ELSEVIER

Analytica Chimica Acta 293 (1994) 183–189

ANALYTICA CHIMICA ACTA

Data analysis in the chemical laboratory Part 1. Analysis of indirect measurements

Milan Meloun [a,*], Jiří Militký [b]

[a] *Department of Analytical Chemistry, University of Chemical Technology, 532 10 Pardubice, Czech Republic*
[b] *Department of Textile Materials, Technical University, 461 17 Liberec, Czech Republic*

(Received 30th March 1993; revised manuscript received 10th November 1993)

Abstract

Response quantities of analytical chemistry investigations of, for instance, concentration or content of substances, viscosity, stability constants or solubility, can be obtained as a non-linear transformation of directly measured quantities or signals. The goal of the indirect measurements analysis is estimation of basic statistical parameters of analytical results from the known non-linear transformation and from the statistical parameters of measured variables. The analysis is based on Taylor series expansion, two-point approximation and Monte Carlo simulation. An algorithm may be applied on any chemical, physical, biological or medical result.

Key words: Error propagation; Indirect measurements; Monte Carlo simulation; Taylor series expansion; Two-points approximation

1. Introduction

A result of a chemical analysis y is often calculated as the known functional transformation $y = G(x_1, \ldots, x_m)$ from a number of directly measured experimental quantities $x_1, \ldots x_m$. Due to various kinds of errors the measured quantities x_i, $i = 1, \ldots, m$, are random variables. Using a basic statistical treatment of measured data the sample means $\bar{x}_i$ and sample variances $s^2(x_i)$, $i = 1, \ldots, m$, are computed. To project these errors into the resultant y is a topic treated in many analytical chemistry texts and is known as error propagation. The well-known formula for the random errors propagation

$$s^2(y) \simeq \left(\frac{dG(x_i)}{dx_i} \right)^2 s^2(x_i) \qquad (1)$$

is based on a number of assumptions: (a) the random variables x_i are uncorrelated; and (b) if y is not a linear function of x_i, then each $s(x_i)$ must be sufficiently small relative to the corresponding mean values of x_i so that the function $G(x_1, \ldots, x_m)$ can be reasonably linearized. If one or more of these assumptions is invalid, Eq. 1 can be suitably corrected but only when making use of a computer. A more difficult problem occurs when the function G(x) is differentiable

* Corresponding author.

SSDI 0003-2670(94)00086-2

with great difficulty only. This problem, however, can be solved by numerical methods.

This article compares the Taylor series expansion, the two-point approximation, and the Monte Carlo simulation for a calculation of the mean $\bar{y}$ and the variance $s^2(y)$ of a response quantity (i.e., an analyte concentration or an analyte content).

2. Theory

The treatment of the indirect measurements under consideration in this paper leads to the following problems:

(1) The estimation of the result of the chemical analysis, i.e., the mean value $\bar{y}$.

(2) The estimation of the total error expressed as a standard deviation of chemical analysis, $s(y)$, from known errors of several measured quantities, $s(x_i)$.

(3) The inverse estimation of limiting errors of measured quantities, $s(x_i)$, from the allowed error of the chemical analysis, $s(y)$.

To express the absolute error of the ith variable x_i, the standard deviation $s(x_i)$ is convenient; for the relative error of x_i the relative standard deviation (or the coefficient of variation) is used

$$\delta(x_i) = s(x_i)/x_i \tag{2}$$

For the first problem, if the experiment and computation of $G(\cdot)$ could be done repeatedly to generate a reasonable statistical sample of y_i values, the information on the random uncertainty in y would be within reach. The computer offers a convenient way of simulating the repetition. It is only necessary to generate new sets of x_i data, and the estimates of mean $\bar{x}_i$, variance $s^2(x_i)$, skewness $g_{1,i}$ and kurtosis $g_{2,i}$ are used, cf. p. 101 in [1].

For the second problem the expression for variance $s^2(y)$ as a function of individual variances $s^2(x_i)$ is used. A simplification can often be achieved using the relative errors.

For the third problem the expression for variance $s^2(y)$ or variation coefficient $\delta(y) = s(y)/\bar{y}$ is used. The basic assumption is that individually measured quantities x_i have the same relative effects.

To solve all three problems the mean $\bar{y}$ and corresponding variance $s^2(y)$ of a function $y = G(x_1, \ldots, x_m)$ must be known. The estimates $\bar{y}$ and $s^2(y)$ may be obtained by any of the following methods: (1) Taylor series expansion of the function $y = G(x_1, \ldots, x_m)$; (2) two-points approximation; and (3) Monte Carlo simulation.

Whereas the method of Taylor series expansion requires knowledge of at least first and second derivatives of the function $G(x_1, \ldots, x_m)$, the remaining two methods can be computer-assisted.

2.1. Method of Taylor series expansion

When a function of random variables is analyzed it should be realized that each non-linear transformation of the random variable distorts its distribution, and therefore changes the dependence of variance on the mean value. In the case when the single measured variable x has a constant variance $s^2(x)$, the results of analysis $y = G(x)$ have a non-constant variance $s^2(y)$, Eq. 1. Moreover, in the multivariate case the sample mean $\bar{y}$ cannot be estimated by direct substitution of sample mean $\bar{x}$ into the function $G(\bar{x})$, i.e.,

$$\bar{y} \neq G(\bar{x}) \tag{3}$$

To estimate the mean $\bar{y}$, the variance $s^2(y)$ and higher statistical moments, the Taylor series expansion of function $G(x)$ can be used.

Suppose that the function $y = G(x_1, \ldots, x_m)$ is known. Let $G(\mathbf{x})$ be doubly differentiable at least. When writing the Taylor series expansion in the neighbourhood of the vector of means $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_m)^T$ we obtain

$$y \approx G(\bar{\mathbf{x}}) + \sum_{i=1}^{m} \frac{\delta G(\mathbf{x})}{\delta x_i}(x_i - \bar{x}_i) + \frac{1}{2}\sum_{i=1}^{m} \frac{\delta^2 G(\mathbf{x})}{\delta x_i^2}(x_i - \bar{x}_i)^2 + \sum_{i=1}^{m-1}\sum_{j=i+1}^{m} \frac{\delta^2 G(\mathbf{x})}{\delta x_i \delta x_j}(x_i - \bar{x}_i)(x_j - \bar{x}_j) + \ldots \tag{4}$$

where all first and second derivatives are calculated for the vector of mean values $\bar{\mathbf{x}}$. By using a mean value operator $E(\cdot)$ at both sides of Eq. 4 the expression for the estimate of mean $\bar{y}$ may be written as

$$\bar{y} \approx G(\bar{\mathbf{x}}) + \frac{1}{2}\sum_{i=1}^{m}\frac{\delta^2 G(\mathbf{x})}{\delta x_i^2}s^2(x_i) + \sum_{i=1}^{m-1}\sum_{j>i}^{m}\frac{\delta^2 G(\mathbf{x})}{\delta x_i \delta x_j}\mathrm{cov}(x_i, x_j) \quad (5)$$

where $\bar{y} = E(y) = E(G(\mathbf{x}))$, $s^2(x_i) = E[(x_i - \bar{x}_i)^2]$ and where $E[(x_i - \bar{x}_i)] = 0$. The symbol $\mathrm{cov}(x_i, x_j)$ stands for the covariance which gives a measure of "linear dependence" between the two variables x_i and x_j.

For computation of variance $s^2(y)$ the linearization based on Taylor expansion is obviously used. More precise is to apply approximation 4 with neglecting the higher moments. (i.e., the skewness and curtosis). The resulting approximate relation for variance is termed the rule of propagation of absolute errors and can be expressed by

$$s^2(y) \approx \sum_{i=1}^{m}\left[\frac{\delta G(\mathbf{x})}{\delta x_i}\right]^2 s^2(x_i) + 2\sum_{i=1}^{m-1}\sum_{j>i}^{m}\frac{\delta G(\mathbf{x})}{\delta x_i}\frac{\delta G(\mathbf{x})}{\delta x_j}\mathrm{cov}(x_i, x_j) + \sum_{i=1}^{m-1}\sum_{j>i}^{m}\frac{\delta^2 G(\mathbf{x})}{\delta x_i \delta x_j}s^2(x_i)s^2(x_j) \quad (6)$$

When the resulting error $s(y)$ is formed from m sources of additive errors, i.e., $G(x_1, \ldots, x_m) = \Sigma x_i$ is linear combination of x_i and each source has its own variance $\delta_i^2(x)$, the following expression for the estimate of error can be used

$$s^2(y) = \sum_{i=1}^{m}\delta^2(x_i) + 2\sum_{i=1}^{m-1}\sum_{j>i}^{m}\mathrm{cov}(x_i, x_j) \quad (7)$$

where $\mathrm{cov}(x_i, x_j)$ again is a measure of linear dependence between the two variables x_i and x_j. There are two limiting cases of estimation of total error of measurements, $s(y)$ from Eq. 7:

(1) The sources of errors are quite independent, so that the covariances $\mathrm{cov}(x_i, x_j)$ are equal to zero. The resulting estimate of an error will be proportional only to the quadratic mean of errors $\delta(x_i)$ coming from m sources,

$$s(y) = \sqrt{\sum_{i=1}^{m}\delta^2(x_i)} \quad (8)$$

(2) The sources of errors are linearly dependent. Then covariances $\mathrm{cov}(x_i, x_j)$ are given by

$$\mathrm{cov}(x_i, x_j) = \sqrt{s^2(x_i)s^2(x_j)}$$

The resulting estimate of the total error will be proportional to the arithmetic mean of errors $\delta(x_i)$ coming from m sources

$$s(y) = \sum_{i=1}^{m}\delta(x_i) \quad (9)$$

For various analytical operations and signal measurements in a chemical laboratory, the function $G(\mathbf{x})$ can be expressed by a power-type relationship

$$y = G(\mathbf{x}) = x_1^{a_1} \cdot x_2^{a_2} \cdot \cdots \cdot x_m^{a_m} = \prod_{i=1}^{m} x_i^{a_i} \quad (10)$$

where a_i are known coefficients usually equal to ∓ 1. The estimation of the absolute error $s(y)$ or $s^2(y)$ by Eq. 6 is then rather complicated. The logarithmic transformation leads to the simpler expression

$$\ln G(\mathbf{x}) = \sum_{i=1}^{m} a_i \cdot \ln x_i \quad (11)$$

Then

$$\frac{\mathrm{d}\ln G(\mathbf{x})}{\mathrm{d}x} = \frac{1}{G(\mathbf{x})}\frac{\mathrm{d}G(\mathbf{x})}{\mathrm{d}x} \quad (12)$$

Substitution of Eq. 12 to Eq. 6 and rearrangement leads to a simplified form for the relative error (variation coefficient)

$$\delta(y) \approx \sqrt{\sum_{i=1}^{m} a_i^2\delta^2(x_i) + 2\sum_{i=1}^{m-1}\sum_{j>i}^{m} a_i a_j r_{ij}\delta(x_i)\delta(x_j)} \quad (13)$$

where r_{ij} represents the correlation coefficient expressing the closeness of linear dependence between variables x_i and x_j. Eq. 13 is called the rule of propagation of relative errors. The quality of estimates $\bar{y}$, $s^2(y)$ and $\delta(y)$ is dependent on the quality of quadratic approximation of the function G(**x**).

Although the estimate $\bar{y}$ is normally sufficiently accurate, some inaccuracy may be found in the estimation $s^2(y)$ [2].

Eq. 13 may be used for estimation of relative errors $\delta(x_i)$ such that a relative error of chemical results $\delta(y)$ will not be greater than the selected value for H in %, i.e., $100 \cdot \delta(y) \leq H$. In solving this inversion problem, the independence of the measured variables x_i and the principle of the same relative influence

$$|a_1|\delta(x_1) \approx |a_2|\delta(x_2) \approx \ldots \approx |a_m|\delta(x_m) \approx H/m$$

are assumed. Here a_i, $i = 1, \ldots, m$, are coefficients of function G(**x**), Eq. 10. For the case of a ratio $G(x_1, x_2) = x_1/x_2$ an estimate of the mean $\bar{y}$ is controlled only by the variance $\delta^2(x_2)$ and not by the variance $\delta^2(x_1)$.

2.2. Method of two-points approximation

Manly's procedure [3] of two-point approximation is based on replacement of the probability distribution of function G(**x**) by the two-points distribution with the same mean and variance. For the case of single x the estimate of the mean is expressed as

$$\bar{y} \approx \{G[\bar{x} + s(x)] + G[\bar{x} - s(x)]\}/2 \quad (14)$$

and the estimate of variance by

$$s^2(y) \approx \{G[\bar{x} + s(x)] - G[\bar{x} - s(x)]\}^2/4 \quad (15)$$

Both simple relations give better results than Taylor's formula for a function of the type in (Eq. 10).

When the function G(**x**) is a function of m independent random variables $x_1, \ldots, x_m$, the summation of Eqs. 14 and 15 can be used

$$\bar{y} \approx \sum_{i=1}^{m} \{G[\bar{x}_i + s(x_i)] + G[\bar{x}_i - s(x_i)]\}/2m \quad (16)$$

and

$$s^2(y) \approx \sum_{i=1}^{m} \{G[\bar{x}_i + s(x_i)] - G[\bar{x}_i - s(x_i)]\}^2 /4m \quad (17)$$

2.3. Method of Monte Carlo simulation

The mean $\bar{y}$ and its variance $s^2(y)$ as a function G(**x**) of random variables x, may be determined by computer-assisted Monte Carlo simulation method. Schwartz [2] showed that this general procedure is well suited for simulation of statistical behaviour of even rather complicated systems. The following steps can be formulated:

(1) Selection of the function G(**x**): for many chemical problems the function G(**x**) is usually known. The great advantage of Monte Carlo simulation method is that the function G(**x**) need not necessarily be expressed in explicit form.

(2) Distribution of measured variables: in chemistry it is usually assumed that measured variables are independent and have normal distribution. Then the Monte Carlo simulation method requires numerical values of quantities $\bar{x}_i$, $s(x_i)$, $i = 1, \ldots, m$, only.

When these values are not available, two limiting values of interval [A,B] in which the variables x_i are expected should be supplied. The approximate probability density function can be then expressed by the parabolic distribution

$$f(x_i) = 6(x_i - A)(B - x_i)/(B - A)^2$$

for $A < x_i < B$. The situation is more complicated when some correlation among the input variables exists. Then the simultaneous distribution of all variables x_i, $i = 1, \ldots, m$, should be specified; this will be simple only for the case of the normal distribution.

(3) Generation of random numbers: most computer software contains a function that will generate pseudo-random numbers from rectangular distribution $R(0,1)$. For two independent random numbers R_j, R_{j+1} the Box–Müller transformation is used to generate two independent random

numbers N_j, N_{j+1}

$$N_j = \sqrt{(-2 \ln R_j)} \sin(2\pi R_{j+1}) \quad (18)$$

$$N_{j+1} = \sqrt{(-2 \ln R_j)} \cos(2\pi R_{j+1}) \quad (19)$$

which have standardized normal distribution. The jth simulated value of the ith variable x_i will be expressed by

$$x^*_{i,j} = N_j s(x_i) + \bar{x}_i \quad (20)$$

(4) The choice of the number of simulations: the rules for the determination of the necessary number of simulations are the same as for the determination of sample size. The minimum number of simulations for the requested $100(1-\alpha)\%$ confidence interval D of the mean is expressed by the relation

$$n_{\min} = \left[4u_{1-\alpha/2} s^2(y)\right]/D^2 + 1 \quad (21)$$

where $u_{1-\alpha/2}$ is the quantile of standardized normal distribution and $s^2(y)$ is the estimate of variance from the first 50 simulations.

(5) The display of results: this step includes a graph of an empirical probability density function of simulated data $\{y_j^*\}$, $j = 1, \ldots, n_{\min}$, and a calculation of the estimates of location and spread, $\bar{y}^*$ and $s^2(\bar{y}^*)$.

3. Computation

The program Propagation-of-Errors calculates the results of indirect measurements or the analytical quantity (concentration, content, etc.) $\bar{y}$ and the variance $s^2(y)$ as a result of several errors concerning various experimental and instrumental operations. In addition to the classical method of Taylor series expansion, two computer-assisted methods can be applied, i.e., the two-points estimation method and the Monte Carlo simulation method. The function G(x) is inserted in the one-row panel using the usual algebraic notations. The maximum number of directly measured variables $(m) = 10$. For these variables the value $\bar{x}_i$ and error $s(x_i)$ are required. For all methods the approximate mean $\bar{y}$, variance $s^2(y)$ and variation coefficient $\delta(y)$ are computed. For Taylor expansion all required derivatives are computed using difference formula. For Monte Carlo simulation the kernel probability density function is also creacted created. The probability density function of normal distribution $N[\bar{y}, s^2(y)]$ is calculated and drawn.

The program Propagation-of-Errors in the package CHEMSTAT is available from the authors up on request.

4. Results

The following samples illustrate the application of the computational Propagation-of-Errors technique.

4.1. Sample 1: error in arsenic content in isotope dilution

Arsenic was determined by the isotope dilution method. The initial specific activity was $a_2 = 3.7 \times 10^4$ s^{-1}. After addition of the standard $m_1 = 5 \times 10^{-7}$ g of arsenic, the specific activity was $a_1 = 5.3 \times 10^6$ s^{-1}. The relative error of the arsenic content in the sample should be estimated supposing that the relative error of weighing is $\delta(m) = 0.03\%$, and the relative error of the activity measurement $\delta(a_1) = \delta(a_2) = 1\%$. The content of arsenic, m_x, in the samples is calculated by

$$G(\cdot) = m_x = m_1(a_1 - a_2)/a_2$$

Because this expression is not in the form of Eq. 10, Eq. 13 cannot be used. Assuming that the quantities m_1, a_1 and a_2 are not correlated, results obtained by three methods of the Propagation-of-Errors program are identical (Table 1).

Table 1
Analysis of indirect measurements in isotope dilution

	Taylor series expansion	Two points estimation	Monte Carlo simulation
$\bar{y}$ (g)	7.1122×10^{-5}	7.1124×10^{-5}	7.1142×10^{-5}
s (y)(%)	1.0130×10^{-6}	1.0132×10^{-6}	1.0433×10^{-6}
δ (y)(%)	1.42	1.42	1.47

4.2. Sample 2: error in indirect viscosity measurements

The viscosity of glycerol is calculated by the Stokes method from the following experimental data: the radius of the ball $r = 0.0112 \mp 0.0001$ cm; the density of the ball $d_0 = 1335$ kg m^{-3}, the density of glycerol $d = 1280$ kg m^{-3}, the trajectory $l_t = 31.23 \mp 0.05$ cm, the time $t = 62.1 \mp 0.2$ s, and the acceleration due to gravity $g = 9.801$ m s^{-1}. Viscosity, η, determined by the Stokes method is calculated from the expression

$$G(\cdot) = \eta = 2gr^2(d_0 - d)t/(9l_t)$$

Because this relation is not of the Eq. 10 type, the relative error cannot be calculated with the use of a simple relationship. Results obtained from three methods of the Propagation-of-Errors program are in good agreement (Table 2).

4.3. Sample 3: correlated errors in solution concentration

A mass (m) of 0.1 g zinc was dissolved in hydrochloric acid and diluted in a standard flask with a volume, V, of 1000 ml. A volume, V_1, of 100 ml of this solution was diluted to a volume, V_2, of 1000 ml. The sample for analysis was prepared by taking $V_3 = 5$ ml and diluting into $V_4 = 25$ ml. The concentration of the resulting sample and its relative error is calculated when the standard deviation of weighing, $s(m)$, is 0.3 mg and for the standard flasks $s(V) = s(V_2) = 0.2$ ml, $s(V_1) = 0.05$ ml, $s(V_3) = 0.005$ ml and $s(V_4) = 0.025$ ml. The concentration c is calculated from

$$G(\cdot) = c = mV_1V_3/(VV_2V_4)$$

Table 2
Analysis of indirect viscosity measurements

	Taylor series expansion	Two points estimation	Monte Carlo simulation
$\bar{y}$ (Pa s)	2.988×10^{-4}	2.988×10^{-4}	2.986×10^{-4}
$s(y)$ (Pa s)	5.443×10^{-6}	5.443×10^{-6}	5.304×10^{-6}
$\delta(y)$ (%)	1.82	1.82	1.78

Table 3
Analysis of indirect measurements in solution concentration

	Taylor series expansion	Two points estimation	Monte Carlo simulation
$\bar{y}$ (g dm^{-3})	2.00×10^{-3}	2.000×10^{-3}	2.000×10^{-3}
$s(y)$ (g dm^{-3})	6.732×10^{-6}	6.732×10^{-6}	6.853×10^{-6}
$\delta(y)$ (%)	0.34	0.34	0.34

Errors in volumes V_2 and V_4 are strongly correlated with errors of volumes V_1 and V_3.

The ideal case when correlation coefficients $r(V_1V_2) = r(V_3V_4) = 1$ is considered first, while other variables are uncorrelated. From Eq. 13 it is

$$\begin{aligned}\delta^2(c) \simeq {} & [s(m)/m]^2 + [s(V)/V]^2 \\ & + [s(V_1)/V_1]^2 + [s(V_2)/V_2]^2 \\ & + [s(V_3)/V_3]^2 + [s(V_4)/V_4]^2 \\ & - 2[s(V_1)/V_1][s(V_2)/V_2] \\ & - 2[s(V_3)/V_3][s(V_4)/V_4]\end{aligned}$$

and numerically $\delta(c) = 0.302\%$.

Then, consider that the correlations between V_1 and V_2,and between V_3 and V_4 are negligible, so that $r(V_1V_2) = r(V_3V_4) = 0$ and then $\delta(c) = 0.336\%$.

Eq. 5 allows the mean concentration $\bar{c}$ to be estimated

$$\begin{aligned}\bar{c} = {} & mV_1V_3/(VV_2V_4) + mV_1V_3[s^2(V)/(V^3V_2V_4) \\ & + s^2(V_2)/(V_2VV_4) + s^2(V_4)/(V_4VV_2)] \\ & - mV_3s(V_1)s(V_2)/(VV_2^2V_4) \\ & - mV_1s(V_3)s(V_4)/(VV_2V_4^2)\end{aligned}$$

where the first term is equal to 2×10^{-6} g cm^{-3}, the second 2.16×10^{-12} g cm^{-3} and the third is 2.2×10^{-12} g cm^{-3}. If the two smaller terms are neglected the mean concentration will be $\bar{c} = 2\times10^{-6}$ g cm^{-3} or 2×10^{-3} g dm^{-3}. Results obtained from the Propagation-of-Errors method for a case $r(V_1V_2) = r(V_3V_4) = 0$ are presented in Table 3. Correlation between volumes V_1 and V_3 and also between V_2 and V_4 diminishes the relative error of the resulting sample concentration.

5. Conclusion

The Propagation-of-Errors program in CHEMSTAT applies three different approaches to analysis of indirect measurements. All three methods calculate the mean $\bar{y}$, standard deviation $s(y)$ and variation coefficient $\delta(y)$ and lead practically to the same results. Application of the algorithm is simple, easy and quite convenient for analytical chemists but also physicists and biologists.

References

[1] M. Meloun, J. Militký and M. Forina, Chemometrics for Analytical Chemistry, Part 1, PC-Aided Statistical Data Analysis, Ellis Horwood, Chichester, 1992.
[2] L.M. Schwartz, Anal. Chem., 47 (1975) 963.
[3] B.F.J. Manly, Biomed. J., 28 (1986) 949.