurtosis or fourth moment test); see pp. 218-236 of Barnett and Lewis (1994), or pp. 89-97 of Verma (2005). The new critical values are compared with the literature values and are shown to be more precise and accurate, enabling thus statistically more reliable applications in many science and engineering fields. We also present a few examples for the application of all normal univariate tests for which we have reported new critical values in this as well as in our earlier paper (Verma and Quiroz-Ruiz, 2006).

DISCORDANCY TESTS

We will not repeat the explanation of discordancy tests; the reader is referred to Barnett and Lewis (1994), Verma (2005), or our earlier paper (Verma and Quiroz-Ruiz, 2006). The nine tests with their 22 variants for which critical values were simulated are listed in Table 1. We note that critical values were available in the literature only for *n* up to 20 for test N2, up to 30 for N8, and up to 50 for N4–k=1,3,4, whereas for most other tests, in spite of the availability for *n* up to 100 (or more), interpolations were occasionally required because tabulated values were not reported for all *n* (see Barnett and Lewis, 1994, or tables A4 to A18 in Verma, 2005).

Tests N1 and N2

We will briefly describe discordancy tests N1 and N2, which are, respectively, the upper (or lower) and extreme outlier tests in a normal sample with both population mean (μ) and population variance (σ^2) unknown (Barnett and Lewis, 1994), because their statistically "erroneous" version has been used as a popular, so-called "two standard deviation" method by numerous workers (*e.g.*, Stoch and Steele, 1978; Ando *et al.*, 1987, 1989; Gladney and Roelandts, 1988, 1990; Gladney *et al.*, 1991, 1992; Itoh *et al.*, 1993; Imai *et al.*, 1995, 1996) to process inter-laboratory data for international geochemical reference materials.

The test statistic of N1 for upper or lower outlier is, respectively:

$$TN1_{(u)} = \frac{(x_{(n)} - \bar{x})}{s} \qquad (1)$$

or

$$TN1_{(l)} = \frac{(\bar{x} - x_{(1)})}{s} \qquad (2)$$

whereas that of N2 is

$$TN2 = Max : \left[\frac{x_{(n)} - \bar{x}}{s}, \frac{\bar{x} - x_{(1)}}{s} \right] \qquad (3)$$

Here, for an ordered array $x_{(1)}$, $x_{(2)}$, $x_{(3)}$,… $x_{(n\text{-}2)}$, $x_{(n\text{-}1)}$, $x_{(n)}$ of *n* observations $x_{(1)}$ is the lowest observation and $x_{(n)}$ the highest one; $\bar{x}$ is the sample mean; and *s* is the sample standard deviation.

Two standard deviation method: a statistically erroneous and outdated version of tests N1 and N2

For the standard normal distribution, the total probability of observations at a distance greater than 1.96σ (*i.e.*, about 2σ) from the mean μ, *i.e.*, outside the ($\mu \pm 1.96\sigma$) is 0.05, in other words, about 95% of the area of the density curve is contained within this range (*e.g.*, Otto, 1999). These considerations of the normal density curve have been used by some workers to fix the limit of 2σ, or more appropriately 2*s* (because the population parameters μ and σ are not known for most experimental data) for testing samples of finite size for discordant outliers, *i.e.*, observations falling within ($\bar{x} \pm 2s$) are retained and those outside this range are rejected, irrespective of the actual value of *n* (where *n*, $\bar{x}$, and *s* are, respectively, the total number of samples, location parameter – sample mean, and scale parameter – sample standard deviation).

This kind of outlier test belongs to a group of old, outdated test procedures (pre-1925!) characterized by two general defects (see for more details pages 30-31, 108-116, and 222-223 in Barnett and Lewis, 1994): they fail to distinguish between population variance (σ) and sample variance (*s*), and, more importantly, they are erroneously based on the distributional behavior of a random sample value rather than on an appropriate sample *extreme* value. Barnett and Lewis (1994, p. 31) go on stating that even a more serious shortcoming of such an outdated procedure is "the failure to recognize that *it is an extreme* $x_{(1)}$ *or* $x_{(n)}$ which (by the very nature of outlier study) *should figure in the test statistic, rather than an arbitrary sample value* x_j". This shortcoming is certainly overcome by tests N1 or N2 above (see equations 1 to 3), in which an extreme value ($x_{(1)}$ or $x_{(n)}$) is tested by the respective test statistic. Thus, an *ad hoc* procedure of a "fixed multiple of standard deviation" leads to rejection of any observation x_j for which ($|x_j - \bar{x}|/s$) is sufficiently large, in fact, >2 for the 2*s* method or >3 for the 3*s* method –but with no regard to the effect of sample size *n* on the distribution form of the statistic. We will further comment on the shortcomings of this procedure after we have presented the new critical values for tests N1 and N2.

Masking and swamping effects and different types of test statistics

We will briefly point out the reasons for our recommendations to apply all outlier tests to a given data set instead of only a few of them (see Verma, 1997, 1998; Verma *et al.*, 1998 for more details).

One problem in testing for a single outlier in a normally distributed sample is the sensitivity to the phenomenon of *masking* (Bendre and Kale, 1987; Barnett and Lewis, 1994). A discordancy test of the most extreme observation (*e.g.*, x_n) is rendered insensitive by the proximity of the next most extreme observation ($x_{n\text{-}1}$), in which case the presence of the latter would have masked the first. Dixon test N7 is especially susceptible to such masking effects (see Verma and Quiroz-Ruiz, 2006, for the test statistic *TN*7), although the N1 statistic (Grubbs, 1950) is probably not too much better. One solution to this problem is to use test statistics that are less sensitive to masking. There are a number of Dixon-like statistics for this purpose (N11-N13; Barnett and Lewis, 1994; Verma and Quiroz-Ruiz, 2006). In the case of test N12, for example, the numerator is the difference between the outlier (x_n or x_1) being tested and its second-nearest neighbor (x_{n-2} or x_3). No masking effect is observed by the measurement value x_{n-1} or x_2.

Outlier masking occurs because there are actually two (or more) outliers, and some statistics, such as N1, work best when testing data sets with a single outlier (*e.g.*, Prescott, 1978). If a data set contains more than one outlier, it is necessary to modify the statistical approach, considering the next outlier assemblages: (1) two or more upper outliers, (2) two or more lower outliers, and (3) a combination of one (or more) upper outlier(s) and one or more lower outlier(s). Caution is, however, required if one is dealing with chemical data obtained from different analytical techniques, in which case other statistical tests, such as F-test, Student t, or ANOVA, should, in fact, be applied prior to the application of discordancy tests (see, *e.g.*, Verma, 1998, 2005). This topic will be dealt with in more detail in a separate paper.

In general, two testing approaches have been applied for these cases (Barnett and Lewis, 1994): (1) the consecutive testing approach, where a test statistic such as N1 or N7 is applied repeatedly to a data set (one outlier at a time); or (2) the block testing approach, where a statistic simultaneously tests k (= 2, 3, 4, or more) observations in the data set.

In the consecutive testing, the most extreme outlier is evaluated; if it gives a positive test result, *i.e.*, if this outlier is declared to be discordant, it is removed from the data set, and then the most extreme remaining outlier is tested. This procedure is repeated until all the outliers are tested, or until an outlier gives a negative test result, *i.e.*, it is not discordant. However, the disadvantage is the susceptibility of this procedure to masking effects. Tests N1 and N7 are certainly poor candidates for this type of testing procedure although tests N14 and N15 (high-order moment tests) should be applied consecutively when more than one outlier is present in a statistical sample (see Barnett and Lewis, 1994). Test N14 was recommended for a one-sided detection, although Iglewicz and Martinez (1982) reported that this test is greatly affected by masking effect, and thus should not be used when more than one outlier is suspected. Both tests N14 and N15 are used for testing an extreme value (two-sided tests). The poor efficiency of these consecutive tests based on the use of standard deviation (s), such as N1 or N2, can be adduced to the fact that value of s is greatly influenced by the presence of discordant outliers. As a consequence, the presence of several outliers may cause a sufficiently large increase in standard deviation, with the result that no outliers are detected (Barnett and Lewis, 1994).

An alternative to consecutive testing is block testing, where a statistic is used to test all k outliers at once (*e.g.*, test N3 or N4; Table 1). Also, some differences in performance have been reported for block-testing statistics (Prescott, 1978; Hayes and Kinsella, 2003). For example, McMillan (1971) suggested that test N4–k=2 is more robust than test N3–k=2. It is important to point out that, assuming that all outliers have been identified, statistics intended for block tests are not susceptible to outlier masking. However, the application of this procedure could generate another problem, known as outlier *swamping* (Barnett and Lewis, 1994). A block test could be insensitive if the second observation x_{n-1} is close to the next neighbor (x_{n-2}) and is not outlying or discordant. The pair x_n, x_{n-1} might not reach discordancy level when tested jointly, even though x_n on its own is discordant in relation to the other n-1 observations (*i.e.*, swamping effect of x_{n-1}). In block testing, all k outliers are labeled as contaminants. What this means in practice is that contaminants belong to a different probability distribution than the rest of the data or they are accepted as "normal" deviated measurements drawn from a different normal distribution. There is no middle ground, as there is in consecutive testing, where it is possible to establish when some outliers are contaminants and when some are not. Thus, there is the possibility that a marginal outlier might be falsely declared a contaminant because it is "carried along" in the block testing procedure by others, more extreme, outliers. Or perhaps a few marginal outliers cause the block testing procedure to fail, which means that the contaminants that are in the block will not be identified. However, some procedures have been proposed to establish k (*i.e.*, the number of contaminants in the sample), free from the masking and swamping effects, when testing upper or lower outliers (Zhang and Wang, 1998). The block testing procedure can also be applied consecutively if more that k outliers are suspected in a given data set.

In summary, more work is needed to better understand the relative merits and usefulness of different test procedures. As pointed out in our earlier paper (Verma and Quiroz-Ruiz, 2006), these evaluations can also be performed empirically or through computer simulations, which is planned to be carried out in a future study. From the practical point of view, however, we can evaluate an experimental data set, using both types of outlier tests –single-outlier (consecutive procedure) as well as multiple-outlier (block procedure) tests– termed here as the "multiple-test" method.

SIMULATION PROCEDURE FOR MORE PRECISE AND ACCURATE CRITICAL VALUES

Our highly precise and accurate Monte Carlo type simulation procedure has already been described in detail (Verma and Quiroz-Ruiz, 2006) and, therefore, will not be repeated here. However, some required changes will be mentioned.

As in our earlier work, our simulations were of sizes 100,000, and were repeated 10 times (each using a different set of 10,000,000 random normal variates) for obtaining the final mean critical value or percentage point and its standard error. However, for test N1, two independent test statistics (one for upper and the other for lower outlier) were simulated and thus 20 independent results could be obtained from

Table 1. Discordance tests for univariate normal samples (modified after Barnett and Lewis, 1994; Verma, 1997, 2005).

Test code [a]	**Value(s) tested**		**Test statistic**	**Test significance**	**Applicability of test** $n_{\min} - n_{\max}$	
					Literature (less precise values)	This work (more precise values)
N1	Upper	$x_{(n)}$	$TN1_{(u)} = (x_{(n)} - \bar{x})/s$	Greater	3 – 147	3 – 100
	Lower	$x_{(1)}$	$TN1_{(l)} = (\bar{x} - x_{(1)})/s$	Greater	3 – 147	3 – 100
N2 (two-sided)	Extreme	$x_{(n)}$ or $x_{(1)}$	$TN2 = Max\colon\{(x_{(n)} - \bar{x})/s,\ (\bar{x} - x_{(1)})/s\}$	Greater	3 – 20	3 – 100
N3	k=2 Upper	$x_{(n)}$, $x_{(n\text{-}1)}$	$TN3_{(2u)} = (x_{(n)} + x_{(n\text{-}1)} - 2\bar{x})/s$	Greater	5 – 100	5 – 100
	k=3 Upper	$x_{(n)}$, $x_{(n\text{-}1)}$, $x_{(n\text{-}2)}$	$TN3_{(3u)} = (x_{(n)} + x_{(n\text{-}1)} + x_{(n\text{-}2)} - 3\bar{x})/s$	Greater	7 – 100	7 – 100
	k=4 Upper	$x_{(n)}$, $x_{(n\text{-}1)}$, $x_{(n\text{-}2)}$, $x_{(n\text{-}3)}$	$TN3_{(4u)} = (x_{(n)} + x_{(n\text{-}1)} + x_{(n\text{-}2)} + x_{(n\text{-}3)} - 4\bar{x})/s$	Greater	9 – 100	9 – 100
	k=2 Lower	$x_{(1)}$, $x_{(2)}$	$TN3_{(2l)} = (2\bar{x} - x_{(1)} - x_{(2)})/s$	Greater	5 – 100	5 – 100
	k=3 Lower	$x_{(1)}$, $x_{(2)}$, $x_{(3)}$	$TN3_{(3l)} = (3\bar{x} - x_{(1)} - x_{(2)} - x_{(3)})/s$	Greater	7 – 100	7 – 100
	k=4 Lower	$x_{(1)}$, $x_{(2)}$, $x_{(3)}$, $x_{(4)}$	$TN3_{(4l)} = (3\bar{x} - x_{(1)} - x_{(2)} - x_{(3)} - x_{(4)})/s$	Greater	9 – 100	9 – 100
N4	k=1 Upper	$x_{(n)}$	$TN4_{(1u)} = S^{2}_{(n)} / S^{2}$	Smaller	3 – 50	3 – 100
	k=2 Upper	$x_{(n)}$, $x_{(n\text{-}1)}$	$TN4_{(2u)} = S^{2}_{(n),(n\text{-}1)} / S^{2}$	Smaller	4 – 149	4 – 100
	k=3 Upper	$x_{(n)}$, $x_{(n\text{-}1)}$, $x_{(n\text{-}2)}$	$TN4_{(3u)} = S^{2}_{(n),(n\text{-}1),(n\text{-}2)} / S^{2}$	Smaller	6 – 50	6 – 100
	k=4 Upper	$x_{(n)}$, $x_{(n\text{-}1)}$, $x_{(n\text{-}2)}$, $x_{(n\text{-}3)}$	$TN4_{(4u)} = S^{2}_{(n),(n\text{-}1),(n\text{-}2),(n\text{-}3)} / S^{2}$	Smaller	8 – 50	8 – 100
	k=1 Lower	$x_{(1)}$	$TN4_{(1l)} = S^{2}_{(1)} / S^{2}$	Smaller	3 – 50	3 – 100
	k=2 Lower	$x_{(1)}$, $x_{(2)}$	$TN4_{(2l)} = S^{2}_{(1),(2)} / S^{2}$	Smaller	4 – 149	4 – 100
	k=3 Lower	$x_{(1)}$, $x_{(2)}$, $x_{(3)}$	$TN4_{(3l)} = S^{2}_{(1),(2),(3)} / S^{2}$	Smaller	6 – 50	6 – 100
	k=4 Lower	$x_{(1)}$, $x_{(2)}$, $x_{(3)}$, $x_{(4)}$	$TN4_{(4l)} = S^{2}_{(1),(2),(3),(4)} / S^{2}$	Smaller	8 – 50	8 – 100
N5	k=2 Upper– lower	$x_{(n)}$, $x_{(1)}$	$TN5_{(ul)} = S^{2}_{(n),(1)} / S^{2}$	Smaller	4 – 100	4 – 100
N6	k=2 Upper– lower	$x_{(n)}$, $x_{(1)}$	$TN6_{(ul)} = (x_{(n)} - x_{(1)})/s$	Greater	3 –1000	3 – 100
N8 (two-sided)	Extreme	$x_{(n)}$ or $x_{(1)}$	$TN8 = Max\colon\{(x_{(n)} - x_{(n\text{-}1)}) / (x_{(n)} - x_{(1)}),$ $(x_{(2)} - x_{(1)}) / (x_{(n)} - x_{(1)})\}$	Greater	4 – 30	4 – 100
N14	Extreme	$x_{(n)}$ or $x_{(1)}$	$TN14 = \left[\dfrac{n^{1/2}\left\{\sum_{i=1}^{n}(x_i - \bar{x})^3\right\}}{\left\{\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}^{3/2}}\right]$	Greater	5 –1000	5 – 100
N15	Extreme	$x_{(n)}$ or $x_{(1)}$	$TN15 = \left[\dfrac{n\left\{\sum_{i=1}^{n}(x_i - \bar{x})^4\right\}}{\left\{\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}^{2}}\right]$	Greater	5 –1000	5 – 100

the same simulation scheme as reported earlier (Verma and Quiroz-Ruiz, 2006). Similarly, because of the availability of the upper or lower version of the statistic (Table 1), 20 results of critical values were obtained for tests N3–k=2,3,4, and N4–k=1,2,3,4.

Barnett and Lewis (1994; pp. 218-221) have shown that the statistics $TN1_{(u)}=(x_{(n)}-\bar{x})/s$ and $TN4_{(1u)}=S^2_{(n)}/S^2$ are really equivalent (for the meaning of $S^2_{(n)}$ and S^2 see pp. 95-96 of Verma, 2005), and bear the following relationships:

$$TN4_{(1u)} = 1 - \frac{n}{(n-1)^2} TN1^2_{(u)} \tag{4}$$

and conversely,

$$TN1_{(u)} = \sqrt{\frac{(n-1)^2}{n}(1 - TN4_{(1u)})} \tag{5}.$$

Similar equations are valid for lower outlier statistics $TN1_{(l)}$ and $TN4_{(1l).}$

The above exact relationships enable us to use the simulation results of $TN1_{(1u)}$ and $TN4_{(1l)}$ to convert them, respectively, for those of $TN1_{(u)}$ and $TN1_{(l)}$, and vice versa. In this way, we obtained 20 more results of critical values from simulations of sizes 100,000 for tests N1 and N4–k=1. Due to the above exact relationships and in order to avoid the "artificial" decrease of standard errors of the mean critical values, we used different starting points (either the first or the second datum in a given simulated normal distribution) for drawing samples of sizes up to 100 (for tests N1 and N4–k=1) from the simulated normal independently and identically distributed IID N(0,1) distributions (see more details in Verma and Quiroz-Ruiz, 2006, and references cited therein).

Thus, in summary, the critical values for tests N1 and N4–k=1 were from 40 simulations of sizes 100,000; for tests N3–k=2,3,4 and N4–k=2,3,4 from 20 such simulations; and for tests N2, N5, N6, N8, N14, and N15 from 10 such simulations.

However, for tests N2, N6, and N15, the standard errors of the mean critical values obtained from 10 simulations of sizes 100,000 were still large (the standard error was generally on the third decimal place, and was the highest for test N15 that involves a fourth order statistic, see the definition of *TN*15 in Table 1). Therefore, we decided to simulate these critical values from 10 simulations, but each of sizes 500,000 (five times greater than the sizes used by Verma and Quiroz-Ruiz, 2006), for which 10 different sets of 50,000,000 random normal variates had to be generated. In fact, for tests N6 and N15 we carried out a total of 20 simulations each of sizes 500,000. Twenty instead of 10 simulations were made possible by varying the starting point for a given set of IID N(0,1) while drawing samples of sizes up to 100.

The final critical values tabulated correspond to the largest simulation sizes for each test. These are: 100,000 for N3–k=2,3,4, N4–k=2,3,4, N5, N8, and N14; and 500,000 for N1, N4–k=1, N2, N6, and N15. The mean ($\bar{x}$) and standard error of the mean ($se_{\bar{x}}$) of all individual simulation results for each n and α were estimated and the best results (with the smallest errors) reported in tabular form. The median critical values were found to be in close agreement with these mean values, ascertaining the simulated critical values to be also normally distributed.

RESULTS OF NEW CRITICAL VALUES

The new critical values for 22 discordancy test variants (Table 1), for each n from 3 (or 4, 5, 6, 7, 8, or 9, depending on the type of statistic to be calculated) up to 100 and α = 0.30, 0.20, 0.10, 0.05, 0.02, 0.01, and 0.005 (corresponding to confidence level of 70% to 99.5%, or equivalently significance level of 30% to 0.5%) are summarized in Tables A1-A14, available from the journal web site (electronic supplement 23-3-01). The footnotes in these tables also present the standard error of the mean ($se_{\bar{x}}$) for these critical values. Thus, for all cases our present values are more reliable (error is on the fourth or even the fifth decimal place, although in some cases a small number –1 or 2– on the third decimal place) than the earlier literature values (compiled by Barnett and Lewis, 1994; and Verma, 2005; see also Pearson and Hartley, 1976). In fact, the errors of these literature values are never precisely known. The errors of our present simulations for all values of α ranged as follows: ~0.0006-0.0011 for test N1 (Table A1); ~0.0002-0.0010 for

Notes to Table 1: [a] Test code (N series) is from Barnett and Lewis (1994). The symbols for test statistics $TN1_{(u)}$, $TN1_{(l)}$, $TN2$, etc. are proposed by Verma (2005). The subscripts $_{(u)}$, $_{(l)}$, $_{(2u)}$ and $_{(2l)}$ are, respectively, upper (the highest), lower (the lowest), upper pair, and lower pair observations. The test statistics are self explanatory except the statistics of the type "*reduced* sum of the squares"/"*total* sum of the squares" for example, $S^2_{(n)}/S^2$ for test N4–k=1, proposed by Grubbs (1950, 1969), which need some explanation. For an ordered array $x_{(1)}, x_{(2)}, x_{(3)},\ldots x_{(n-2)}, x_{(n-1)}, x_{(n)}$ the S^2 term is calculated using all data $S^2 = \sum_{i=1}^{n}(x_{(i)} - \bar{x})^2$, where $\bar{x}$ is the arithmetic mean ($\bar{x} = \sum_{i=1}^{n} x_{(i)}/n$), whereas $S^2_{(n)}$ is computed from the (n-1) remaining data $x_{(1)}, x_{(2)}, x_{(3)},\ldots x_{(n-2)}, x_{(n-1)}$, after eliminating the highest datum to be tested $x_{(n)}$ (see the subscript $_{(n)}$ in the term $S^2_{(n)}$) as follows: $S^2_{(n)} = \sum_{i=1}^{n-1}(x_{(i)} - \bar{x}_n)^2$ where $\bar{x}_n = \sum_{i=1}^{n-1} x_{(i)}/(n-1)$. The other statistics of the type $S^2_{(n)}/S^2$, such as $S^2_{(1)}/S^2$ or $S^2_{(n)(n\text{-}1)}/S^2$ are calculated in a similar manner. For more details, see Verma (2005) . Finally, note that, in the present work, the $n_{\max}$ has been increased to 100 for all tests (see Tables A1-A14 in the electronic supplement 23-03-01), and when critical values were already available for this $n_{\min} - n_{\max}$ range, the new values are shown to be more precise and accurate. Additional references for information on these tests: Grubbs (1950, 1969); Dixon (1950); King (1953); Quesenberry and David (1961); David and Paulson (1965); Pearson and Hartley (1976); Shapiro *et al.* (1968); Stefansky (1971); McMillan (1971); Grubbs and Beck (1972); Tietjen and Moore (1972, 1979); Moran and McMillan (1973); Rosner (1975); Tiku (1975, 1977); Prescott (1978, 1979); Hawkins (1979); Srivastava (1980).

test N2 (Table A2), ~0.0002–0.0009 for N3–k=2 (Table A3); ~0.0002–0.0010 for N3–k=3 (Table A4); ~0.0003–0.0010 for N3–k=4 (Table A5); ~0.00002–0.0001 for N4–k=1 (Table A6; note most errors on the fifth decimal place); ~0.00007–0.0003 for N4–k=2 (Table A7; note some errors on the fifth decimal place); ~0.00008–0.0003 for N4–k=3 (Table A8; note some errors on the fifth decimal place); ~0.00007–0.0003 for N4–k=4 (Table A9; note some errors on the fifth decimal place); ~0.00008–0.0004 for N5–k=2 (Table A10; note some errors on the fifth decimal place); ~0.0002–0.0008 for N6–k=2 (Table A11); ~0.0001–0.0005 for N8 (Table A12); ~0.0002–0.0009 for N14 (Table A13); and ~0.0002–0.0025 for N15 (Table A14).

As for our earlier tables for the six Dixon tests (Verma and Quiroz-Ruiz, 2006), these new critical value data, along with their individual uncertainty estimates, are available in other formats such as *txt* or *Excel* or *Statistica*, on request from any of the authors (S.P. Verma spv@cie.unam.mx, or A. Quiroz-Ruiz aqr@cie.unam.mx).

In spite of this important observation, *i.e*., our simulation results being more precise than the earlier literature values, we decided to compare the literature critical values for the most commonly used $\alpha = 0.05$ (5% SL) and 0.01 (1% SL) with our results to highlight the similarities and differences between the two sets. These comparisons are graphically presented in: Figure 1 for single-outlier tests N1, N2, N4–k=1, N8, N14, and N15; Figure 2 for two multiple-outlier tests N3–k=2,3,4 and N4–k=2,3,4; and Figure 3 for the remaining two multiple-outlier tests N5–k=2 and N6–k=2. The differences between our newly simulated critical values and the literature data (Figures 1-3) are as follows (listed in ascending order of these differences): up to ~0.15% for test N1, ~0.2% for test N6(k=2), ~0.4% for tests N2 and N3–k=4, ~0.7% for test N3–k=3, ~0.8% for test N15, ~1% for tests N3–k=2, N4–k=1, and N4–k=2, ~2% for test N8, ~6% for test N4–k=3, ~8% for test N14, ~15% for test N4–k=4, and ~20% for test N5(k=2).

These differences are generally larger for 1% SL (99% CL, or $\alpha = 0.01$) values than for 5% SL (95% CL, or $\alpha = 0.05$). We attribute these differences to the inaccuracy of literature values (some of them were generated by just one simulation of small sizes of 10,000 only; for others, simulation procedure was not specified) as compared to the present work based on 10 to 40 independent simulations of very large sizes (from 100,000 requiring 10,000,000 random normal variates and IID N(0,1) values to 500,000 requiring 50,000,000 random normal variates and IID N(0,1) values).

For single-outlier test N1 there seem to be small systematic differences for 1% SL values for *n* >50 (Figure 1a) between our values and the literature values (Grubbs and Beck, 1972); for these sizes (*n* >50), the literature values are systematically somewhat greater than our new simulated values.

The less-precise literature values for another single-outlier test N2 (available for *n* only up to 20) differ from our new values by < 0.4% (Figure 1b).

For multiple-outlier test N3 the differences between our newly simulated values and the literature are somewhat larger, up to about 1% for the k=2 variant (Figure 2a), about 0.7% for k=3 (Figure 2b), and about 0.4% for k=4 (Figure 2c). This may be due to the fact that the literature values were generated from a single simulation of sizes 10,000 (Barnett and Lewis, 1994) –much smaller than our present simulations of sizes 500,000 repeated up to 20 times. These differences are even larger than the assumed minimum error of ±0.01 shown as blue-dashed and red-dotted lines in Figure 2 (see numerous %Difference symbols fall above or below these curves).

For multiple-outlier test N4 these differences are still larger than for tests N1 to N3, up to about 1.1% for the *k*=1 (single-outlier version of test N4) and *k* = 2 variants (Figures 1c and 2d), up to about 6% for *k* = 3 (Figure 2e), and up to about 15% for k = 4 (Figure 2f).

For multiple-outlier test N5–k=2, the differences of up to 20% (Figure 3a) are due to much smaller sizes of 10,000 for the literature critical values (Barnett and Lewis, 1994).

For multiple-outlier test N6–k=2, on the other hand, very small differences of up to about 0.2% (Figure 3b) were observed; the reason for such a close agreement is not clear.

For single-outlier Dixon-type test N8 (King, 1953), also known as Dixon Q-test, the differences are generally quite large, reaching up to about 2%. This is because the literature values were simulated using small sizes of 10,000.

Single-outlier tests N14 (skewness) and N15 (kurtosis) are specially recommended to be used consecutively (Barnett and Lewis, 1994). Although for test N14 the differences reach up to about 8%, most critical values for *n* >20 differ only by a much smaller amount (Figure 1e).

Finally, for single-outlier test N15, the differences are relatively small (up to ~0.8%; Figure 1f).

APPLICATIONS IN SCIENCE AND ENGINEERING

The tests (Table 1) after extending their applicability to samples of sizes up to 100, can be applied to all examples earlier summarized by us (Verma and Quiroz-Ruiz, 2006). These include: (1) Agricultural and Soil Sciences (Stevens *et al.*, 1995; Batjes, 2005; Lugo-Ospina *et al.*, 2005; Luedeling *et al.*, 2005); (2) Aquatic environmental research (Thomulka and Lange, 1996); (3) Astronomy (Taylor, 2000); (4) Biology (Linkosalo *et al.*, 1996; Schaber and Badeck, 2002); (5) Biomedicine and Biotechnology (Freeman *et al.*, 1997; Woitge *et al.*, 1998; Sevransky *et al.*, 2005); (6) Chemistry (Zaric and Niketic, 1997); (7) Electronics (Jakubowska and Kubiak, 2005); (8) Ecology (Yurewicz, 2004); (9) Geochronology (Bartlett *et al.*, 1998; Wang *et al.*, 1998;

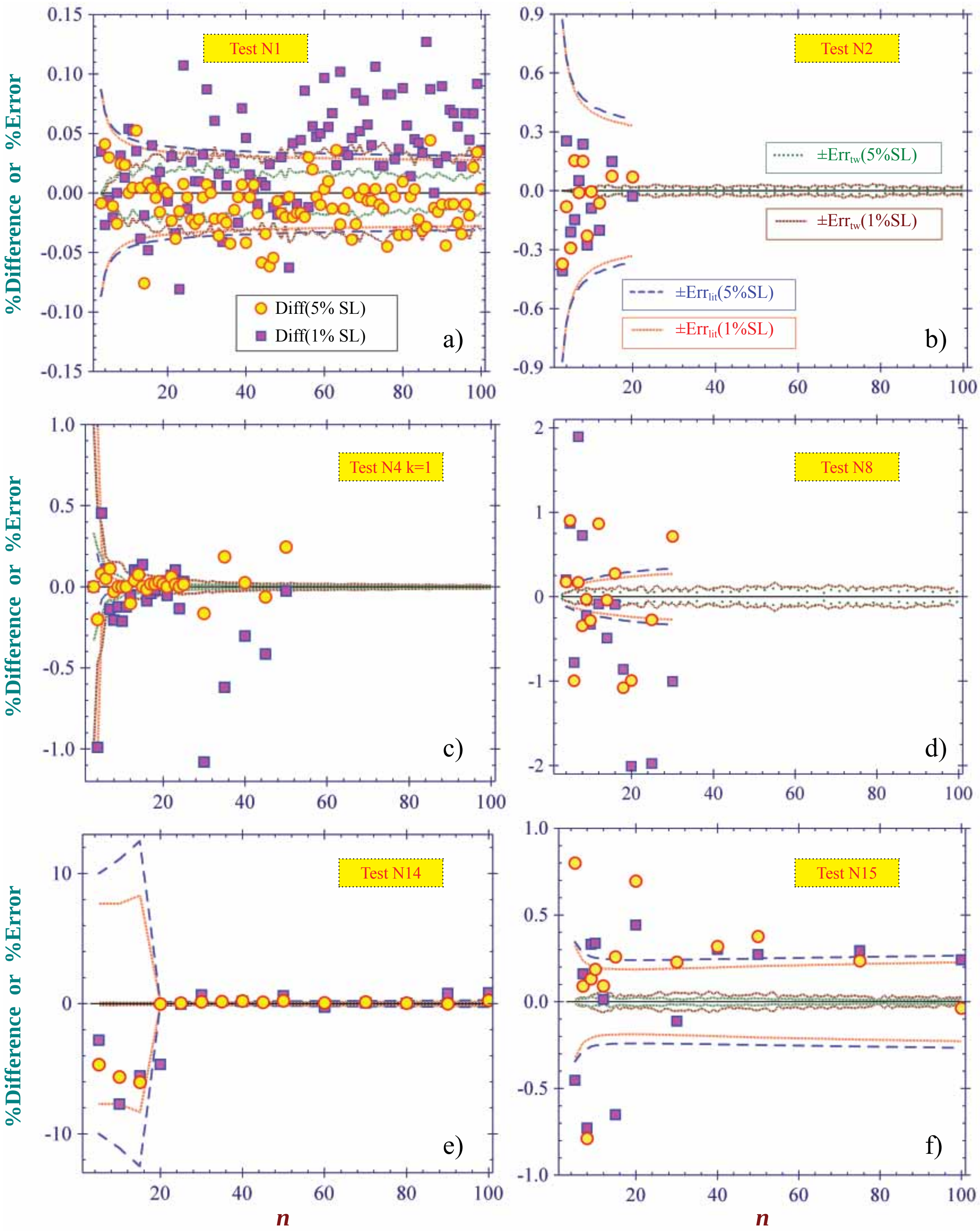

Figure 1. Comparison of new simulation results for single-outlier tests, with the literature critical values for most frequently-used $\alpha = 0.05$ and 0.01. The y-axis either represents the %Difference, i.e., [$100 \cdot (CV_{lit}\text{-}CV_{tw})/CV_{tw}$] being the % difference between the literature critical values (CV_{lit} – critical value from Barnett and Lewis, 1994, or Verma, 2005, see also Pearson and Hartley, 1976, and other references cited in the footnote of Table 1) and the present critical values (CV_{tw} – critical value obtained in this work), or %Error, i.e., assumed % error in literature values (see below) or estimated % error of our present simulations (the standard error of critical value expressed in percent). The %Difference parameter is expressed assuming a red-yellow circle for 5%SL critical values and a blue-pink square for 1%SL critical values. The %Error parameter is expressed for the literature data by a dotted line (blue for 5%SL; i.e., ±Err$_{lit}$(5%SL)) and a dashed line (red for 1%SL; i.e., ±Err$_{lit}$(1%SL)); because no errors were explicitly reported for the literature critical values, the %Error is the minimum error of 1 on the last digit reported for a given critical value. This %Error parameter for the present simulation is the actual standard error of the mean expressed in percent and is shown using dotted lines (green for 5%SL; i.e., ±Err$_{tw}$(5%SL), and dark red for 1%SL; i.e., ±Err$_{tw}$(1%SL)). Reference line for 0% difference is shown as a horizontal solid line. Note the %Error for the present simulations (±Err$_{tw}$(5%SL) and ±Err$_{tw}$(1%SL)) always lie close to this 0% difference horizontal line. These symbols used are also explained in parts a) and b). Parts a) to f) are, respectively, for tests N1, N2, N4–k=1, N8, N14, and N15.

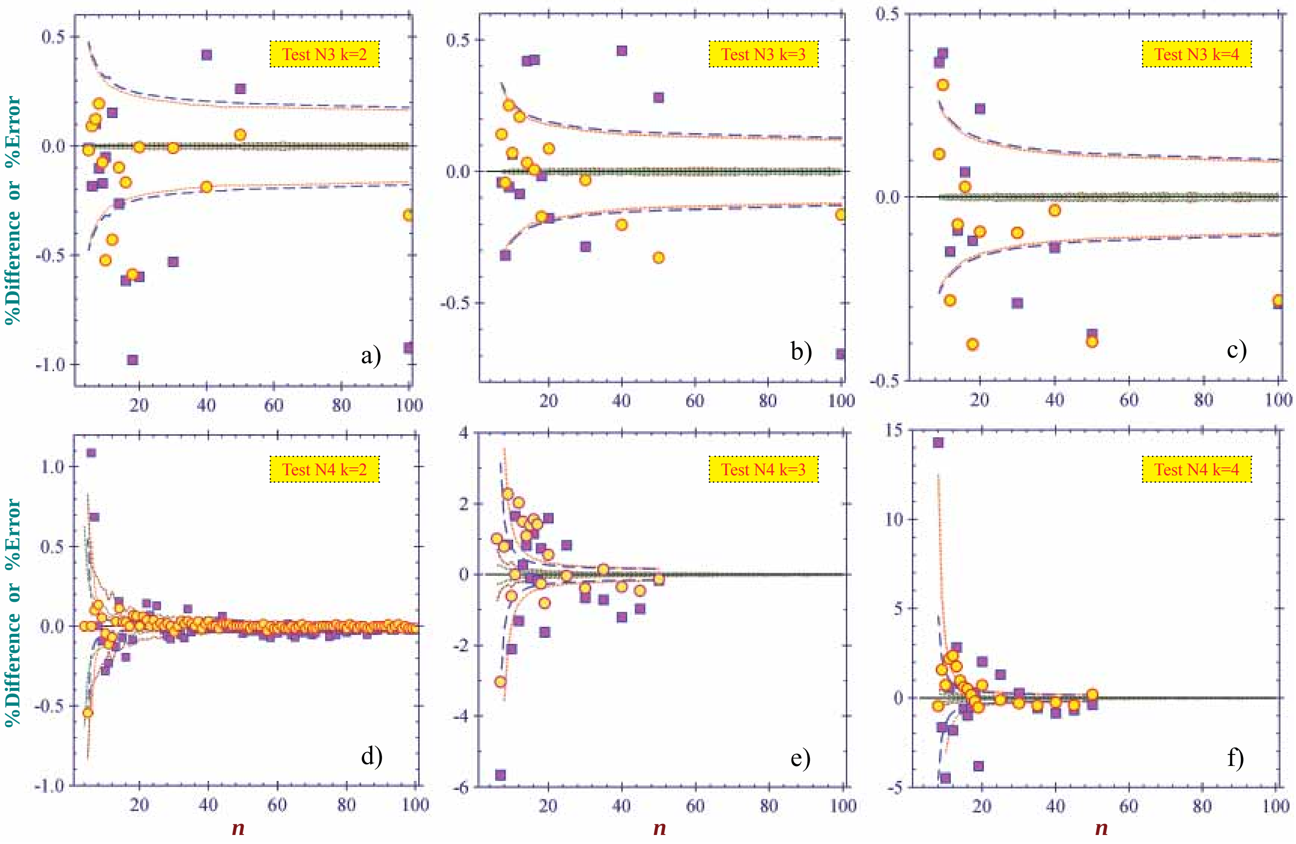

Figure 2. Comparison of new simulation results for multiple-outlier tests, with the literature critical values for α = 0.05 and 0.01. Parts a) to f) are, respectively, for tests N3–k=2, N3–k=3, N3–k=4, N4–k=2, N4–k=3, and N4–k=4. For more explanation see Figure 1.

Dougherty-Page and Bartlett, 1999); (10) Geodesy (Kern *et al.*, 2005); (11) Geochemistry (Treviño-Cázares *et al.*, 2005); (12) Isotope Geology (Morán-Zenteno *et al.*, 1998); (13) Medical science and technology (Tigges *et al.*, 1999; Hofer and Murphy 2000; Reed *et al.*, 2002; Stancak *et al.*, 2002; Cooper *et al.*, 2006); (14) Meteorology (Graybeal *et al.*, 2004); (15) Paleontology (Alberdi and Corona-M., 2005; Esquivel-Macías *et al.*, 2005; Ifrim *et al.*, 2005; Villaseñor *et al.*, 2005); (16) Petroleum hydrocarbons and organic compounds in sediment samples (Villeneuve *et al.*, 2002); (17) Quality assurance and assessment programs in Biology and Biomedicine (Ihnat, 2000; Patriarca *et al.*, 2005), in cement industry (Sieber *et al.*, 2002), in Food Science and Technology (In't Veld, 1998: Suhren and

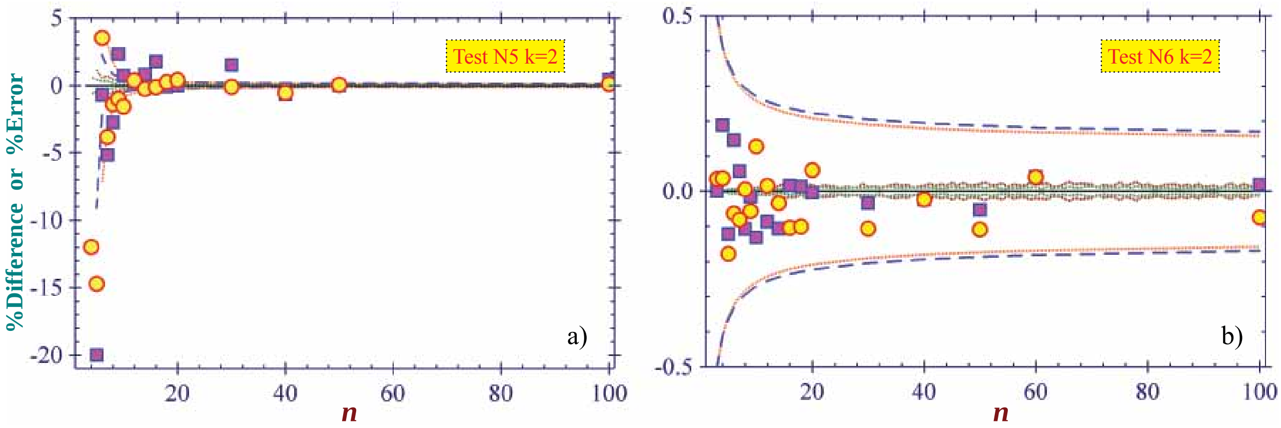

Figure 3. Comparison of new simulation results for multiple-outlier tests, with the literature critical values for α = 0.05 and 0.01. Parts a) to b) are, respectively, for tests N5–k=2 and N6–k=2. For more explanation see Figure 1.

Walte, 2001; Langton *et al.*, 2002; Morabito *et al.*, 2004; Villeneuve *et al.*, 2004), in Environmental and Pollution Research (Dybczyński *et al.*, 1998; Gill *et al.*, 2004), in Nuclear Science (Lin *et al.*, 2001), in Rock Chemistry (Velasco-Tapia *et al.*, 2001; Lozano and Bernal, 2005), in Soil Science (Hanson *et al.*, 1998; Verma *et al.*, 1998), and in Water Research (Holcombe *et al.*, 2004); (18) Structural Geology (Dávalos-Álvarez *et al.*, 2005); (19) Water Resources (Buckley and Georgianna, 2001); and (20) Zoology (Harcourt *et al.*, 2005).

Further, as was suggested by Shoemaker *et al.* (1996) for the Dixon tests, our new critical value tables will be equally useful for applying these discordancy tests for identifying outliers in linear regressions such as those employed by: Verma (2006) for trace element inverse modeling; Guevara *et al.* (2005) and Santoyo *et al.* (2006) for instrumental calibration purposes; and Verma *et al.* (2006) for fluid chemistry and geotermometric temperatures.

Users of a number of internet sites (*e.g.*, SCNPHT, 2006; WORM Database, 2006; and ESMP, 2006) will also benefit from the incorporation of these new tables of critical values and the respective tests into these systems.

Specific examples

We now present a number of examples or case histories to illustrate the use of all discordancy tests for which new critical values have been obtained in this work and by Verma and Quiroz-Ruiz (2006). We have designed and used a *Statistica* spreadsheet to apply all 15 tests (N1 to N15) with 33 test variants. A computer program is currently under preparation for an easy use of our multiple-test method. For these applications, we chose the strict confidence level of 99% (*i.e.*, we used the 99% CL, or 1% SL, or 0.01 α column; see the respective critical values in Tables A1-A14 in the electronic supplement to present work; and tables 2-7 of Verma and Quiroz-Ruiz, 2006).

Example 1: Petroleum hydrocarbons in a sediment reference sample

We used the example of IAEA-417 (IAEA–International Atomic Energy Agency) used by Villeneuve *et al.* (2002) for an inter-laboratory study to highlight our multiple-test method and compare its performance with the box-and-whisker plot method used by the original authors. The individual data for six selected compounds (phenanthrene, chrysene, fluorenthene, pyrene, benz(a) anthracene, and benz(a) pyrene) were summarized in our earlier paper (see table 9 in Verma and Quiroz-Ruiz, 2006, compiled from the original report by Villeneuve *et al.*, 2002). Our multiple-test method consists of applying all nine single-outlier (with 13 test variants) and seven multiple-outlier tests (with 20 test variants) to a given set of data (see Table 2). Note one test (N4) is common to both these categories; in fact, we have used 15 (and not 16) tests –N1 to N15 (because precise critical values for test N16 are still not available; see Barnett and Lewis, 1994, or Verma, 2005). The performance of these tests for detecting discordant outliers in each set of petroleum hydrocarbon data is summarized in Table 2, in which all applied tests, along with the tests that were successful or unsuccessful in detecting outliers, in the categories of single- and multiple-outlier tests are listed.

After the identification of discordant outliers, these were eliminated and the tests applied to the remaining data until no more outliers were detected by any of the 15 tests or 33 test variants. The concentration data along with the basic statistical information are summarized in Table 3. It is not surprising to see such highly dispersed data for a given compound in the same sample distributed and analyzed by laboratories around the world (see the initial range of individual values that differ by nearly two orders of magnitude); unfortunately, this is the “state-of-the-art” in geochemistry! The same type of situation should exist in other science or engineering fields as well (see for example, the numerous references cited above for “Quality Assurance and Assessment Programs”).

More discordant outliers were detected by the present multiple-test method than the box-and-whisker plot method used by the original authors (Villeneuve *et al.*, 2002) in the data for most compounds listed in Table 3. Note the initial mean and standard deviation data strongly differ from the final statistical parameters.

We conclude that the multiple-test method exemplified in this work can be advantageously used in future to arrive at the final statistical parameters in such inter-laboratory studies. This multiple-test method was already proposed and documented by Verma and collaborators (Verma, 1997, 1998, 2005; Verma *et al.*, 1998; Guevara *et al.*, 2001; Velasco *et al.*, 2000; Velasco-Tapia *et al.*, 2001). The availability of new, precise and accurate critical values for sample sizes up to 100 (this work; and Verma and Quiroz-Ruiz, 2006) makes this proposal much more powerful than the other methods such as the box-and-whisker plot method.

Example 2: Two chemical elements in a geochemical reference material from Japan

We present the example of just two elements (a major element or oxide MgO and a trace element Zr) in a rock reference material peridotite JP-1 from Japan. The same kind of reasoning will be valid for other elements in JP-1 and for all other international geochemical reference materials. The individual data were downloaded from the Geological Survey of Japan–Geochemical Reference Samples Database (GSJ-GRS, 2006) and are presented in the footnote of Table 2. The results of the application of our multiple-test method are summarized in Tables 2 and 3. Unfortunately, no information is available on the currently used method of data processing for the Japanese geochemical reference materials; the only available information was taken from Imai *et al.* (1995). These authors had used the two standard deviation method for outlier detection, which was already criticized by Verma and collaborators (Verma, 1997, 1998, 2005; Verma *et al.*, 1998; Guevara *et al.*, 2001; Velasco *et al.*, 2000; Velasco-Tapia *et al.*, 2001).

Example 3: The need of outlier tests in other geoscience studies

We now briefly comment on the need of using the above multiple-test method in numerous geoscience studies. As the third example, we applied the multiple-test method to the oxygen isotope data for the Los Azufres hydrothermal system (see individual data in the 4th column of table 3 of Torres-Alvarado, 2002) to test if there were any outliers in these data. Our application of the multiple-test method showed no outliers in this dataset, implying that the conventional location and scale parameters –mean and standard deviation– can be safely used to handle these data (see Verma, 2005 for more details).

This and the other two examples (Tables 2 and 3) clearly show the need for processing the raw univariate data by the multiple-test method, irrespective of if or not any outliers were eventually detected. This is true, for example, for the recent studies by Mendoza-Amézquita *et al.* (2006) and Ramos-Arroyo and Siebe-Grabach (2006). These authors did not actually report the individual data; consequently, the multiple-test method could not be applied by us. However, they reported mean and standard deviation values as indicators of central tendency (location) and dispersion (scale) parameters, for which it is *always* advisable to first apply our multiple-test method to identify any discordant outliers and only then use these location and scale parameters for interpretation purposes (see Verma, 2005 for more details). The discordant outliers, if present, are of much value to further understand the geological processes (again, see Verma, 2005 for more details).

As yet other application examples, we may mention the chemical data for a given mineral or rock type such as

Table 2. Examples of the performance of discordance tests as applied to diverse databases.

Database [a]	**Applied tests (99% CL)**		**Successful tests**		**Unsuccessful tests**	
	Single outlier-type	**Multiple outlier-type**	**Single outlier-type**	**Multiple outlier-type**	**Single outlier-type**	**Multiple outlier-type**
IAEA-417 Phenanthrene	N1, N2, N4–k=1, N7-N10, N14, N15	N3–k=2,3,4, N4–k=2,3,4, N5, N6, N11-N13	N1, N2, N4–k=1, N7-N10, N14, N15	N3–k=2,3,4, N4–k=2,3,4, N5, N11-N13	---	N6
IAEA-417 Chrysene	N1, N2, N4–k=1, N7-N10, N14, N15	N3–k=2,3,4, N4–k=2,3,4, N5, N6, N11-N13	N1, N2, N4–k=1, N14, N15	N3–k=2,3,4, N4–k=2,3,4, N5, N6, N11-N13	N7-N10	---
IAEA-417 Fluorenthene	N1, N2, N4–k=1, N7-N10, N14, N15	N3–k=2,3,4, N4–k=2,3,4, N5, N6, N11-N13	N1, N2, N4–k=1, N14, N15	N3–k=2,3,4, N4–k=2,3,4, N5, N6, N11-N13	N7-N10	---
IAEA-417 Pyrene [b]	N1, N2, N4–k=1, N7-N10, N14, N15	N3–k=2,3,4, N4–k=2,3,4, N5, N6, N11-N13	N1, N2, N4–k=1, N7, N9, N10, N14, N15	N3–k=2,3,4, N4–k=2,3,4, N5, N11	N8	N6, N12, N13
IAEA-417 Benz(a) Anthracene	N1, N2, N4–k=1, N7-N10, N14, N15	N3–k=2,3,4, N4–k=2,3,4, N5, N6, N11-N13	N1, N2, N4–k=1, N7-N10, N14, N15	N3–k=2,3,4, N4–k=2,3,4, N5, N6, N11-N13	---	---
IAEA-417 Benz(a) Pyrene	N1, N2, N4–k=1, N7-N10, N14, N15	N3–k=2,3,4, N4–k=2,3,4, N5, N6, N11-N13	---	---	N1, N2, N4–k=1, N7-N10, N14, N15	N3–k=2,3,4, N4–k=2,3,4, N5, N6, N11-N13
Peridotite JP-1 MgO [c]	N1, N2, N4–k=1, N7-N10, N14, N15	N3–k=2,3,4, N4–k=2,3,4, N5, N6, N11-N13	---	N4–k=2,3,4, N5, N6	N1, N2, N4–k=1, N7-N10, N14, N15	N3–k=2,3,4, N11-N13
Peridotite JP-1 Zr [c]	N1, N2, N4–k=1, N7-N10, N14, N15	N3–k=2,3,4, N4–k=2,3,4, N5, N6, N11-N13	N1, N2, N4–k=1, N14, N15	N3–k=2,3,4, N4–k=2,3,4, N11-N13	N7-N10	N5, N6

[a] For the individual data for IAEA-417 reference sample reproduced as such from the original report (Villeneuve *et al.*, 2002), see table 9 in Verma and Quiroz-Ruiz (2006); the JP-1 data were from Japan Geochemical Reference Material Website (JGRMW, 2006). [b] The second cycle also detected discordant outliers. --- signifies no test in this category. [c] The ordered data array for MgO in JP-1 was: 41.12, 42.784, 42.8, 42.96, 43.5, 43.53, 43.9, 43.91, 44.06, 44.08, 44.26, 44.3, 44.35, 44.38, 44.5, 44.56, 44.6, 44.61, 44.72, 44.72, 44.72, 44.74, 44.76, 44.77, 44.8, 44.81, 44.86, 44.9, 44.94, 45.04, 45.12, 45.15, 45.34, 45.84, 45.91, 46.05, 46.18, 46.24, 46.6, 47.26, 48.0; and for Zr it was: 3, 3.9, 3.9, 4, 4, 4, 4.4, 4.7, 4.8, 5.09, 5.25, 5.34, 5.5, 5.8, 6, 6, 6, 6, 6.9, 7, 7, 7, 7, 7.9, 8, 8, 8.2, 9, 9.13, 9.7, 9.9, 10, 10, 11, 11, 12, 12.2, 16, 21, 25. The tests were applied to these datasets.

Table 3. Results of six petroleum hydrocarbons in IAEA-417 sediment reference sample (Villeneuve *et al.*, 2002), two chemical elements in reference samples peridotite JP-1 (Japan Geochemical Reference Material Website, 2006), and application of our multiple-test method (15 tests with 33 test variants).

Compound	**Initial statistics**				**Final statistics (this work)**					**Final statistics (literature)**				
	n_{in}	$\bar{x}_{in}$	s_{in}	R_{in}	o_f	n_f	$\bar{x}_f$	s_f	R_f	o_l	n_l	$\bar{x}_l$	s_l	R_l
Phenanthrene	45	4400	2900	852 – 16400	5	40	3800	1300	1090 – 6310	2	43	3900	1500	852 – 7572
Chrysene	45	4700	4500	140 – 22500	5	40	3600	1500	180 – 6270	3	42	3600	1700	140 – 6870
Fluoranthene	49	8700	5900	883 – 36250	5	44	7600	2500	2477 – 12290	2	47	7700	3000	883 – 14540
Pyrene	48	7500	5100	462 – 28950	7	41	6100	2000	1475 – 10228	5	43	6000	2200	462 – 10570
Benz[a]Anthracene	42	3600	2300	60 – 15000	5	37	3200	1000	1018 – 4757	2	40	3200	1200	60 – 5370
Benz[a]Pyrene	44	2800	1200	6.3 – 5160	0	44	2800	1200	6.3 – 5160	0	44	2800	1200	6.3 – 5160
Peridotite JP–1 – MgO [a]	41	44.7	1.2	41.1 – 48	8	33	44.7	0.6	43.5 – 46.18	—	24	44.60	0.64	—
Peridotite JP–1 – Zr [a]	40	8.0	4.5	3 – 25	4	36	6.8	2.4	3 – 12	—	—	5.92	—	—

For the individual data reproduced as such from the original report, see Verma and Quiroz-Ruiz (2006). *n*: number of data; *x*: mean; *s*: standard deviation; *R*: range; *o*: number of discordant outliers detected by a given method; the subscripts : $_{in}$, $_f$ and $_l$ refer to the initial, final (after applying all discordance tests including those presented by Verma and Quiroz-Ruiz, 2006; this work), and box-and-whisker plot method (literature, Villeneuve *et al.*, 2002); the difference between n_{in} and n_f gives the number of discordant outliers detected by the discordance tests presented in this work and Verma and Quiroz-Ruiz (2006); similarly, the difference between n_{in} and n_l gives the number of discordant outliers detected by the box-and-whisker plot method presented by Villeneuve *et al.* (2002). [a] No information on the method used for outlier detection (JGRMW, 2006); partial statistical information taken from Imai *et al.* (1995), which is referred to on this website.

those presented by Vattuone *et al.* (2005), Rodríguez (2005), or Lozano and Bernal (2005). We suggest that these raw data should first be processed using this outlier-scheme (multiple-test method) and only then their mean and standard deviation can be used as the proper central tendency and dispersion parameters for a given mineral or rock type.

Finally, as an example, we may mention that the application of our multiple-test method may also be highly advantageous for correctly processing the grain size data of sands recently reported by Kasper-Zubillaga and Carranza-Edwards (2005), in which the mean, skewness and kurtosis parameters were used (after applying some other tests) for interpreting these data. These authors used a probably inappropriate Kolmogorov-Smirnov one-sample test for normality, which is best applied when the mean and standard deviation of the normal distribution are known *a priori* and not estimated from the data (Statistica for Windows, 1998). The Levene's test used by these authors has also certain restrictions and conditions to be fulfilled.

In summary, therefore, the multiple-test method proposed and exemplified in our paper is strongly recommended to be used for experimental data under the assumption that the data are drawn from a normal distribution and departure from this assumption due to any contamination or presence of discordant outliers can be properly handled by tests N1 to N15 (15 tests with their 33 variants). Unfortunately, most commercially available software packages do not pay due attention to these fundamental tests. Nevertheless, the computer program currently under preparation by our group will greatly facilitate the use of our multiple-test method. In the mean while, the interested persons can use the *Statistica* spreadsheet prepared by us or an old computer program SIPVADE (Verma *et al.*, 1998) to process their experimental data.

Two standard deviation method

Comparison with a valid multiple-outlier test

We present additional arguments against the use of the "two standard deviation" method (henceforth called the 2*s* method) by comparing the performance of this method with that of just one multiple-outlier test N3–k=2,3,4 (with three variants). For this purpose, we applied this test to the inter-laboratory data on the geochemical reference material granodiorite GSP-1 from U.S.A., compiled and processed by Gladney *et al.* (1992) using the 2*s* method. In order to make this comparison objective, test N3–k=2,3,4 was applied at the 95% confidence level (or 5% significance level) to exactly the same compiled data that had been evaluated by Gladney *et al.* (1992).

The results obtained for the trace element data with initial number of observations between 20 and 100, are shown in Table 4. The inter-laboratory data show a large scatter for most elements as seen from the wide range of values as well as from large standard deviation values. Test N3–k=2,3,4 detected, in general, more discordant outlier than the 2*s* method (Table 4). A graphical comparison of the results is presented in Figure 4. Those data that fall close to the diagonal line in Figures 4a and 4b or close to the horizontal "zero value" line in Figures 4c and 4d show no significant differences between the two methods. However, many elements plot away from these reference lines, and document the superiority of test N3–k=2,3,4 as compared to the 2*s* method (Figure 4).

It is beyond any reasonable doubt that the combined performance of all the tests proposed in our multiple-test method should be at least similar to, if not better than, the test N3–k=2,3,4 evaluated here. Therefore, the 2*s* method can be safely abandoned, and our proposal of multiple-test method adopted for all future work, especially for evaluating inter-laboratory compositional data.

Not recommended for samples of finite size

We present more arguments against the use of the 2*s* method to process inter-laboratory compositional data (or for that matter, any other kind of univariate data). We have plotted in Figure 5 the 95% and 99% critical values for tests N1 and N2. The use of the 2*s* method is shown schematically as the solid horizontal line, which means that all individual data falling anywhere below this line (in the field marked A) will be considered as legitimate or valid observations, whereas any observations lying above this line in any of the fields B, C, or D will be considered discordant outliers. Note the sample size-independent nature of the 2*s* method represented schematically by the "critical values" lying on the horizontal line in Figure 5 –an erroneous property according to Barnett and Lewis (1994). As was explained earlier, tests N1 and N2 are the statistically-correct versions of this 2*s* method; the corresponding critical values strongly depend on the sample size (see the different critical value "curves" –not "straight lines"– in Figure 5). The observa-

Table 4. Comparison of the 2*s* method (Gladney *et al.*, 1992) with the multiple-outlier test N3–k=2,3,4 (this work), using trace element data (with initial *n* between 20 and 100) for the geochemical reference material granodiorite GSP-1.

Compound	**Initial statistics**				**Final statistics (this work)**					**Final statistics (literature)**				
	n_{in}	$\bar{x}_{in}$	s_{in}	R_{in}	o_f	n_f	$\bar{x}_f$	s_f	R_f	o_l	n_l	$\bar{x}_l$	s_l	R_l
Ag [a]	20	0.22	0.65	0.015 – 3	5	15	0.090	0.013	0.075 – 0.12	5	15	0.086	0.013	n.r.
Be	28	1.4	0.6	0.5 – 3.6	12	16	1.19	0.10	1.0 – 1.32	4	24	1.35	0.27	n.r.
C	26	315	62	82 – 900	7	19	320	60	232 – 420	6	20	315	62	n.r.
Cd	24	0.061	0.024	0.016 – 0.12	6	18	0.058	0.008	0.048 – 0.072	5	19	0.058	0.008	n.r.
Cl	28	333	33	267 – 400	0	28	333	33	267 – 400	3	25	332	24	n.r.
Cs	55	1.3	1.2	0.33 – 9	16	39	0.99	0.12	0.77 – 1.2	10	45	1.02	0.19	n.r.
Dy	45	5.7	1.2	3.2 – 10	8	37	5.6	0.6	4.2 – 6.8	6	39	5.5	0.7	n.r.
Er	26	2.8	0.8	1.7 – 4.7	4	22	2.58	0.45	1.7 – 3.3	2	24	2.7	0.6	n.r.
Eu	82	1.4	3.5	1.4 – 3.5	12	70	2.35	0.22	1.95 – 2.9	13	69	2.34	0.21	n.r.
F (%)	51	0.36	0.06	0.194 – 0.6633	12	39	0.366	0.017	0.327 – 0.400	7	44	0.363	0.024	n.r.
Ga	57	21.5	4.3	10 – 35	6	51	21.7	2.8	15 – 27	4	53	21	3	n.r.
Gd	48	13.4	2.6	8 – 25	6	42	13.0	1.4	10.5 – 16	4	44	12.9	1.4	n.r.
Hf	44	15.2	1.9	9.7 – 18.7	4	40	15.6	1.4	12.9 – 18.7	3	41	15.5	1.4	n.r.
Hg (ppb)	33	26	19	1.5 – 106	14	19	16.2	2.7	11.01 – 21	4	29	22	10	n.r.
Ho	23	1.3	0.7	0.7 – 3.9	7	16	1.00	0.15	0.7 – 1.3	2	21	1.14	0.30	n.r.
Li	50	33	7	19.3 – 63	4	46	31.3	4.5	19.3 – 44	5	45	32	4	n.r.
Lu	55	0.23	0.06	0.14 – 0.44	4	51	0.216	0.041	0.14 – 0.31	5	50	0.214	0.039	n.r.
Nb	41	26	7	1.8 – 36	15	26	27.9	1.6	25.3 – 30	6	35	26	3	n.r.
Nd	80	196	28	99 – 300	8	72	196	16	164 – 234	7	73	196	17	n.r.
Pb	97	53	13	0.23 – 85	12	85	55	6	40 – 68.1	9	88	54	7	n.r.
Pr	31	53	11	20 – 82	6	25	54	5	44.4 – 61	3	28	52	6	n.r.
Sb	39	4	5	1.38 – 32	6	33	3.24	0.35	2.42– 4	3	36	3.2	0.4	n.r.
Sc	70	7.2	3.6	4 – 30	12	58	6.2	0.7	4 – 7.9	6	64	6.4	1.0	n.r.
Sm	84	26.2	3.3	15 – 35	5	79	26.2	2.7	20 – 82	6	76	26	2	n.r.
Sn	33	14	43	2 – 253	2	31	6.5	1.9	2 – 10	3	30	6.7	1.7	n.r.
Ta	42	1.08	0.36	0.63 – 2.2	8	34	0.94	0.16	0.63 – 1.26	5	37	0.97	0.19	n.r.
Tb	60	1.4	0.5	0.4 – 3.9	8	52	1.37	0.18	1.03 – 1.8	8	52	1.34	0.17	n.r.
Th	93	105	16	22.65 – 144	16	77	106	5	95 – 120	10	83	106	6	n.r.
Tl	30	1.41	0.32	0.71 – 2.1	0	30	1.41	0.32	0.71 – 2.1	2	28	1.4	0.3	n.r.
Tm	26	0.6	0.6	0.1 – 2.9	4	22	0.35	0.13	0.1 – 0.63	6	20	0.34	0.11	n.r.
U	63	2.2	0.6	0.1 – 2.9	7	56	2.22	0.30	0.1 – 2.9	7	56	2.20	0.29	n.r.
V	90	53	11	17 – 101	8	82	53	7	36 – 70	7	83	53	7	n.r.
Y	70	30	9	8.4 – 59	13	57	28.3	4.4	20 – 41	12	58	28	5	n.r.
Yb	85	1.76	0.41	0.9 – 3.0	4	81	1.71	0.35	0.9 – 2.45	5	80	1.74	0.34	n.r.

For the individual data not reproduced here, see Gladney *et al.* (1992). *n*: number of data; *x*: mean; *s*: standard deviation; *R*: range; *o*: number of discordant outliers detected by a given method; the subscripts $_{in}$, $_f$ and $_l$ refer to the initial, final (after applying test N3–k=2, 3, and 4; see Table 1 and Tables A3-A5 in the electronic supplement), and 2*s* method (literature, Gladney *et al.*, 1992); the difference between n_{in} and n_f gives the number of discordant outliers detected by the discordance test N3 presented in this work; similarly, the difference between n_{in} and n_l gives the number of discordant outliers detected by the 2*s* method presented by Gladney *et al.* (1992). n.r.: not reported (by the original authors). [a] The first highly discordant outlier was detected using single-outlier test N1 (the application of the multiple-outlier test N3–k=2,3,4 appears to be adversely affected by the swamping effect of the nearest neighbor). The final statistical data calculated in this work were rounded following the criteria put forth by Verma (2005), whereas those from the literature as included as reported by the authors of the initial compilation.

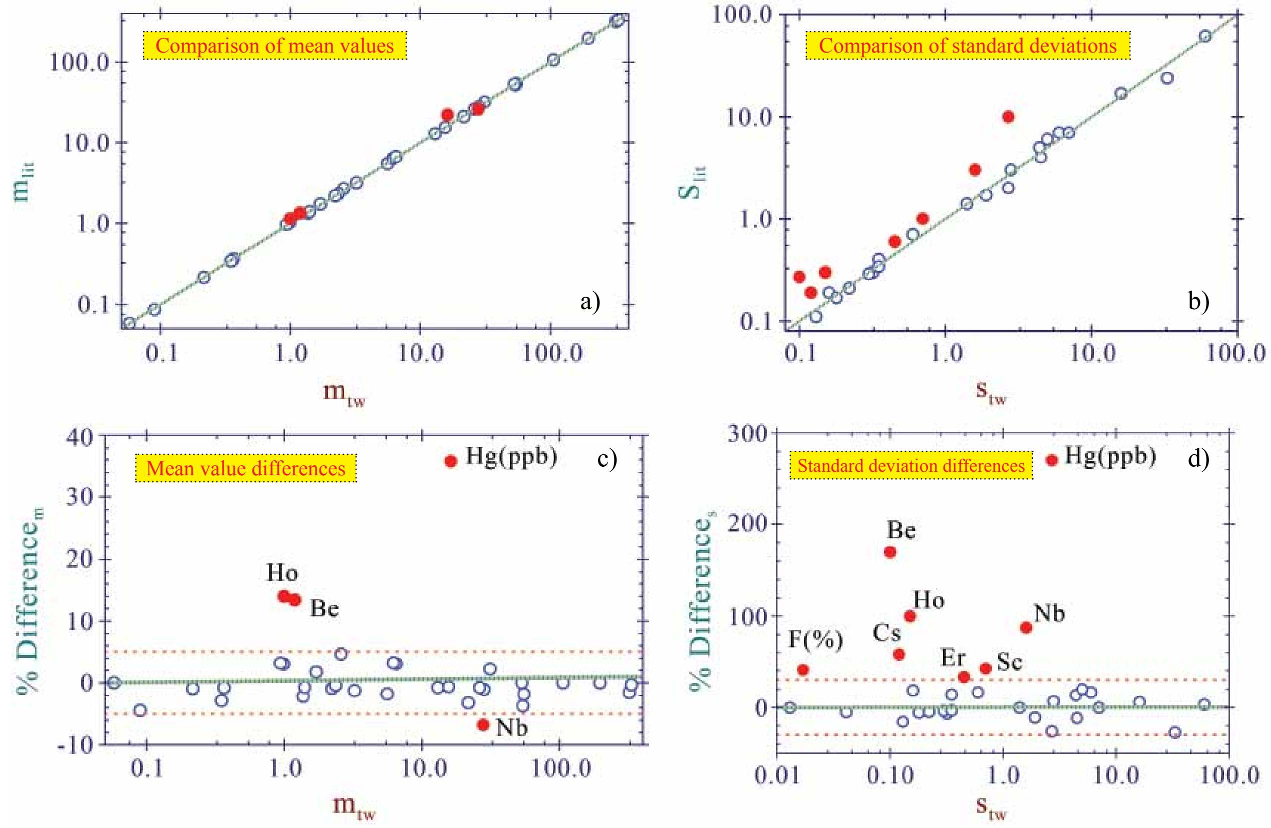

Figure 4. Comparison of the performance of 2*s* method with that of multiple-outlier test N3–k=2,3,4 at the 95% confidence level, using trace element data (with *n* between 20 and 100) for international geochemical reference material granodiorite GSP-1. The abbreviations used are: m: mean; s: standard deviation; $_{tw}$: this work; $_{lit}$: literature value (Gladney *et al.*, 1992). Open circles are for elements that show smaller differences for the two methods, whereas filled circles represent elements with large differences for them. The diagonal lines in Figures (a) and (b) are of equal values for mean and standard deviation data for the two methods, whereas the heavy dotted horizontal "zero value" lines in Figures (c) and (d) are for equal mean and standard deviation values for the two methods. The other horizontal lines in Figure (c) are for 5% differences between the two methods, whereas in (d) these are for 30% differences. The elements with larger differences are marked with element symbols. (a) Comparison of mean values obtained from 21 method by Gladney *et al.* (1992) with those obtained in this work using test N3–k=2,3,4; (b) Comparison of standard deviation values obtained from 2*s* method by Gladney *et al.* (1992) with those obtained in this work; (c) Comparison of mean values obtained from 2*s* method by Gladney *et al.* (1992) with those obtained in this work using test N3–k=2,3,4; the %Difference$_m$, *i.e.*, [100·(m_{lit}-m_{tw})/m_{tw}] being the % difference between the literature mean value (m_{lit}: mean value from Gladney *et al.*, 1992) and the present mean value (m_{tw}: mean value obtained in this work) for a given element; (d) Comparison of standard deviation values obtained from 2*s* method by Gladney *et al.* (1992) with those obtained in this work; the %Difference$_s$, *i.e.*, [100·(s_{lit}-s_{tw})/s_{tw}] being the % difference between the literature standard deviation value (s_{lit}: standard deviation value from Gladney *et al.*, 1992) and the present standard deviation value (s_{tw}: standard deviation value obtained in this work) for a given element.

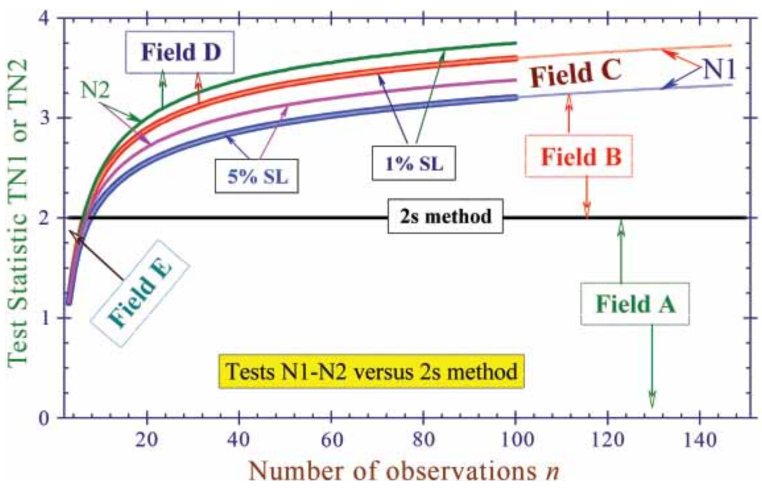

Figure 5. Test statistic TN1 or TN2 (see Table 1 and the text for more details) as a function of the number of observations (*n*). For test N1, 5% SL and 1% SL critical values are shown by blue and red dotted curves, respectively, whereas those for test N2 are in magenta and green solid curves. The N1 curves (blue dotted line curve for 5% SL and red dotted line curve for 1% SL) beyond *n* = 100, were constructed from the literature critical value data (Grubbs and Beck, 1972). The "two standard deviation" (2*s*) method is shown schematically by a horizontal thick solid black line. The fields A to E are as follows: Field A represents the field below the 2*s* horizontal line; Field B is the field below 5% SL curve; Field C is above 5% SL curve and below 1% SL curve; Field D is above the 1% SL curve; Field E is above any of these curves but below the 2*s* horizontal line. See text for the discussion on the erroneous 2*s* method.

tions below a given curve, falling in both fields A and B for 95% CL, or in fields A, B, and C for 99% CL, will be considered as legitimate data, whereas those falling above these curves (shown in blue for test N1 95%CL, in red for test N1 99%CL, in magenta for test N2 95%CL, and in green for test N2 99%CL) would be discordant outliers. Therefore, in most cases, legitimate observations (those falling in fields B or C) will be erroneously eliminated by the 2*s* method, whereas some (those falling in field E) will be erroneously retained by the 2*s* method. This reasoning combined with the problems mentioned earlier, clearly show that the 2*s* method should be abandoned in future for data handling of samples of finite sizes such as those encountered in most experimental work, including the study of international geochemical reference materials. Instead, our present multiple-test method should be adopted for these purposes.

CONCLUSIONS

We have used our established and well-tested Monte Carlo type simulation procedure for generating new, precise and accurate critical values for nine discordancy tests with 22 test variants for sample sizes up to 100. Occasionally, these values were found to be similar to (differences <0.4%), but more precise and accurate than the existing literature values; for most tests, however, larger differences (~0.4–20%) were observed between the two sets.

These new critical values will be of great use in many diverse fields of science and engineering. Three specific examples are presented to highlight the use of these new critical values as well as those recently published by Verma and Quiroz-Ruiz (2006).

The multiple-test method outlined in the present work seems to perform better than the box-and-whisker plot method used for processing inter-laboratory data on petroleum hydrocarbon compounds.

Finally, the frequently used "two standard deviation" (2*s*) method is shown to be statistically erroneous and less efficient in detection of outliers in comparison to our proposed multiple-test method, and should, therefore, be abandoned. The multiple-test method (15 tests with 33 variants) is shown to be highly suited for handling of experimental data, including those for international geochemical reference materials.

ACKNOWLEDGEMENTS

This research was partly supported by the Sistema Nacional de Investigadores (Mexico), through a scholarship to A. Quiroz-Ruiz as the first author's (SPV) "Ayudantes de Investigador Nacional Nivel 3". The revision of this manuscript benefited from constructive comments from two anonymous reviewers.

REFERENCES

Alberdi, M.T., Corona-M., E., 2005, Revisión de los gonfoterios en el Cenozoico tardío de México: Revista Mexicana de Ciencias Geológicas, 22(2), 246–260.

Ando, A., Mita, N., Terashima, S., 1987, 1986 values for fifteen GSJ rock reference samples, "Igneous rock series": Geostandards Newsletter, 11(1), 159-166.

Ando, A., Kamioka, H., Terashima, S., Itoh, S., 1989, 1988 values for GSJ rock reference samples, " Igneous rock series": Geochemical Journal, 23, 159-166.

Barnett, V., Lewis, T., 1994, Outliers in Statistical Data: Chichester, John Wiley, Third edition, 584 p.

Bartlett, J.M., Dougherty-Page, J.S., Harris, N.B.W., Hawkesworth, C.J., Santosh, M., 1998, The application of single zircon evaporation and model Nd ages to the interpretation of polymetamorphic terrains: an example from the Proterozoic mobile belt of south India: Contributions to Mineralogy and Petrology, 131(2-3), 181–195.

Batjes, N.H., 2005, Organic carbon stocks in the soils of Brazil: Soil Use and Management, 21(1), 22–24.

Bendre, S.M., Kale, B.K., 1987, Masking effect for outliers in normal samples: Biometrika, 74(4), 891-896.

Buckley, J.A., Georgianna, T.D., 2001, Analysis of statistical outliers with application to whole effluent toxicity testing: Water Environment Research, 73(5), 575–583.

Cooper S.J., Trinklein N.D., Anton E.D., Nguyen L., Myres R.M., 2006, Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome: Genome Research 16 (1), 1-10.

Dávalos-Álvarez, O.G., Nieto-Samaniego, A.F., Alaniz-Álvarez, S.A., Gómez-González, J.M., 2005, Las fases de deformación cenozoica en la región de Huimilpan, Querétaro, y su relación con la sismicidad local: Revista Mexicana de Ciencias Geológicas, 22(2), 129–147.

David, H.A., Paulson, A.S., 1965, The performance of several tests for outliers: Biometrika, 52, 429-436.

Dixon, W.J., 1950, Analysis of extreme values: Annals of Mathematical Statistics, 21(4), 488–506.

Dixon, W.J., 1951, Ratios involving extreme values: Annals of Mathematical Statistics, 22(1), 68–78.

Dixon, W.J., 1953, Processing data for outliers: Biometrics, 9(1), 74–89.

Dougherty-Page, J.S., Bartlett, J.M., 1999, New analytical procedures to increase the resolution of zircon geochronology by the evaporation technique: Chemical Geology, 153(1-4), 227–240.

Dybczyński, R., 1980, Comparison of the effectiveness of various procedures for the rejection of outlying results and assigning consensus values in interlaboratory programs involving determination of trace elements or radionuclides: Analytica Chimica Acta, 117(1), 53–70.

Dybczyński, R., Tugsavul, A., Suschny, O., 1979, Soil-5, a new IAEA certified reference material for trace element determinations: Geostandards Newsletter, 3(1), 61–87.

Dybczyński, R., Polkowska-Motrenko, H., Samczynski, Z., Szopa, Z., 1998, Virginia tobacco leaves (CTA-VTL-2) - new Polish CRM for inorganic trace analysis including microanalysis: Fresenius Journal of Analytical Chemistry, 360(3-4), 384–387.

Esquivel-Macías C., León-Olvera, R.G., Flores-Castro, K., 2005, Caracterización de una nueva localidad fosilífera del Jurásico Inferior con crinoides y amonites en el centro-oriente de México: Revista Mexicana de Ciencias Geológicas, 22(1), 97–114.

ESMP, 2006, Detection and Accommodation of Outliers in Normally Distributed Data Sets (online) by Fallon A., Spada, C.: Environmental Sampling and Monitoring Primer, <http://ewr.cee.vt.edu/environmental/teach/smprimer/outlier/outlier.html>.

Freeman, B. D., Quezado, Z., Zeni, F., Natanson, C., Danner, R.L., Banks, S., Quezado, M., Fitz, Y., Bacher, J., Eichacker, P.Q., 1997, rG-CSF reduces endotoxemia and improves survival during E-coli pneumonia: Journal of Applied Physiology, 83(5), 1467–1475.

Gill, U., Covaci, A., Ryan, J.J., Emond, A., 2004, Determination of persistent organohelogenated pollutants in human hair reference material (BCR 397); an interlaboratory study: Analytical and Bioanalytical Chemistry, 380(7-8), 924–929.

Gladney, E.S., Roelandts, I., 1988, 1987 compilation of elemental concentration data for USGS BIR-1, DNC-1 and W-2: Geostandards Newsletter, 12(1), 63–118.

Gladney, E.S., Roelandts, I., 1990, 1988 compilation of elemental concentration data for USGS geochemical exploration reference materials GXR-1 to GXR-6: Geostandards Newsletter, 14(1), 21–118.

Gladney, E.S., Jones, E.A., Nickell, E.J., 1992, 1988 compilation of elemental concentration data for USGS AGV-1, GSP-1 and G-2: Geostandards Newsletter, 16(2), 111–300.

Gladney, E.S., Jones, E.A., Nickell, E.J., Roelandts, I., 1991, 1988 compilation of elemental concentration data for USGS DTS-1, G-1, PCC-1, and W-1: Geostandards Newsletter, 15(2), 199–396.

Graybeal, D.Y., DeGaetano, A.T., Eggleston, K.L., 2004, Improved quality assurance for historical hourly temperature and humidity; development and application to environmental analysis: Journal of Applied Meteorology, 43(11), 1722–1735.

Grubbs, F.E., 1950, Sample criteria for testing outlying observations: Annals of Mathematical Statistics, 21(1), 27-58.

Grubbs, F.E., 1969, Procedures for detecting outlying observations in samples: Technometrics, 11(1), 1-21.

Grubbs, F.E., Beck, G., 1972, Extension of sample sizes and percentage points for significance tests of outlying observations: Technometrics, 14(4), 847–854.

GSJ-GRS, 2006, GSJ Geochemical Reference samples DataBase (online): National Institute of Advanced Industial Science and Technology, Geochemical Standards Database, <www.aist.go.jp/RIODB/geostand/>, accessed on April 6, 2006.

Guevara, M., Verma, S.P., Velasco-Tapia, F., 2001, Evaluation of GSJ intrusive rocks JG1, JG2, JG3, JG1a, and JGb1: Revista Mexicana de Ciencias Geológicas, 18(1), 74–88.

Guevara, M., Verma, S.P., Velasco-Tapia, F., Lozano-Santa Cruz, R., Girón, P., 2005, Comparison of linear regression models for quantitative geochemical analysis; Example of X-ray fluorescence spectrometry: Geostandards and Geoanalytical Research, 29(3), 271-284.

Hanson, D., Kotuby-Amacher, J., Miller, R.O., 1998, Soil analysis; Western States proficiency testing program for 1996: Fresenius Journal of Analytical Chemistry, 360(3-4), 348–350.

Harcourt, A.H., Coppeto, S.A., Parks, S.A., 2005, The distribution-abundance (density) relationship; its form and causes in a tropical mammal order, Primates: Journal of Biogeography, 32(4), 565–579.

Hawkins, D.M., 1979, Fractiles of an extended multiple outlier test: Journal of Statistical and Computational Simulations, 8, 227-236.

Hayes, K, Kinsella, T., 2003, Spurious and non-spurious power in performance criteria for tests of discordancy: Journal of the Royal Statistical Society, Series D (The Statistician), 52 (1), 69-82.

Hofer, J.D., Murphy, J.R., 2000, Structured use of the median in the analytical measurement process: Journal of Pharmaceutical and Biomedical Analysis, 23(4), 671–686.

Holcombe, G., Lawn, R., Sargent, M., 2004, Improvements in efficiency of production and traceability for certification of reference materials: Accreditation and Quality Assurance, 9(4-5), 198–204.

Ifrim, C., Stinnesbeck, W., Schafhauser, A., 2005, Maastrichtian shallow-water ammonites of northwestern Mexico: Revista Mexicana de Ciencias Geológicas, 22(1), 48–64.

Ihnat, M., 2000, Performance of NAA methods in an international interlaboratory reference material characterization campaign: Journal of Radioanalytical and Nuclear Chemistry, 245(1), 73–80.

Iglewicz, B., Martinez, J., 1982, Outlier detection using robust measures of scale: Journal of Statistical Computation and Simulation, 15, 285-293.

Imai, N., Terashima, S., Itoh, S., Ando, A., 1995, 1994 compilation of analytical data for minor and trace elements in seventeen GSJ geochemical reference samples, "Igneous rock series": Geostandards Newsletter, 19(2), 135–213.

Imai, N., H., S., Terashima, S., Itoh, S., Ando, A., 1996, 1996 Compilation of analytical data on nine GSJ geochemical reference samples, "sedimentary rock series": Geostandards Newsletter, 20 (2), 165-216.

In't Veld, P.H., 1998, The use of reference materials in quality assurance programmes in food microbiology laboratories: International Journal of Food Microbiology, 45(1), 35–41.

Itoh, S., Terashima, S., Imai, N., Kamioka, H., Mita, N., Ando, A., 1993, 1992 compilation of analytical data for rare-earth elements, scandium, yttrium, zirconium and hafnium: Geostandards Newsletter, 17(1), 5-79.

Jakubowska, M., Kubiak, W.W., 2005, Removing spikes from voltammetric curves in the presence of random noise: Electroanalysis, 17(18), 1687-1694.

JGRMW, 2006, Japan Geochemical Reference Material Website, www.aist.go.jp/RIODB/Geostand/; accessed on April 6, 2006.

Kasper-Zubillaga, J.J., Carranza-Edwards, A., 2005, Grain size discrimination between sands of desert and coastal dunes from northwestern Mexico: Revista Mexicana de Ciencias Geológicas, 22(3), 383-390.

Kern, M., Preimesberger, T., Allesch, M., Pail, R., Bouman, J., Koop, R., 2005, Outlier detection algorithms and their performance in GOCE gravity field processing: Journal of Geodesy, 78(9), 509–519.

King, E.P., 1953, On some procedures for the rejection of suspected data: Journal of American Statistical Association, 48(263), 531-533.

Langton, S.D., Chevennement, R., Nagelkerke, N., Lombard, B., 2002, Analysing collaborative trials for qualitative microbiological methods; accordance and concordance: International Journal of Food Microbiology, 79(3), 175–181.

Lin, Z., Inn, K.G.W., Filliben, J.J., 2001, An alternative statistical approach for interlaboratory comparison data evaluation: Journal of Radioanalytical and Nuclear Chemistry, 248(1), 163–173.

Linkosalo, T., Hakkinen, R., Hari, P., 1996, Improving the reliability of a combined phenological time series by analyzing observation quality: Tree Physiology, 16(7), 661–664.

Lozano, R., Bernal, J.P., 2005, Assessment of eight new geochemical reference materials for XRF major and trace element analysis: Revista Mexicana de Ciencias Geológicas, 22(3), 329-344.

Luedeling, E., Nagieb, M., Wichern, F., Brandt, M., Deurer, M., Buerkert, A., 2005, Drainage, salt leaching and physico-chemical properties of irrigated man-made terrace soils in a mountain oasis of northern Oman: Geoderma, 125(3-4), 273–285.

Lugo-Ospina, A., Dao, T.H., Van Kessel, J.A., Reeves III, J.B., 2005, Evaluation of quick tests for phosphorus determination in dairy manures: Environmental Pollution, 135(1), 155–162.

McMillan, R.G., 1971, Tests for one or two outliers in normal samples with unknown variance: Technometrics, 13, 87-100.

Mendoza-Amézquita, E., Armienta-Hernández, M.A., Ayora, C., Soler, A., Ramos-Ramírez, E., 2006, Potencial lixiviación de elementos traza en jales de las minas La Asunción y Las Torres, en el Distrito Minero de Guanajuato, México: Revista Mexicana de Ciencias Geológicas, 23(1), 75–83.

Morabito, R., Massanisso, P., Cámara, C., Larsson, T., Frech, W., Kramer, K.J.M., Bianchi, M., Muntau, H., Donard, O.F.X., Lobinski, R., McSheehy, S., Pannier, F., Potin-Gautier, M., Gawlik, B.M., Bowadt, S., Quevauviller, P., 2004, Towards a new certified reference material for butyltins, methylmercury and arsenobetaine in oyster tissue: Trends in Analytical Chemistry, 23(9), 664–676.

Moran, M.A., McMillan, R.G., 1973, Tests for one or two outliers in normal samples with unknown variance; a correction: Technometrics, 15(3), 637-640.

Morán-Zenteno, D.J., Alba-Aldave, L.A., Martínez-Serrano, R.G., Reyes-Salas, M.A., Corona-Esquivel, R., Angeles-García, S., 1998, Stratigraphy, geochemistry and tectonic significance of the Tertiary volcanic sequences of the Taxco-Quetzalapa region, southern Mexico: Revista Mexicana de Ciencias Geológicas, 15(2), 167–180.

Otto, M., 1999, Chemometrics. Statistics and Computer Application in Analytical Chemistry: Weinheim, Wiley-VCH, 314 p.

Patriarca, M., Chiodo, F., Castelli, M., Corsetti, F., Menditto, A., 2005, Twenty years of the Me.Tos. project; an Italian national external quality assessment scheme for trace elements in biological fluids: Microchemical Journal, 79(1-2), 337–340.

Pearson, E.S., Hartley, H.O., 1976, Biometrika Tables for Statisticians, v. I: England, Biometrika Trust, third edition, 263 p.

Prescott, P., 1978, Examination of the behaviour of tests for outliers when more than one outlier is present: Applied Statistics, 27(1), 10-25.

Prescott, P., 1979, Critical values for a sequential test for many outliers: Applied Statistics, 28(1), 36-39.

Quesenberry, C.P., David, H.A., 1961, Some tests for outliers: Biometrika, 48, 379-387.

Ramos-Arroyo, Y.R., Siebe-Grabach, C.D., 2006, Estrategia para identificar jales con potencial de riesgo ambiental en un distrito minero; estudio de caso en el Distrito de Guanajuato, Mexico: Revista Mexicana de Ciencias Geológicas, 23(1), 54–74.

Reed, D.S., Smoll, J., Gibbs, P., Little, S.F., 2002, Mapping of antibody responses to the protective antigen of Bacillus anthracis by flow cytometric analysis: Cytometry, 49(1), 1–7.

Rodríguez, S.R., 2005, Geology of Las Cumbres volcanic complex, Puebla and Veracruz states, México: Revista Mexicana de Ciencias Geológicas, 22(2), 181–198.

Rosner, B., 1975, On the detection of many outliers: Technometrics, 17, 221-227.

Santoyo, E., Guevara, M., Verma, S.P., 2006, Determination of lanthanides in international geochemical reference materials by reversed-phase high performance liquid chromatography; An application of error propagation theory to estimate total analysis uncertainties: Journal of Chromatography A, 1118(1), 73-81.

Schaber, J., Badeck, F. W., 2002, Evaluation of methods for the combination of phenological time series and outlier detection: Tree Physiology, 22(14), 973–982.

SCNPHT, 2006, Statistics for Chemists, [Non-parametric Hypothesis Tests] (online): <http://www.webchem.science.ru.nl/cgi-bin/Stat/HypT/nphypt.pl>.

Sevransky, J., Vandivier, R.W., Gerstenberger, E., Correa, R., Ferantz, V., Banks, S.M., Danner, R.L., Eichacker, P.Q., Natanson, C., 2005, Prophylactic high-dose N$^{\omega}$-monomethyl-L-arginine prevents the late cardiac dysfunction associated with lethal tumor necrosis factor-α challenge in dogs: Shock, 23(3), 281–288.

Shapiro, S.S., Wilk, M.B., Chen, H.J., 1968, A compartive study of various tests for normality: Journal of American Statistical Association, 63, 1343-1371.

Shoemaker, D.P., Garland, C.W., Nibler, J.W., 1996, Experiments in Physical Chemistry. Sixth edition: New York, McGraw Hill, 778 p.

Sieber, J., Broton, D., Fales, C., Leigh, S., MacDonald, B., Marlow, A., Nettles, S., Yen, J., 2002, Standards reference materials for cements: Cement and Concrete Research, 32(12), 1899–1906.

Srivastava, M.S., 1980, Effect of equicorrelation in detecting a spurious observation: Canadian Journal of Statistics, 8(2), 249-251.

Stancak, A., Hoechstetter, K., Tintera, J., Vrana, J., Rachmanova, R., Kralik, J., Scherg, M., 2002, Source activity in the human secondary somatosensory cortex depends on the size of corpus callosum: Brain Research, 936(1-2), 47–57.

Statistica for Windows, 1998, General Convention and Statistics I. Second edition: Tulsa, OK, Statsoft, 1878 p.

Stefansky, W., 1971, Rejecting outliers by maximum normed residual: Annals of Mathematical Statistics, 42(1), 35-45.

Stevens, R.J., O'Bric, C.J., Carton, O.T., 1995, Estimating nutrient content of animal slurries using electrical conductivity: Journal of Agricultural Science, 125(2), 233–238.

Stoch, H., Steele, T.W., 1978, Analyses, by several laboratories, of three ferromanganese slags: Network of Independent Monitors (NIM) South Africa Report 1965, 34 p.

Suhren G., Walte H.G., 2001, Determination of precision data of the bactoscan FC-method by an interlaboratory study: Kieler Milchwirtschaftliche Forschungsberichte 53(4): 269-282.

Taylor, B.J., 2000, A statistical analysis of the metallicities of nine old superclusters and moving groups: Astronomy and Astrophysics, 362, 563–579.

Thomulka, K.W., Lange, J.H., 1996, A mixture toxicity study employing combinations of tributyltin chloride, dibytyltin dichloride, and tin chloride using the marine bacterium vibrio harveyi as the test organism: Ecotoxicology and Environmental Safety, 34(1), 76–84.

Tietjen, G.L., Moore, R.H., 1972, Some Grubbs-type statistics for the detection of several outliers: Technometrics, 14(3), 583-597.

Tietjen, G.L., Moore, R.H., 1979, Corrigendum "Some Grubbs-type statistics for the detection of several outliers": Technometrics, 21(3), 396.

Tigges, M., Iuvone, P.M., Fernández, A., Sugrue, M.F., Mallorga, P.J., Laties, A.M., Stone, R.A., 1999, Effects of muscarinic cholinergic receptor antagonists on postnatal eye growth of rhesus monkeys: Optometry and Vision Science, 76(6), 397–407.

Tiku, M.L., 1975, A new statistic for testing suspected outliers: Communications in Statistics, 4(8), 737-752.

Tiku, M.L., 1977, Rejoinder; Comment on 'A new statistic for testing suspected outliers': Communications in Statistics, Theory and Methods, A6(14), 1417-1422.

Torres-Alvarado, I.S., 2002, Chemical equilibrium in hydrothermal systems; the case of Los Azufres geothermal field, Mexico: International Geology Review, 44(7), 639-652.

Treviño-Cázares, A., Ramírez-Fernández, J.A., Velasco-Tapia, F., Rodríguez-Saavedra, P., 2005, Mantle xenoliths and their host magmas in the Eastern Alkaline Province (NE Mexico): International Geology Review, 47(12), 1260-1286.

Vattuone, M.E., Latorre, C.O., Leal, P.R., 2005, Polimetamorfismo de muy bajo a bajo grado en rocas volcánicas jurásico-cretácicas al sur de Cholila, Chubut, Patagonia Argentina: Revista Mexicana de Ciencias Geológicas, 22(3), 315–328.

Velasco, F., Verma, S.P., Guevara, M., 2000, Comparison of the performance of fourteen statistical tests for detection of outlying values in geochemical reference material databases: Mathematical Geology, 32(4), 439–464.

Velasco-Tapia, F., Guevara, M., Verma, S.P., 2001, Evaluation of concentration data in geochemical reference materials: Chemie der Erde, 61(1), 69–91.

Verma, S.P., 1997, Sixteen statistical tests for outlier detection and rejection in evaluation of International Geochemical Reference Materials; example of microgabbro PM-S: Geostandards Newsletter, Journal of Geostandards and Geoanalysis, 21(1), 59–75.

Verma, S.P., 1998, Improved concentration data in two international geochemical reference materials, USGS basalt BIR-1 and GSJ peridotite JP-1) by outlier rejection: Geofísica Internacional, 37(3), 215–250.

Verma, S.P., 2005, Estadística Básica para el Manejo de Datos Experimentales; Aplicación en la Geoquímica (Geoquimiometría): México, D.F., Universidad Nacional Autónoma de México, 186 p.

Verma, S.P., 2006, Extension-related origin of magmas from a garnet-bearing source in the Los Tuxtlas volcanic field, Mexico: International Journal of Earth Sciences (Geologische Rundschau), 95(5), 871-901.

Verma, S.P., Quiroz-Ruiz, A., 2006, Critical values for six Dixon tests for outliers in normal samples up to sizes 100, and applications in science and engineering: Revista Mexicana de Ciencias Geológicas, 23(2), 133-161.

Verma, S.P., Orduña-Galván, L.J., Guevara, M., 1998, SIPVADE; A new computer programme with seventeen statistical tests for outlier detection in evaluation of international geochemical reference materials and its application to Whin Sill dolerite WS-E from England and Soil-5 from Peru: Geostandards Newsletter, Journal of Geostandards and Geoanalysis, 22(2), 209–234.

Verma, S.P., Pandarinath, K., Santoyo, E., González-Partida, E., Torres-Alvarado, I.S., Tello, E., 2006, Fluid chemistry and temperatures prior to exploitation at the Las Tres Vírgenes geothermal field, Mexico: Geothermics, 35(2), 156-180.

Villaseñor, A.B., González-León, C.M., Lawton, T.F., Aberhan, M., 2005, Upper Jurassic ammonites and bivalves from Cucurpe Formation, Sonora (Mexico): Revista Mexicana de Ciencias Geológicas, 22(1), 65–87.

Villeneuve, J.-P., de Mora, S.J., Cattini, C., 2002, World-wide and regional intercomparison for the determination of organochlorine compounds and petroleum hydrocarbons in sediment sample IAEA-417: Vienna, Austria, International Atomic Energy Agency, Analytical Quality Control Services, 136 p.

Villeneuve, J.-P., de Mora, S., Cattini, C., 2004, Determination of organochlorinated compounds and petroleum in fish-homogenate sample IAEA-406; results from a worldwide interlaboratory study: Trends in Analytical Chemistry, 23(7), 501–510.

Wang, X.-D., Söderlund, U., Lindh, A., Johansson, L., 1998, U-Pb and Sm-Nd dating of high-pressure granulite- and upper amphibolite facies rocks from SW Sweden: Precambrian Research, 92(4), 319–339.

Woitge, H.W., Scheidt-Nave, C., Kissling, C. Leidig-Bruckner, G., Meyer, K., Grauer, A., Scharla, S.H., Ziegler, R., Seibel, M.J., 1998, Seasonal variation of biochemical indexes of bone turnover; Results of a population-based study: Journal of Clinical and Endocrinological Metabolism, 83(1), 68–75.

WORM Database, 2006, < http://www.wormbase.org/>

Yurewicz, K.L., 2004, A growth/mortality trade-off in larval salamanders and the coexistence of intraguild predators and prey: Oecologia, 138(1), 102–111.

Zaric S., Niketic S.R., 1997, The anisotropic π-effect of the nitro group in ammine-nitro cobalt(III) complexes: Polyhedron, 16(20), 3565–3569.

Zhang, J, 1998, Tests for multiple upper or lower outliers in an exponential sample: Journal of Applied Statistics, 25(2), 245-255.

Zhang, J, Wang, X., 1998, Unmasking test for multiple upper or lower outliers in normal samples: Journal of Applied Statistics, 25(2), 257-261.

Manuscript received: March 18, 2006
Corrected manuscript received: August 9, 2006
Manuscript accepted: August 23, 2006