

Critical values for 33 discordancy test variants for outliers in normal samples of very large sizes from 1,000 to 30,000 and evaluation of different regression models for the interpolation and extrapolation of critical values

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, Temixco 62580, Mexico.

* spv@cie.unam.mx

ABSTRACT

In this final paper of a series of four, using our well-tested simulation procedure we report new, precise, and accurate critical values or percentage points (with four to eight decimal places) of 15 discordancy tests with 33 test variants, and each with seven significance levels $\alpha = 0.30, 0.20, 0.10, 0.05, 0.02, 0.01, \text{ and } 0.005$, for normal samples of very large sizes n from 1,000 to 30,000, viz., 1,000(50) 1,500(100) 2,000(500) 5,000(1,000) 10,000(10,000) 30,000, i.e., 1,000 (steps of 50) 1,500 (steps of 100) 2,000 (steps of 500) 5,000 (steps of 1,000) 10,000 (steps of 10,000) 30,000. The standard error of the mean is also reported explicitly and individually for each critical value. As a result, the applicability of these discordancy tests is now extended to practically all sample sizes (up to 30,000 observations or even greater). This final set of critical values for very large sample sizes would cover any present or future needs for the application of these discordancy tests in all fields of science and engineering. Because the critical values were simulated for only a few sample sizes between 1,000 and 30,000, six different regression models were evaluated for the interpolation and extrapolation purposes, and a combined natural logarithm-cubic model was shown to be the most appropriate. This is the first time in the literature that a log-transformation of the sample size n before a polynomial fit is shown to perform better than the conventional linear to polynomial regressions hitherto used. We also use 1,402 unpublished datasets from quantitative proteomics to show that our multiple-test method works more efficiently than the MAD_Z robust outlier method used for processing these data and to illustrate thus the usefulness of our final work on these lines.

Key words: outlier methods, normal sample, Monte Carlo simulations, critical value tables, Dixon tests, Grubbs tests, skewness, kurtosis, statistics, regression equations, log-transformation, proteomics.

RESUMEN

En este trabajo final de una serie de cuatro, usando nuestro procedimiento de simulación bien establecido reportamos nuevos valores críticos o puntos porcentuales, precisos y exactos (con cuatro a ocho puntos decimales) de 15 pruebas de discordancia con 33 variantes y cada uno con siete niveles de significancia $\alpha = 0.30, 0.20, 0.10, 0.05, 0.02, 0.01 \text{ y } 0.005$, para muestras normales de tamaños muy grandes n de 1,000 a 30,000, viz., 1,000(50) 1,500(100) 2,000(500) 5,000(1,000) 10,000(10,000) 30,000, esto es, 1,000 (pasos de 50) 1,500 (pasos de 100) 2,000 (pasos de 500) 5,000 (pasos de 1,000) 10,000 (pasos de 10,000) 30,000. Se reporta también el error estándar de la media en forma explícita e individual para cada valor crítico. Como consecuencia, la aplicabilidad de estas pruebas de discordancia ha sido extendida a prácticamente cualquier tamaño de muestra estadística (hasta 30,000 observaciones o aún

mayores). Este conjunto final de valores críticos para tamaños muy grandes cubrirá cualquier necesidad presente o futura de aplicación de estas pruebas de discordancia en todos los campos de las ciencias e ingenierías. Dado que los valores críticos fueron simulados para pocos tamaños de muestra entre 1,000 y 30,000, seis modelos de regresión diferentes fueron evaluados para la interpolación y extrapolación de los datos y se demostró que un modelo combinado de logaritmo natural-cúbico es el más apropiado. Es la primera vez en la literatura mundial que se demuestra que una transformación logarítmica del tamaño de muestra n antes de un ajuste polinomial resulta mejor que los ajustes convencionales desde lineal hasta polinomial de tercer grado usados a la fecha. Finalmente, usamos 1,402 conjuntos de datos de la proteómica cuantitativa con el fin de demostrar que nuestro método de pruebas múltiples funciona más eficientemente que el método robusto MAD_Z usado para procesar estos datos y, de esta manera, ilustrar la utilidad de nuestro trabajo final en estas líneas.

Palabras clave: métodos de valores desviados, muestra normal, simulaciones Monte Carlo, tablas de valores críticos, pruebas de Dixon, pruebas de Grubbs, sesgo, curtosis, estadística, ecuaciones de regresión, transformación-log, proteómica.

INTRODUCTION

Three recent papers (Verma and Quiroz-Ruiz, 2006a, 2006b; Verma *et al.*, 2008a) have reported a highly precise and accurate Monte Carlo type simulation procedure for $N(0,1)$ random normal variates and presented new, precise, and accurate critical values for 7 significance levels $\alpha = 0.30, 0.20, 0.10, 0.05, 0.02, 0.01,$ and 0.005 , and for sample sizes n up to 1,000 for 15 discordancy tests with 33 variants. These tests were summarized by Verma *et al.* (2008a) and therefore will not be repeated here. For greater $n (>1,000)$, practically no critical values are available in the literature for any of these tests (Barnett and Lewis, 1994; Verma, 2005).

It may be pointed out that the critical values simulated by Verma and Quiroz-Ruiz (2006a, 2006b) and by Verma *et al.* (2008a) are for testing the discordancy of outliers in *normal* statistical samples under the assumption of some kind of a contamination model (see Barnett and Lewis, 1994 and Verma *et al.*, 2008b for details on the possible contamination models). The outliers are simply extreme observations, irrespective of their discordancy, for example, an upper outlier $x_{(n)}$ or a lower outlier $x_{(1)}$ in an ordered sample array of n observations or data $x_{(1)}, x_{(2)}, x_{(3)}, x_{(n-2)}, x_{(n-1)}, x_{(n)}$. In an “uncontaminated” normal sample these outlying observations will ideally not be discordant whereas in a “contaminated” normal sample they are likely to be identified as discordant. A “statistical” sample (without any assumption for the population from which it was drawn) can actually come from any distribution such as a beta or a gamma distribution and not *necessarily* from a normal distribution. Only under the assumption that a statistical sample was drawn from a normal population and was probably contaminated in some way, it is true that the outliers in this sample should be tested using the discordancy tests that have been especially designed for normal samples (Barnett and Lewis, 1994). As an example, a statistical sample of experimental data (such as geochemical data) is most likely drawn from a normal population (Verma, 2005), in which the outliers can be tested as discordant (or not discordant)

using the discordancy tests along with the critical values for *normal* samples (*e.g.*, Verma and Quiroz-Ruiz, 2006a, 2006b; Verma *et al.*, 2008a). For a statistical sample drawn from a different distribution the discordancy tests especially designed for *that* particular distribution along with the corresponding critical values, if available, will have to be used. Thus, the critical values for 33 discordancy test variants have been simulated for outliers in *normal* samples, with the possibility of their application for discordancy of outliers in *statistical* samples *assumed* to be drawn from a normal population.

In inter-laboratory analytical studies for quality control purposes, the number of individual data (n) for a given chemical element in a reference material (RM) seldom exceeds 1,000, but this might become more common in future. In those cases, at present the multiple-test method (see Verma *et al.*, 2008a and references therein) is not likely to be appropriately applicable due to the unavailability of precise critical values for $n > 1,000$ for these discordancy tests. New critical values could therefore be proposed for $n > 1,000$ through an adequate statistical methodology. Requirements of critical values for very large $n (>1,000)$ also exist in an altogether different field of molecular and cellular proteomics (Murray Hackett, written communication, June 2007 and February 2008).

For the present work, we have included most discordancy tests for normal univariate samples (15 tests with 33 test variants; see Table 1 in Verma *et al.*, 2008a) for simulating new, precise, and accurate critical values for the same 7 significance levels ($\alpha = 0.30$ to 0.005) and for very large sample sizes n , *viz.*, 1,000(50)1,500(100)2,000(500)5,000(1,000)10,000(10,000)30,000, using a highly precise and accurate simulation procedure described earlier (Verma and Quiroz-Ruiz, 2006a, 2006b; Verma *et al.*, 2008a). The above is a rather standard nomenclature to express the availability of critical values (see, for example, Barnett and Lewis, 1994) and has been used by us in the past (*e.g.*, Verma and Quiroz-Ruiz, 2006a, 2006b). As an example, the expression “1,000(50)1,500” actually means

that the critical values were simulated for the sample sizes of 1,000 to 1,500, with the sample size steps of 50, *i.e.*, for $n = 1,000; 1,050; 1,100; 1,150; 1,200; 1,250; 1,300; 1,350; 1,400; 1,450; \text{ and } 1,500$. Therefore, our present simulation was for the sample sizes of: 1,000 (steps of 50) 1,500 (steps of 100) 2,000 (steps of 500) 5,000 (steps of 1,000) 10,000 (steps of 10,000) 30,000.

The importance of our present work resides in the fact that to date, precise critical values are available for sample sizes only up to 1,000 (Verma and Quiroz-Ruiz, 2006a, 2006b; Verma *et al.*, 2008a). Even the sophisticated regression equations obtained by the artificial neural network (ANN) methodology presented by Verma *et al.* (2008a) are *only* valid for interpolation purposes, *i.e.*, for sample sizes $n \leq 1,000$, and are *not* recommended to be used for extrapolation purposes, *i.e.*, not for sample sizes $n > 1,000$, because the extrapolation is always a less accurate operation than the interpolation. Furthermore, no critical values for $n > 1,000$ were available to test the quality of these regression equations for extrapolation purposes. Therefore, the present work fulfills the gap of the still much needed critical values for very large sample sizes (1,000 to 30,000).

These results would be useful in all fields of science and engineering, especially in molecular and cellular proteomics and also for quality control purposes. For the first time, we evaluate six different regression models for the interpolation and extrapolation of critical values and show that the natural log-transformation (the \ln function) of sample size (n) combined with a polynomial fit provides the best regression model for interpolating (and also for extrapolating) the critical value (CV) data. We close our final paper in this series by demonstrating that the multiple-test method works more efficiently than the MAD_Z robust outlier method hitherto practiced for processing quantitative data on proteins.

DISCORDANCY TESTS AND THE SIMULATION PROCEDURE

Details on discordancy tests can be found in Barnett and Lewis (1994), Verma (2005), or the papers by Verma and Quiroz-Ruiz (2006a, 2006b) and Verma *et al.* (2008a). The 15 tests with their 33 variants for which critical values were simulated include the Dixon tests, the Grubbs tests, and the skewness and kurtosis tests (see tab. 1 in Verma *et al.*, 2008a).

Our highly precise and accurate Monte Carlo type simulation procedure has already been described in detail (Verma and Quiroz-Ruiz, 2006a, 2006b). The modifications reported by Verma *et al.* (2008a) to improve the precision of critical values were also incorporated in the present simulation. We may remind, however, that for very large sample sizes as many as 20 chains of 6×10^{10} random variates of high-quality (as judged from the tests summarized by Verma and Quiroz-Ruiz, 2006a) were generated.

RESULTS OF NEW CRITICAL VALUES (CV)

Both the standard error of the mean ($se_{\bar{x}}$) and the mean (\bar{x}) critical values for 33 discordancy test variants, for $n = 1,000(50)1,500(100)2,000(500)5,000(1,000)10,000(10,000)30,000$ and $\alpha = 0.30, 0.20, 0.10, 0.05, 0.02, 0.01,$ 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-A40 (40 tables in the electronic supplement; 20 odd-numbered tables for $se_{\bar{x}}$ and 20 even-numbered tables for \bar{x}). The present values cannot be compared with literature data because the latter are not available for such large sample sizes.

As for our earlier tables for sample sizes up to 1,000 (Verma and Quiroz-Ruiz, 2006a, 2006b; Verma *et al.*, 2008a), these new critical value data, along with their individual uncertainty estimates, are available in other formats such as *txt*, *Excel*, or *Statistica*, on request from the authors (S.P. Verma, spv@cie.unam.mx, or A. Quiroz-Ruiz, aqr@cie.unam.mx). Similarly, the regression equations (see below) can also be obtained in a *doc* file with plain text format.

EVALUATION OF REGRESSION MODELS (n -CV and $\ln(n)$ -CV axes)

Six different regression models were fitted to obtain regression equations for the interpolation and extrapolation purposes. These include: three models of the (n -CV) type (linear; quadratic; and cubic) and three of the combined natural logarithm-transformed- n ($\ln(n)$ -CV) type (combined \ln -linear; combined \ln -quadratic; and combined \ln -cubic). Other models, such as those based on logarithm-base 10, are likely to provide similar results as the natural logarithm-transformed models, and therefore are not evaluated here. Similarly, more complex models based on the ANN methodology were already successfully applied by Verma *et al.* (2008a) and are also not evaluated in the present paper. A comparison of these ANN and more complex higher order models should be the subject of a separate paper.

For illustration purposes, we selected two powerful single-outlier tests (skewness N14 and kurtosis N15; see *e.g.*, Velasco and Verma, 1998; Velasco *et al.*, 2000; Verma, 2005). The results for the skewness test N14 are presented in Table 1, whereas those for the kurtosis test N15 are included in Table A41 (see electronic supplement). The 27 simulated critical values (n from 1,000 to 10,000; see Table A38 for N14 and Table A40 for N15 in the electronic supplement) for a given test variant and significance level (α) were used to obtain six different fitted equations for each significance level.

The quality of these fits is shown by the parameter (R^2) called the multiple-correlation coefficient (Bevington and Robinson, 2003). This is simply an extrapolation of the well-known concept of the linear-correlation coef-

ficient r , which characterizes the correlation between two variables at a time, to include multiple correlations, such as polynomial correlations, between groups of variables taken simultaneously. The parameter r is useful for testing whether one particular variable should be included in the theoretical function that is fitted to the data whereas the parameter R^2 characterizes the fit of the data to the entire function (Bevington and Robinson, 2003). Thus, a comparison of the R^2 for different functions is useful in optimizing the theoretical functional forms such as those evaluated in the present work.

To evaluate the equations for the interpolation and extrapolation purposes, the parameters $\{\sum(SIM-FitEq)^2\}_{int}$ and $\{\sum(SIM-FitEq)^2\}_{ext}$, respectively were computed and reported in Table 1 for test N14 and Table A41 for test N15. These equations can be used to compute the interpolated critical values for all n between 1,000 and 10,000 and if required, the extrapolated critical values for greater n , although the latter values will be, as expected, more approximate than the interpolated values.

However, to evaluate the six regression models and to decide which of the fitted equation would be better to use for the interpolation or extrapolation purposes, we have plotted the R^2 parameter (Figure 1a,b), the interpolation residuals (Figure 1c,d), and the extrapolation residuals (Figure 1e,f) for these six models. Similarly, the simulated critical values for these two tests (N14 and N15) along with the interpolated and extrapolated critical values for the simpler models (linear to cubic equations) are plotted in Figures 2a,b and 2c,d, respectively. The results for the natural log-transformed linear to cubic models are presented in Figures 2e,f and 2g,h, respectively.

From the examination of Figures 1 and 2 as well as of Tables 1 and A41, the following conclusions can be drawn:

(1) The simple linear model is probably the worst because this model provides the lowest R^2 values and the highest interpolation residuals (see the results for the Fitting Model L in Figure 1a-d), although the linear extrapolation residuals are smaller than for the quadratic and cubic models (compare L, Q and C models in Figure 1e,f). Linear models are sometimes employed to interpolate critical values, especially using only two simulated data for n adjacent to the missing critical value (e.g., Verma *et al.*, 1998). The values of linear correlation coefficients (r , which is equivalent to R for linear regressions), although statistically significant at the 99% confidence level (see Bevington and Robinson, 2003; Verma, 2005), are certainly relatively low ($R^2 < 1$). Further, strictly speaking the behavior of critical values as a function of n is certainly not linear (Figure 2a,b). This can be easily demonstrated by the application of proper statistical tests such as those practiced by Andaverde *et al.* (2005) for bottom hole temperature data.

(2) The more complex quadratic and cubic models used by some researchers (e.g., Rorabacher, 1991) are also not recommended because of relatively low R^2 (< 1) and

fairly large interpolation and extrapolation errors (see Figure 1c-f; also Figure 2a-d). In fact, these models provide totally unrealistic extrapolated critical values (Figure 2c,d).

(3) All linear to cubic models (without the log-transformation) are, therefore, inadequate for interpolation (Figure 2a,b) and absurd for extrapolation (Figure 2c-d) purposes.

(4) Undoubtedly, as judged by these criteria significantly better results are obtained with the combined natural logarithm-linear or polynomial models than the respective simple linear or polynomial models, and the best ones are for the combined logarithm-cubic model (compare the lnL, lnQ and lnC models respectively with the L, Q and C models in Figure 1a-f and Figure 2a-h). The R^2 approaches 1 (this being theoretically the maximum possible value for R^2 ; see Figure 1a,b), the interpolation errors are lower than those for the simpler models (Figure 1c,d), the extrapolation errors are the lowest (Figure 1e,f), and the interpolation and extrapolation curves better fit the simulated data (compare Figure 2c,d with Figure 2a,b for interpolation and Figure 2g,h with Figure 2c,d for extrapolation). In fact, the natural log-transformed cubic fitted equations (Tables 1 and A41) are the best for both interpolation (Figure 2e,f) and extrapolation (Figure 2g,h) purposes.

We emphasize that this is the first time that a natural logarithm-transformation of the sample size variable (n) has been proposed and its effects evaluated. This transformation is shown to perform much better in combination with the respective simpler linear or polynomial model. The latter simpler models without any log-transformation are hitherto practiced in the literature (Bugner and Rutledge, 1990; Rorabacher, 1991; Verma *et al.*, 1998). We suggest that, in future, the natural logarithm-cubic models (or of higher order terms) should be used routinely to obtain the interpolated or extrapolated critical values whenever required. The criterion should be to practically reach the theoretical value of 1 for R^2 by including all statistically meaningful terms in the natural log-transformed polynomial regression model.

APPLICATIONS IN SCIENCE AND ENGINEERING

The discordancy tests after extending their applicability to samples of sizes now up to 30,000 (or even greater), can be applied to practically any univariate, bivariate and even multivariate examples (for the latter two, after computing the studentized residuals) in all scientific or engineering fields. Similarly, for applications in studies related to quality control in Earth Sciences, such as those presented by Verma (2004), Guevara *et al.* (2001), and Velasco-Tapia *et al.* (2001), the new critical values and regression equations will certainly be useful and will cover all future needs. The examples such as Torres-Alvarado (2002), Colombo *et al.* (2007), Méndez-Ortiz *et al.* (2007), Nagarajan *et al.* (2007, 2008), Ramos-Leal *et al.* (2007), Salleh *et al.* (2007),

Table 1. Regression equations (six different models) fitted to 27 simulated critical values of test N14 of one extreme outlier (for n between 1,000 and 10,000; Table A38 from electronic supplement), and their evaluation for interpolation (for $1,000 > n > 10,000$) and extrapolation ($n > 10,000$) purposes (see Table A38 for those n for which critical values were simulated and for which interpolated or extrapolated critical values were required).

CL / SL / α	Type of fit	R^2 $\{\sum(SIM-FitEq)^2\}_{int}$ $\{\sum(SIM-FitEq)^2\}_{ext}$	Interpolation or extrapolation equation
70% / 30% / 0.30	L	0.844868; 3.22×10^4 ; 5.02×10^{-3} **	$[CV_{TN14}^{0.30}]_L = (0.0376 \pm 0.0011) - (3.09 \times 10^{-6} \pm 2.7 \times 10^{-7}) \cdot n$
	Q	0.968824; 6.52×10^{-5} ; 7.37×10^{-2}	$[CV_{TN14}^{0.30}]_Q = (0.0449 \pm 0.0009) - (8.2 \times 10^{-6} \pm 5 \times 10^{-7}) \cdot n$ $+ (5.2 \times 10^{-10} \pm 5 \times 10^{-11}) \cdot n^2$
	C	0.992071; 1.7×10^{-5} ; 1.57**	$[CV_{TN14}^{0.30}]_C = (0.0514 \pm 0.0009) - (1.47 \times 10^{-5} \pm 8 \times 10^{-7}) \cdot n$ $+ (2.07 \times 10^{-9} \pm 1.9 \times 10^{-10}) \cdot n^2 - (1.00 \times 10^{-13} \pm 1.2 \times 10^{-14}) \cdot n^3$
	lnL	0.981870; 3.76×10^{-5} ; 1.51×10^{-4}	$[CV_{TN14}^{0.30}]_{lnL} = (0.1215 \pm 0.0026) - (0.01205 \pm 0.00033) \cdot [\ln(n)]$
	lnQ	0.999815; 4.29×10^{-6} ; 1.16×10^{-5}	$[CV_{TN14}^{0.30}]_{lnQ} = (0.3064 \pm 0.0038) - (0.0587 \pm 0.0010) \cdot [\ln(n)]$ $+ (0.00291 \pm 0.00006) \cdot [\ln(n)]^2$
	lnC	0.999996; 9.89×10^{-6} ; 1.89×10^{-7}	$[CV_{TN14}^{0.30}]_{lnC} = (0.543 \pm 0.007) - (0.1477 \pm 0.0028) \cdot [\ln(n)]$ $+ (0.01405 \pm 0.00035) \cdot [\ln(n)]^2 - (0.000462 \pm 0.000015) \cdot [\ln(n)]^3$
80% / 20% / 0.20	L	0.844604; 8.33×10^4 ; 1.30×10^{-2} **	$[CV_{TN14}^{0.20}]_L = (0.0604 \pm 0.0017) - (4.97 \times 10^{-6} \pm 4.3 \times 10^{-7}) \cdot n$
	Q	0.968743; 1.68×10^{-4} ; 0.188	$[CV_{TN14}^{0.20}]_Q = (0.0721 \pm 0.0014) - (1.31 \times 10^{-5} \pm 9 \times 10^{-7}) \cdot n$ $+ (8.3 \times 10^{-10} \pm 9 \times 10^{-11}) \cdot n^2$
	C	0.992058; 4.39×10^{-5} ; 4.0**	$[CV_{TN14}^{0.20}]_C = (0.0826 \pm 0.0015) - (2.36 \times 10^{-5} \pm 1.4 \times 10^{-6}) \cdot n$ $+ (3.34 \times 10^{-9} \pm 3.1 \times 10^{-10}) \cdot n^2 - (1.61 \times 10^{-13} \pm 2.0 \times 10^{-14}) \cdot n^3$
	lnL	0.981772; 1.00×10^{-4} ; 4.11×10^{-4}	$[CV_{TN14}^{0.20}]_{lnL} = (0.1952 \pm 0.0041) - (0.0194 \pm 0.0005) \cdot [\ln(n)]$
	lnQ	0.999816; 4.04×10^{-6} ; 5.30×10^{-5}	$[CV_{TN14}^{0.20}]_{lnQ} = (0.493 \pm 0.006) - (0.0944 \pm 0.0015) \cdot [\ln(n)]$ $+ (0.00469 \pm 0.00010) \cdot [\ln(n)]^2$
	lnC	0.999998; 3.89×10^{-7} ; 3.01×10^{-6}	$[CV_{TN14}^{0.20}]_{lnC} = (0.873 \pm 0.009) - (0.2377 \pm 0.0034) \cdot [\ln(n)]$ $+ (0.02261 \pm 0.00043) \cdot [\ln(n)]^2 - (0.000744 \pm 0.000018) \cdot [\ln(n)]^3$

and Rodríguez-Ríos *et al.* (2007) were already pointed out in our earlier paper (Verma *et al.*, 2008a). Another recent example concerns the rainfall data from South India (Yadava *et al.*, 2007), for which the authors presented statistical inferences. This statistical treatment could have been certainly improved if the concepts and methodology presented in our earlier work (Verma, 1997, 2005; Verma and Quiroz-Ruiz, 2006a, 2006b; Verma *et al.*, 2008a) were followed. Similarly, these statistical principles can be (or have been) applied to the petrographic and geochemical data for sandstone samples from an Iranian oil field by Jafarzadeh and Hosseini-Barzi (2008), chemical data for minerals from Argentina by Montenegro and Vattuone (2008) and Vattuone

et al. (2008) and from Mexico by Vargas-Rodríguez *et al.* (2008), geothermal fluid chemistry data from Turkey by Palabiyik and Serpen (2008) and those compiled from all around the world by Díaz-González *et al.* (2008), nitrate pollution data in water samples from Jordan by Obeidat *et al.* (2008), and geochemical data for igneous rocks and minerals from the Mexican Volcanic Belt by Meriggi *et al.* (2008).

In summary, therefore, we emphasize, as in our earlier papers, that the multiple-test method originally proposed by Verma (1997) and exemplified in our four papers (Verma and Quiroz-Ruiz, 2006a, 2006b; Verma *et al.*, 2008a; this work) is a recommended procedure to process experimental

Table 1. (Cont.) Regression equations (six different models) fitted to 27 simulated critical values of test N14 of one extreme outlier (for n between 1,000 and 10,000; Table A38 from electronic supplement), and their evaluation for interpolation (for $1,000 > n > 10,000$) and extrapolation ($n > 10,000$) purposes (see Table A38 for those n for which critical values were simulated and for which interpolated or extrapolated critical values were required).

CL / SL / α	Type of fit	R^2 $\{\sum(SIM-FitEq)^2\}_{int}$ $\{\sum(SIM-FitEq)^2\}_{ext}$	Interpolation or extrapolation equation	
90% / 10% / 0.10	L	0.844266; 1.94x10 ⁻³ ; 3.05x10 ⁻² **	$[CV_{TN14}^{0.10}]_L = (0.0920 \pm 0.0027) - (7.6 \times 10^{-6} \pm 7 \times 10^{-7}) \cdot n$	
	Q	0.968498; 3.93x10 ⁻⁴ ; 0.441	$[CV_{TN14}^{0.10}]_Q = (0.1099 \pm 0.0022) - (2.00 \times 10^{-5} \pm 1.3 \times 10^{-6}) \cdot n$ $+ (1.27 \times 10^{-9} \pm 1.3 \times 10^{-10}) \cdot n^2$	
	C	0.991957; 1.00x10 ⁻⁴ ; 9.5 **	$[CV_{TN14}^{0.10}]_C = (0.1260 \pm 0.0023) - (3.61 \times 10^{-5} \pm 2.1 \times 10^{-6}) \cdot n$ $+ (5.1 \times 10^{-9} \pm 5 \times 10^{-10}) \cdot n^2 - (2.46 \times 10^{-13} \pm 3.0 \times 10^{-14}) \cdot n^3$	
	lnL	0.981617; 2.38x10 ⁻⁴ ; 8.54x10 ⁻⁴	$[CV_{TN14}^{0.10}]_{lnL} = (0.298 \pm 0.006) - (0.0295 \pm 0.0008) \cdot \ln(n)$	
	lnQ	0.999802; 3.70x10 ⁻⁶ ; 1.11x10 ⁻⁴	$[CV_{TN14}^{0.10}]_{lnQ} = (0.754 \pm 0.010) - (0.1445 \pm 0.0024) \cdot \ln(n)$ $+ (0.00718 \pm 0.00015) \cdot \ln(n)^2$	
	lnC	0.999998; 4.24x10 ⁻⁵ ; 2.77x10 ⁻⁵	$[CV_{TN14}^{0.02}]_{lnC} = (2.250 \pm 0.023) - (0.622 \pm 0.009) \cdot \ln(n)$ $+ (0.0602 \pm 0.0011) \cdot \ln(n)^2 - (0.002012 \pm 0.000044) \cdot \ln(n)^3$	
	95% / 5% / 0.05	L	0.843707; 3.23x10 ⁻³ ; 4.93x10 ⁻² **	$[CV_{TN14}^{0.05}]_L = (0.1183 \pm 0.0034) - (9.7 \times 10^{-6} \pm 8 \times 10^{-7}) \cdot n$
		Q	0.965563; 6.57x10 ⁻⁴ ; 0.742	$[CV_{TN14}^{0.05}]_Q = (0.1413 \pm 0.0029) - (2.57 \times 10^{-5} \pm 1.7 \times 10^{-6}) \cdot n$ $+ (1.64 \times 10^{-9} \pm 1.7 \times 10^{-10}) \cdot n^2$
		C	0.991834; 1.75x10 ⁻⁴ ; 16**	$[CV_{TN14}^{0.05}]_C = (0.1620 \pm 0.0029) - (4.66 \times 10^{-5} \pm 2.7 \times 10^{-6}) \cdot n$ $+ (6.6 \times 10^{-9} \pm 6 \times 10^{-10}) \cdot n^2 - (3.17 \times 10^{-13} \pm 3.9 \times 10^{-14}) \cdot n^3$
lnL		0.981379; 3.85x10 ⁻⁴ ; 1.50x10 ⁻³	$[CV_{TN14}^{0.05}]_{lnL} = (0.383 \pm 0.008) - (0.0380 \pm 0.0010) \cdot \ln(n)$	
lnQ		0.999786; 5.71x10 ⁻⁶ ; 1.98x10 ⁻⁴	$[CV_{TN14}^{0.05}]_{lnQ} = (0.973 \pm 0.013) - (0.1868 \pm 0.0033) \cdot \ln(n)$ $+ (0.00930 \pm 0.00020) \cdot \ln(n)^2$	
lnC		0.999998; 2.08x10 ⁻⁴ ; 9.78x10 ⁻⁵	$[CV_{TN14}^{0.05}]_{lnC} = (1.780 \pm 0.015) - (0.491 \pm 0.006) \cdot \ln(n)$ $+ (0.0473 \pm 0.0007) \cdot \ln(n)^2 - (0.001579 \pm 0.000030) \cdot \ln(n)^3$	

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 (all 15 tests with their 33 variants, or only those selected for this purpose). The multiple-test method was already shown to perform better than both the box-and-whisker plot and the “two standard deviation” (2s) methods used for processing interlaboratory data on RMs for quality control purposes (Verma *et al.*, 2008a). Here in the following, we show

that our method performs better than the MAD_Z outlier detection method.

A new set of examples from proteomics

One area where the new critical values and the combined natural logarithm-cubic equations will be useful is molecular and cellular proteomics (Xia *et al.*, 2006; unpublished data from Xia *et al.* were kindly provided by these

Table 1. (Cont.) Regression equations (six different models) fitted to 27 simulated critical values of test N14 of one extreme outlier (for n between 1,000 and 10,000; Table A38 from electronic supplement), and their evaluation for interpolation (for $1,000 > n > 10,000$) and extrapolation ($n > 10,000$) purposes (see Table A38 for those n for which critical values were simulated and for which interpolated or extrapolated critical values were required).

CL / SL / α	Type of fit	R^2 $\{\sum(SIM-FitEq)^2\}_{int}$ $\{\sum(SIM-FitEq)^2\}_{ext}$	Interpolation or extrapolation equation	
98% / 2% / 0.02	L	0.843313; 5.07x10 ⁻³ ; 7.83x10 ⁻² **	$[CV_{TN14}^{0.02}]_L = (0.1480 \pm 0.0043) - (1.22 \times 10^{-5} \pm 1.1 \times 10^{-6}) \cdot n$	
	Q	0.968050; 1.03x10 ⁻³ ; 1.15	$[CV_{TN14}^{0.02}]_Q = (0.1768 \pm 0.0036) - (3.22 \times 10^{-5} \pm 2.1 \times 10^{-6}) \cdot n$ $+ (2.05 \times 10^{-9} \pm 2.1 \times 10^{-10}) \cdot n^2$	
	C	0.991769; 2.82x10 ⁻⁴ ; 25**	$[CV_{TN14}^{0.02}]_C = (0.2028 \pm 0.0037) - (5.84 \times 10^{-5} \pm 3.4 \times 10^{-6}) \cdot n$ $+ (8.3 \times 10^{-9} \pm 8 \times 10^{-10}) \cdot n^2 - (4.0 \times 10^{-13} \pm 5 \times 10^{-14}) \cdot n^3$	
	lnL	0.981219; 6.14x10 ⁻⁴ ; 2.44x10 ⁻³	$[CV_{TN14}^{0.02}]_{lnL} = (0.479 \pm 0.010) - (0.0476 \pm 0.0013) \cdot [\ln(n)]$	
	lnQ	0.999778; 8.07x10 ⁻⁶ ; 2.76x10 ⁻⁴	$[CV_{TN14}^{0.02}]_{lnQ} = (1.222 \pm 0.017) - (0.2348 \pm 0.0042) \cdot [\ln(n)]$ $+ (0.01169 \pm 0.00026) \cdot [\ln(n)]^2$	
	lnC	0.999998; 5.54x10 ⁻⁴ ; 2.52x10 ⁻⁵	$[CV_{TN14}^{0.02}]_{lnC} = (2.250 \pm 0.023) - (0.622 \pm 0.009) \cdot [\ln(n)]$ $+ (0.0602 \pm 0.0011) \cdot [\ln(n)]^2 - (0.002012 \pm 0.000044) \cdot [\ln(n)]^3$	
	99% / 1% / 0.01	L	0.842963; 6.56x10 ⁻³ ; 0.102**	$[CV_{TN14}^{0.01}]_L = (0.168 \pm 0.005) - (1.39 \times 10^{-5} \pm 1.2 \times 10^{-6}) \cdot n$
		Q	0.967836; 1.34x10 ⁻³ ; 1.49	$[CV_{TN14}^{0.01}]_Q = (0.2006 \pm 0.0041) - (3.66 \times 10^{-5} \pm 2.4 \times 10^{-6}) \cdot n$ $+ (2.33 \times 10^{-9} \pm 2.4 \times 10^{-10}) \cdot n^2$
		C	0.991673; 3.71x10 ⁻⁴ ; 31**	$[CV_{TN14}^{0.01}]_C = (0.2302 \pm 0.0042) - (6.64 \times 10^{-5} \pm 3.9 \times 10^{-6}) \cdot n$ $+ (9.4 \times 10^{-9} \pm 9 \times 10^{-10}) \cdot n^2 - (4.5 \times 10^{-13} \pm 6 \times 10^{-14}) \cdot n^3$
lnL		0.981060; 7.91x10 ⁻⁴ ; 3.05x10 ⁻³	$[CV_{TN14}^{0.01}]_{lnL} = (0.544 \pm 0.012) - (0.0540 \pm 0.0015) \cdot [\ln(n)]$	
lnQ		0.999762; 1.40x10 ⁻⁴ ; 5.48x10 ⁻⁴	$[CV_{TN14}^{0.01}]_{lnQ} = (1.390 \pm 0.020) - (0.267 \pm 0.005) \cdot [\ln(n)]$ $+ (0.01332 \pm 0.00031) \cdot [\ln(n)]^2$	
lnC		0.999996; 1.31x10 ⁻⁶ ; 2.90x10 ⁻⁵	$[CV_{TN14}^{0.01}]_{lnC} = (2.596 \pm 0.033) - (0.722 \pm 0.012) \cdot [\ln(n)]$ $+ (0.0702 \pm 0.0016) \cdot [\ln(n)]^2 - (0.00236 \pm 0.00006) \cdot [\ln(n)]^3$	

researchers, which are used to illustrate this application).

We processed 1,402 protein datasets for the application of our multiple-test method for the sample sizes n between 3 and 23,773. The datasets were divided into three groups according to n as follows: (a) very large sizes 1,001-30,000 (49 cases of protein data with sizes between 1,002 and 23,773); (b) medium sizes 101-1,000 (451 cases of protein data with sizes between 101 and 999); and (c) small sizes 3-100 (902 cases of protein data with sizes be-

tween 3 and 100). The criterion for this selection depended directly on the sample sizes for which new, precise critical values were simulated in different papers (see, respectively, this paper for very large sample sizes >1000 ; Verma *et al.*, 2008a for medium sample sizes 101-1,000; and Verma and Quiroz-Ruiz, 2006a, 2006b for small sample sizes up to 100). All 33 discordancy tests were applied to cases in (a) and (b) using the most precise critical values and equations (presented in Verma *et al.*, 2008a and this study) whereas

Table 1. (Cont.) Regression equations (six different models) fitted to 27 simulated critical values of test N14 of one extreme outlier (for n between 1,000 and 10,000; Table A38 from electronic supplement), and their evaluation for interpolation (for $1,000 > n > 10,000$) and extrapolation ($n > 10,000$) purposes (see Table A38 for those n for which critical values were simulated and for which interpolated or extrapolated critical values were required).

CL / SL / α	Type of fit	R^2 $\{\sum(SIM-FitEq)^2\}_{int}$ $\{\sum(SIM-FitEq)^2\}_{ext}$	Interpolation or extrapolation equation
99.5% / 0.5% / 0.005	L	0.842902; 8.08x10 ⁻³ ; 0.125**	$[CV_{TN14}^{0.005}]_L = (0.186 \pm 0.005) - (1.54 \times 10^{-5} \pm 1.3 \times 10^{-6}) \cdot n$
	Q	0.967785; 1.66x10 ⁻³ ; 1.84	$[CV_{TN14}^{0.005}]_Q = (0.223 \pm 0.005) - (4.07 \times 10^{-5} \pm 2.7 \times 10^{-6}) \cdot n$ $+ (2.59 \times 10^{-9} \pm 2.7 \times 10^{-10}) \cdot n^2$
	C	0.991634; 4.39x10 ⁻⁴ ; 39**	$[CV_{TN14}^{0.005}]_C = (0.255 \pm 0.005) - (7.37 \times 10^{-5} \pm 4.3 \times 10^{-6}) \cdot n$ $+ (1.04 \times 10^{-8} \pm 1.0 \times 10^{-9}) \cdot n^2 - (5.0 \times 10^{-13} \pm 6 \times 10^{-14}) \cdot n^3$
	lnL	0.981028; 9.76x10 ⁻⁴ ; 3.81x10 ⁻³	$[CV_{TN14}^{0.005}]_{lnL} = (0.604 \pm 0.013) - (0.0600 \pm 0.0017) \cdot [\ln(n)]$
	lnQ	0.999757; 1.46x10 ⁻⁵ ; 4.40x10 ⁻⁴	$[CV_{TN14}^{0.005}]_{lnQ} = (1.544 \pm 0.022) - (0.297 \pm 0.006) \cdot [\ln(n)]$ $+ (0.01480 \pm 0.00034) \cdot [\ln(n)]^2$
	lnC	0.999996; 3.64x10 ⁻⁴ ; 9.44x10 ⁻⁶	$[CV_{TN14}^{0.005}]_{lnC} = (2.895 \pm 0.039) - (0.806 \pm 0.015) \cdot [\ln(n)]$ $+ (0.0785 \pm 0.0018) \cdot [\ln(n)]^2 - (0.00264 \pm 0.00008) \cdot [\ln(n)]^3$

CL : Confidence level (%); SL : Significance level (%); α : Significance level; CV: critical value; Type of fit refers to the six different models as follows: L: linear in n -CV axes; Q: quadratic in n -CV axes; C: cubic in n -CV axes; lnL: logarithm-linear in $\ln(n)$ -CV axes; lnQ: logarithm-quadratic in $\ln(n)$ -CV axes; lnC: logarithm-cubic in $\ln(n)$ -CV axes. Thus, six different regression models were evaluated (see text for more details).

$\{\sum(SIM-FitEq)^2\}_{int}$ = sum of squares of residuals for $n = 1,000$ to $n = 10,000$ (for interpolation purposes); $\{\sum(SIM-FitEq)^2\}_{ext}$ = sum of squares of residuals for $n = 20,000$ and $n = 30,000$ (for extrapolation purposes). Thus, two sets of fitting quality parameter were used for the evaluation of fitted equations. The first parameter $\{\sum(SIM-FitEq)^2\}_{int}$ is the total sum of squares of the difference between the simulated critical value (SIM) and that (FitEq) predicted by the equation for the 27 simulated values corresponding to $n = [1,000(50)1,500(100)2,000(500)5,000(1,000)10,000]$ for a given CL and for a given regression model (see Table A38 for the SIM values for $n = 1,000$ to $10,000$ used for this fitting). The second parameter $\{\sum(SIM-FitEq)^2\}_{ext}$ is for extrapolation of these equations to predict two critical values for n of 20,000 and 30,000. $\{\sum(SIM-FitEq)^2\}_{ext}$ value identified by two asterisks (**) was obtained from negative critical values (both extrapolated critical values for sizes 20,000 and 30,000 were found to be negative), which is not realistic, and therefore in this case, this parameter is meaningless as a quality parameter. Note that independent equations were fitted for each confidence level (70% to 99.5%) or significance level α (0.30 to 0.005).

As an example, $CV_{TN14}^{0.30}$ in the interpolation equation is the critical value (CV) for test TN14 and significance level $\alpha = 0.30$ obtained by a simple linear regression model. The parameter n is the sample size of the critical value to be computed from the equation for a given significance level (α). $[CV_{TN14}^{0.05}]_{lnC}$ and $[CV_{TN14}^{0.01}]_{lnC}$ are the most commonly used critical values and the corresponding CL/SL/ α are shown in bold face. Note also that Verma (1997) recommended the strict level of $\alpha = 0.1$ be used in application of the multiple-test method. The other CV values in these equations are similarly explained. Finally, note that the coefficients in all equations are reported as rounded values depending on the respective errors as suggested by Verma (2005).

only 13 single-outlier test variants were applied to cases in (c) using the critical values published earlier (Verma and Quiroz-Ruiz, 2006a, 2006b). It is important to note that our multiple-test method was applied at the strict 99% confidence level as originally proposed by Verma (1997). The numbers of discordant outliers detected were registered individually for all 1,402 cases. Similarly, the corresponding numbers for MAD_Z robust outlier method were also counted in the unpublished database by Xia *et al.* No attempt was made to present and compare the central tendency and dispersion parameters from these two statistical methods simply because the protein data made available to us are still unpublished. This can be done in Xia *et al.* paper itself (if those authors would decide to do so), or after they publish their results of the MAD_Z outlier method.

In Figure 3(a-c) we schematically present an objective comparison of our multiple-test method (MTM) with the MAD_Z method practiced by Xia *et al.* (unpublished database). In most cases (Figure 3a-c), especially for medium (Figure 3b) and very large sizes (Figure 3a), our method detects a greater number of discordant outliers than the MAD_Z method. For very small datasets ($n \leq 5$), both methods detect practically no discordant outliers in protein databases.

Thus, the multiple-test method initially proposed by Verma (1997) along with the new, precise critical values and relevant interpolation and extrapolation equations (Verma and Quiroz-Ruiz, 2006a, 2006b; Verma *et al.* 2008a; this work) can be advantageously used to process such large databases as the 1,402 protein cases presented here.

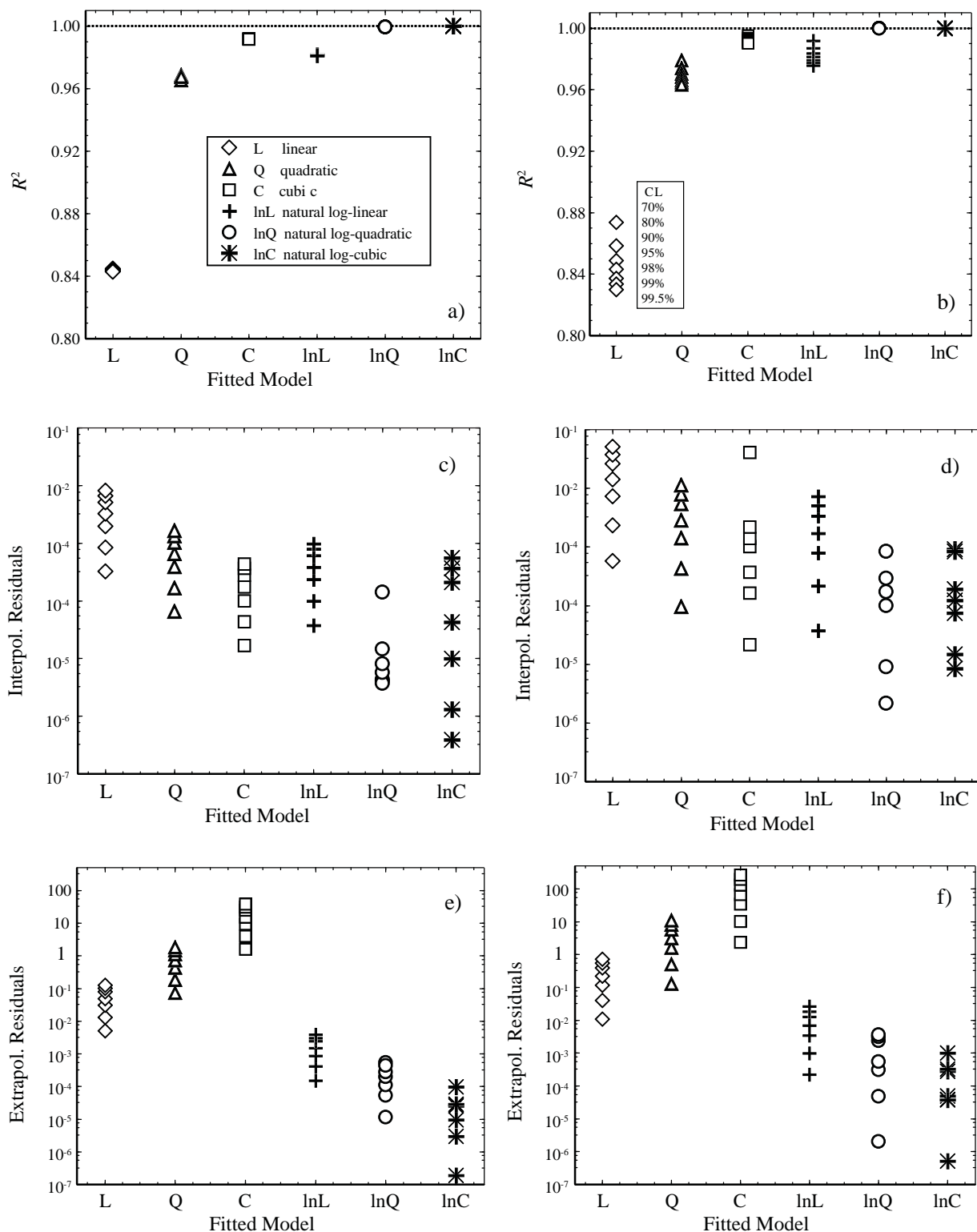


Figure 1. Graphic evaluation of fitting parameters for the six regression models: L: linear in n -CV axes; Q: quadratic in n -CV axes; C: cubic in n -CV axes; lnL: natural logarithm-transformed n with linear fit, *i.e.*, linear in $\ln(n)$ -CV axes; lnQ: natural logarithm-transformed n with quadratic fit, *i.e.*, quadratic in $\ln(n)$ -CV axes; lnC: natural logarithm-transformed n with cubic fit, *i.e.*, cubic in $\ln(n)$ -CV axes. n : sample size; CV: critical value. Results for all seven significant levels $\alpha = 0.30, 0.20, 0.10, 0.05, 0.02, 0.01,$ and 0.005 , corresponding to sample sizes 1,000(50)1,500(100)2,000(500)5,000(1,000)10,000, were used (see Tables 1 and A41 for more details). (a) Parameter R^2 for the six models applied to 27 critical value data (Table A38) for test N14, note the horizontal dotted line at $R^2=1$ represents the maximum possible value for this parameter, the inset explains the abbreviations used in the x-axis (Fitted model); (b) Same as (a), but for test N15, CL: confidence level; (c) Parameter $\{\sum(SIM-FitEq)^2\}_{in}$ (Interpolated Residuals) for 27 critical values for test N14, note six orders of magnitude scale; (d) Same as (c), but for test N15; (e) Parameter $\{\sum(SIM-FitEq)^2\}_{ex}$ (Extrapolated Residuals) for 2 critical values for test N14 ($n=20,000$ and $30,000$), note more than nine orders of magnitude scale; and (f) Same as (e), but for test N15.

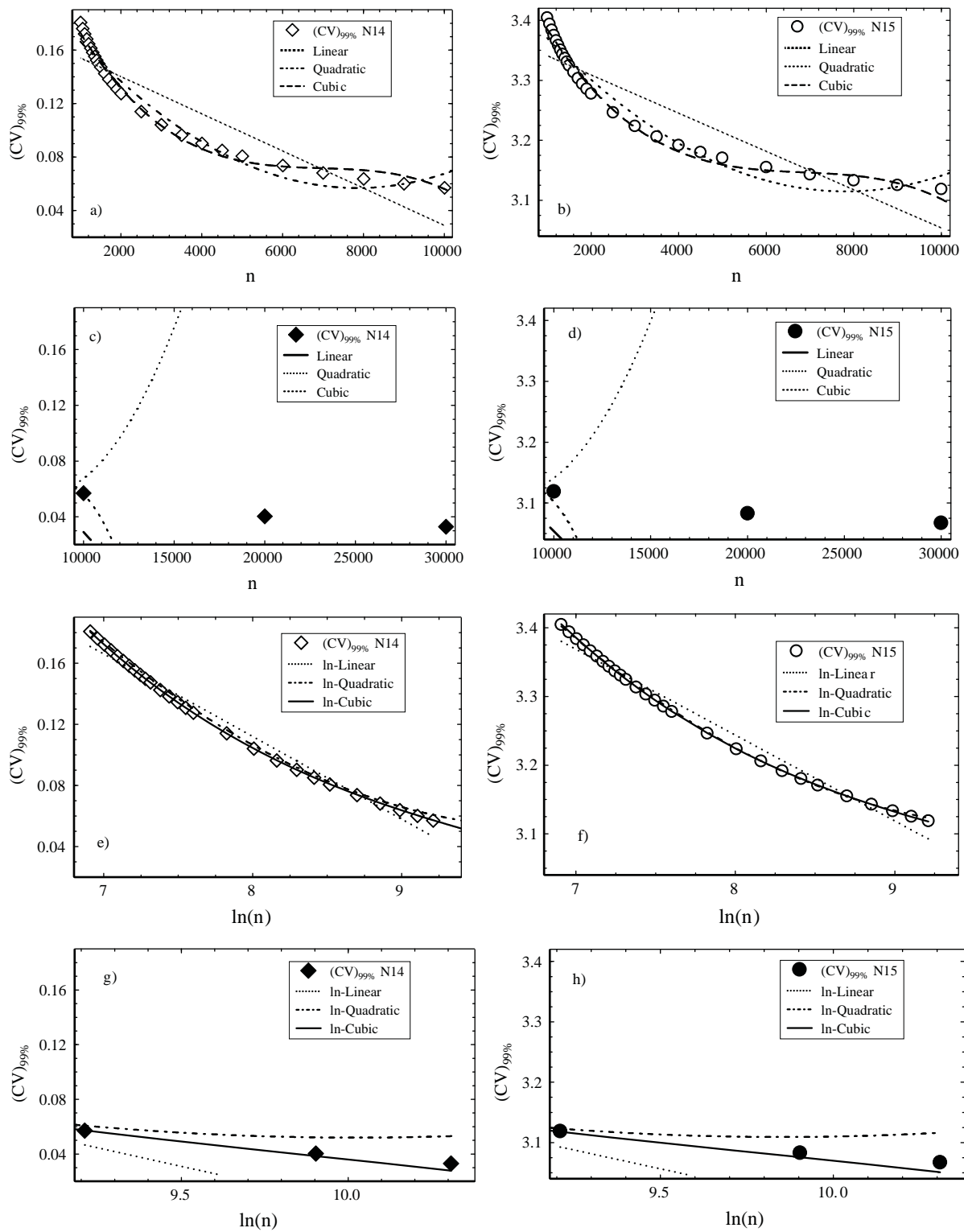


Figure 2. Interpolation and extrapolation curves (drawn for the six regression models, see Figure 1 for more details) and the critical value data for the recommended significance level $\alpha = 0.01$, and sample sizes 1,000(50)1,500(100)2,000(500)5,000(1,000)10,000. Note that $\alpha = 0.01$ is the recommended significance level to be used to test experimental data for possible discordant outliers as recommended by Verma (1997). Also included are two greater sample sizes of 20,000 and 30,000. (a) Test N14 (see Table 1 for more information); (b) Test N15 (see Table A41 for more information).

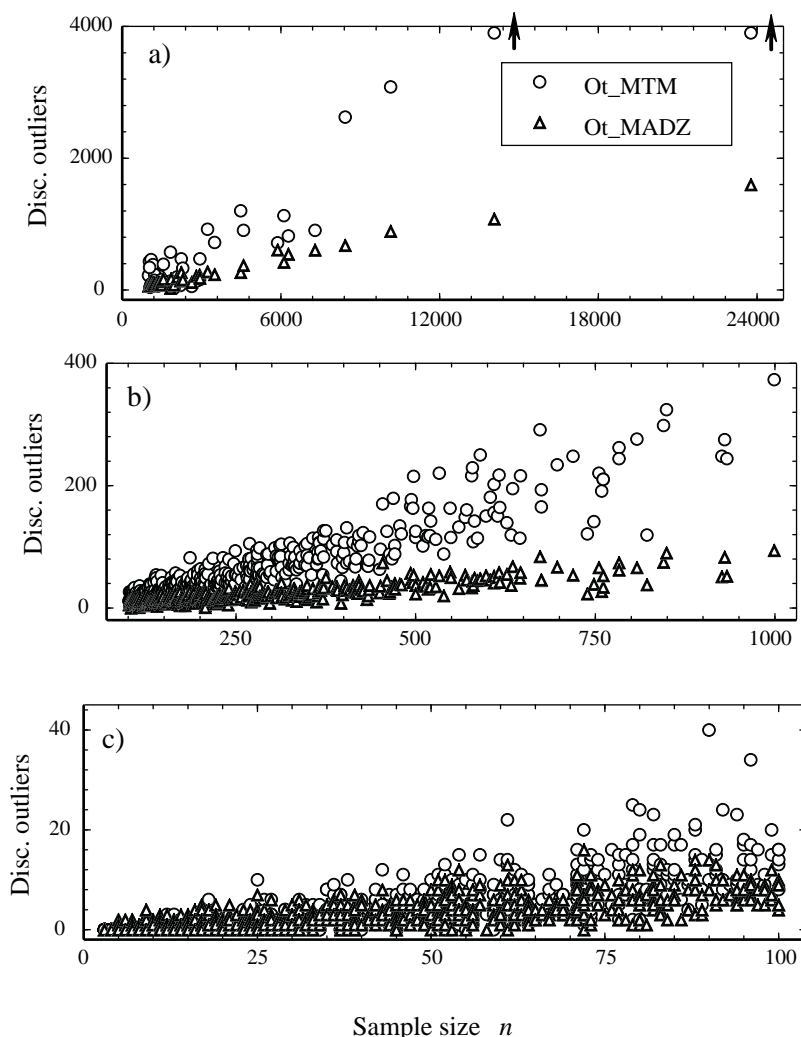


Figure 3. Comparison of the multiple-test method (MTM) with the MAD_Z robust outlier method practiced for quantitative protein data (unpublished data by Xia *et al.*). The x-axis is for sample size n whereas the y-axis gives the actual number of discordant (disc.) outliers (open circles are for discordant outliers by the multiple-test method – Ot_MTM, whereas open diamonds are for the MAD_Z robust method – Ot_MADZ). Such large numbers of discordant outliers (Figures 3a-c) might also suggest that the protein data distribution might be of some other type, such as log-normal, but this can be explored more freely after (or during) the publication of these (unpublished) protein data by the original authors (Xia *et al.*). (a) Very large sample sizes (49 cases), note that for two protein samples of largest sizes of 14,073 and 23,773, the outliers detected as discordant by the multiple-test method (MTM) were outside the plot as shown (*i.e.*, much greater number of outliers were detected by the MTM); (b) Medium sample sizes (451 cases); and (c) Small sample sizes (902 cases).

FUTURE WORK

The present paper closes the series of four papers published on the subject of new critical values for the existing 15 discordancy tests with 33 test variants, because new, precise and accurate critical values (with appropriate interpolation and extrapolation equations) have now been generated for all sample sizes up to 30,000. These tests include 13 single outlier versions ($k=1$) and 20 multiple outlier variants (up to $k=4$). The proposed interpolation and extrapolation equations can also be improved, if necessary, by exploring higher order regression fits than the cubic fits presented here. In fact, the extrapolation needs could be totally eliminated by obtaining new sets of equations based on all critical values

from 1,000 (or a smaller size) up to 30,000.

For such large n (several thousands!) these tests with $k=1$ to $k=4$, at first glance, may not appear to be of much use, because a much greater number of discordant outliers (> 4) may actually be present in such datasets, and new test statistics (with significantly greater k) may have to be proposed and investigated. Surprising, however, these tests seem to perform satisfactorily well for the proteomics data (see Figure 3a for very large sample sizes).

A suitable computer program that could facilitate the application of these discordancy tests on a routine basis would be much useful in all areas of science and engineering, including the geosciences. A preliminary version of this program is already available although its improvements

are underway for enabling its publication (Verma and Díaz-González, in preparation).

Nevertheless, a proper evaluation of efficiencies (*e.g.*, Velasco and Verma, 1998; Velasco *et al.*, 2000) or probabilities as well as of type I or type II errors may have to be first undertaken (see, *e.g.*, Barnett and Lewis, 1994; Hays and Kinsella, 2003; Efstathiou, 2006). This type of work, already in progress, will constitute a great step forward in the application of discordancy tests and a proper handling of experimental data, especially those involving large sample sizes. In fact, the first paper on these lines is already in press (Verma *et al.*, 2008b). Finally, a proper evaluation of the two broad groups of methods – robust methods and outlier-based methods – for the statistical handling of experimental data (Verma, 2005) would be of interest to the scientific and engineering community.

And some of the ideas expressed in this final section should constitute a new series of papers of great interest to all those involved in the handling and interpretation of experimental data.

CONCLUSIONS

In this final paper, we have used our established and well-tested Monte Carlo type simulation procedure for generating new, precise, and accurate critical values for 15 discordancy tests with 33 test variants for sample sizes up to 30,000. These new critical values will cover any future needs of many diverse fields of science and engineering, including molecular and cellular proteomics and quality control in Earth Sciences. In fact, the example set for proteomics presented here demonstrates the usefulness of the multiple-test method. Much work remains to be done to evaluate the performance of discordancy tests and to propose new test statistics, especially for very large sample sizes.

ACKNOWLEDGEMENTS

This final manuscript prepared as a result of the invitation to the first author extended during 2005 by the editor-in-chief Susana Alaniz-Álvarez, closes the series of four papers on the subject of new critical values for 33 test variants. We are also much grateful to Qiangwei Xia, Murray Hackett and John A. Leigh for the use of their unpublished dataset, being a part of the manuscript in preparation by Xia *et al.* We acknowledge their kindness because this application example from quantitative proteomics based on their unpublished data is likely to enhance the importance of our work in areas other than the geosciences. Finally, we are much grateful to the official reviewers of this paper – Qiangwei Xia (USA), Constantinos E. Efstathiou (Greece), and an individual who decided to remain anonymous – for their highly positive evaluation of our final paper in this series.

APPENDIX A. SUPPLEMENTARY DATA

Tables A1-A59 can be found at the journal web site <<http://satori.geociencias.unam.mx/>>, in the table of contents of this issue (electronic supplement 25-3-01).

REFERENCES

- Andaverde, J., Verma, S.P., Santoyo, E., 2005, Uncertainty estimates in the calculation of static formation temperatures in boreholes and evaluation of regression models: *Geophysical Journal International*, 160(3), 1112-1122.
- Barnett, V., Lewis, T., 1994, *Outliers in Statistical Data*: Chichester, John Wiley, Third edition, 584 p.
- Bevington, P.R., Robinson, D.K., 2003, *Data Reduction and Error Analysis for the Physical Sciences*: McGrawHill, NewYork, Third edition, 320 p.
- Bugner, E., Rutledge, D.N., 1990, Modelling of statistical tables for outlier tests: *Chemometrics and Intelligent Laboratory Systems*, 9(3), 257-259.
- Colombo, F., Pannunzio-Miner, E.V., Gay, H.D., Lira, R., Dorais, M.J., 2007, Barbosalita y lipscombbita en Cerro Blanco, Córdoba (Argentina): descripción y génesis de fosfatos secundarios en pegmatitas con triplita y apatita: *Revista Mexicana de Ciencias Geológicas*, 24(1), 120-130.
- Díaz-González, L., Santoyo, E., Reyes-Reyes, J., 2008, Tres nuevos geotermómetros mejorados de Na/K usando herramientas computacionales y geoquimiométricas: aplicación a la predicción de temperaturas de sistemas geotérmicos: *Revista Mexicana de Ciencias Geológicas*, 25(3), in press.
- Efstathiou, C.E., 2006, Estimation of type I error probability from experimental Dixon's "Q" parameter on testing for outliers within small size data sets: *Talanta*, 69(5), 1068-1071.
- Guevara, M., Verma, S.P., Velasco-Tapia, F., 2001, Evaluation of GSI intrusive rocks JG1, JG2, JG3, JG1a, and JGb1 by an objective outlier rejection statistical procedure: *Revista Mexicana de Ciencias Geológicas*, 18(1), 74-88.
- Hays, K., Kinsella, A., 2003, Spurious and non-spurious power in performance criteria for tests of discordancy: *The Statistician*, 52(1), 69-82.
- Jafarzadeh, M., Hosseini-Barzi, M., 2008, Petrography and geochemistry of Ahwaz sandstone member of Asmari Formation, Zagros, Iran: implications on provenance and tectonic setting: *Revista Mexicana de Ciencias Geológicas*, 25(2), 247-260.
- Méndez-Ortiz, B.A., Carrillo-Chávez, A., Monroy Fernández, M.G., 2007, Acid rock drainage and metal leaching on mine waste material (tailings) from a Pb-Zn-Ag skarn deposit: Environmental assessment through static and kinetic laboratory tests: *Revista Mexicana de Ciencias Geológicas*, 24(2), 161-169.
- Meriggi, L., Macías, J.L., Tommasini, S., Capra, L., Conticelli, S., 2008, Heterogeneous magmas of the Quaternary Sierra Chichinautzin Volcanic Field (central Mexico): the role of an amphibole-bearing mantle and magmatic evolution processes: *Revista Mexicana de Ciencias Geológicas*, 25(2), 197-216.
- Montenegro, T., Vattuone, M.E., 2008, Asociaciones minerales de muy bajo grado metamórfico vinculadas a alteración hidrotermal, sudoeste de Trevelin, Chubut, Argentina: *Revista Mexicana de Ciencias Geológicas*, 25(3), 302-313.
- Nagarajan, R., Madhavaraju, J., Nagendra, R., Armstrong-Altrin, J.S., Moutte, J., 2007, Geochemistry of Neoproterozoic shales of the Rabanpalli Formation, Bhima Basin, Northern Karnataka, southern India: implications for provenance and paleoredox conditions: *Revista Mexicana de Ciencias Geológicas*, 24(2), 150-160.
- Nagarajan, R., Sial, A.N., Armstrong-Altrin, J.S., Madhavaraju, J., Nagendra, R., 2008, Carbon and oxygen isotope geochemistry of Neoproterozoic limestones of the Shahabad Formation, Bhima Basin, Karnataka, southern India: *Revista Mexicana de Ciencias*

- Geológicas, 25(2), 225-235.
- Obeidat, M.M., Ahmad, F.Y., Hamouri, N.A.A., Massadeh, A.M., Athamneh, F.S., 2008, Assessment of nitrate contamination of karst springs, Bani Kanana, northern Jordan: *Revista Mexicana de Ciencias Geológicas* 25(3), in press.
- Palabiyik, Y., Serpen, U., 2008, Geochemical Assessment of Simav Geothermal Field, Turkey: *Revista Mexicana de Ciencias Geológica* 25(3), in press.
- Ramos-Leal, J.A., Durazo, J., González-Morán, T., Juárez-Sánchez, F., Cortés-Silva, A., Johannesson, K.H., 2007, Evidencias hidrogeoquímicas de mezclas de flujos regionales en el acuífero de la Muralla, Guanajuato: *Revista Mexicana de Ciencias Geológicas*, 24 (3), 293-305.
- Rodríguez-Ríos, R., Aguillón-Robles, A., Leroy, J.L., 2007, Evolución petrológica y geoquímica de un complejo de domos topacíferos en el campo volcánico de San Luis Potosí (México): *Revista Mexicana de Ciencias Geológicas*, 24(3), 328-343.
- Rorabacher, D.B., 1991, Statistical treatment for rejection of deviant values: critical values of Dixon's "Q" parameter and related subrange ratios at the 95% confidence level: *Analytical Chemistry*, 63(2), 139-146.
- Salleh, S.H., Rosales, E., Flores de la Mota, I., 2007, Influence of different probability based models on oil prospect exploration decision making: A case from southern Mexico: *Revista Mexicana de Ciencias Geológicas*, 24(3), 306-317.
- 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.
- Vargas-Rodríguez, Y.M., Gómez-Vidales, V., Vázquez-Labastida, E., García-Borquez, A., Aguilar-Sahagun, G., Murrieta-Sánchez, H., Salmon, M., 2008, Caracterización espectroscópica, química y morfológica y propiedades superficiales de una montmorillonita Mexicana: *Revista Mexicana de Ciencias Geológicas*, 25(1), 135-144.
- Vattuone, M.E., Leal, P.R., Crosta, S., Berbeglia, Y., Gallegos, E., Martínez-Dopico, C., 2008, Paragénesis de zeolitas alcalinas en un afloramiento de basaltos olivínicos amigdaloides de Junín de Los Andes, Neuquén, Patagonia, Argentina: *Revista Mexicana de Ciencias Geológicas* 25(3), in press.
- Velasco, F., Verma, S.P., 1998, Importance of skewness and kurtosis statistical tests for outlier detection and elimination in evaluation of Geochemical Reference Materials: *Mathematical Geology*, 30(1), 109-128.
- 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-Geochemistry*, 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, M.P., 2004, A revised analytical method for HCO₃⁻ and CO₃²⁻ determinations in geothermal waters: an assessment of IAGC and IAEA interlaboratory comparisons: *Geostandards and Geoanalytical Research*, 28(3), 391-409.
- 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., Quiroz-Ruiz, A., 2006a, 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., Quiroz-Ruiz, A., 2006b, Critical values for 22 discordancy test variants for outliers in normal samples up to sizes 100, and applications in science and engineering: *Revista Mexicana de Ciencias Geológicas*, 23(3), 302-319; with electronic tables available at <http://satori.geociencias.unam.mx>.
- 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., Quiroz-Ruiz, A., Díaz-González, L., 2008a, Critical values for 33 discordancy test variants for outliers in normal samples up to sizes 1000, and applications in quality control in Earth Sciences: *Revista Mexicana de Ciencias Geológicas*, 25(1), 82-96, with 209 pages of electronic supplement 25-1-01 Critical values for 33 discordancy tests, available at <http://satori.geociencias.unam.mx>.
- Verma, S.P., Díaz-González, L., González-Ramírez, R., 2008b, Relative efficiency of single-outlier discordancy tests for processing geochemical data on reference materials and application in instrumental calibrations by a weighted least-squares linear regression model: *Geostandards and Geoanalytical Research* (in press).
- Xia, Q.W., Hendrickson, E.L., Zhang, Y., Wang, T.S., Taub, F., Moore, B.C., Porat, I., Whitman, W.B., Hackett, M., Leigh, J.A., 2006, Quantitative proteomics of the Archaeon *Methanococcus maripaludis* validated by microarray analysis and real time PCR: *Molecular & Cellular Proteomics*, 5 (5), 868-881.
- Yadava, M.G., Ramesh, R., Pandarinath, K., 2007, A positive 'amount effect' in the Sahayadri (Western Ghats) rainfall: *Current Science*, 93(4), 560-564.

Manuscript received: March 14, 2008

Corrected manuscript received: May 14, 2008

Manuscript accepted: May 14, 2008