DIVERGENCE JUDGEMENT OF RELIABILITY OF ANALYTICAL METHODS*

K. ECKSCHLAGER

Institute of Inorganic Chemistry,
Czechoslovak Academy of Sciences, 250 68 Řež

Received July 5th, 1976

On the basis of the divergence measure a characteristic is defined enabling to judge the reliability of analytical methods whose results have a normal, logarithmic-normal, or Poisson distribution.

Analytical methods can be judged according to the reliability of their results and to how the reliability of the results depends on the conditions of the analysis and on the composition of the analysed sample. This, of course, is possible under the assumption that the reliability of the analytical results can be expressed by a suitable quantitative characteristic. The reliability of the analytical results is given by their simultaneous accuracy and correctness. Accurate are such results of parallel determinations which agree well among themselves; their accuracy is usually characterized by the standard deviation of the distribution of the results, σ, or by its relative value $\sigma_{\mathrm{rel}} = \sigma/\xi$, where ξ means the true content of the determined component. Such results which on the average agree with the true content of the determined component are then correct. Their correctness is characterized by the difference between the found most probable value, μ, and the true content of the determined component, *i.e.*, $|\xi - \mu|$, or better by the value of $|\xi - \mu|/\sigma$.

McFarren and coworkers[1] defined a so-called total error as a measure of the reliability of analytical results. Their definition was later modified[2,3] so that the total error represents a practically utilizable but not quite perfect characteristic of the reliability of results. In defining the total error, the true content of the determined component is considered as a known constant. We shall show the possibility of using the Kullback's divergence measure[4] in deriving another characteristic of the reliability of analytical results which has certain advantages as compared with the total error. Besides that, we take into account that the true content of the determined component is never known exactly, hence that it is to be considered as a random variable attaining values only little different from the mean value, ξ.

THEORETICAL

The judgement of the reliability of analytical results can be based on the determination of the dissimilarity of those two distributions which characterize, first, the true value as a random quantity and, second, the distribution of the analytical results. Such judgement can be done best with the use of the divergence measure[4,5], which

[*] Part XI in the series Theory of Information as Applied to Analytical Chemistry; Part X: This Journal *42*, 225 (1977).

we define to this purpose as

$$T(p_0, p) = \int_{x_1}^{x_2} p_0(x) \ln [p_0(x)/p(x)] \, dx \, . \tag{1}$$

The probability density, $p_0(x)$, characterizes the true content of the determined component with the mean value ξ, and the probability density $p(x)$ the distribution of the analytical results with the expected value μ and variance σ^2. The integration limits x_1 and x_2 define the narrowest interval for which $\int_{x_1}^{x_2} p_0(x) \, dx = 1$. The interval $\langle x_1, x_2 \rangle$ must be, of course, very narrow if it has to give a good approximation of the true value. Then we have

$$T(p_0, p) = \int_{x_1}^{x_2} p_0(x) \ln p_0(x) \, dx - \int_{x_1}^{x_2} p_0(x) \ln p(x) \, dx \tag{2a}$$

and

$$T(p_0, p) \approx -\ln p(\xi) + \int_{x_1}^{x_2} p_0(x) \ln p_0(x) \, dx \, . \tag{2b}$$

If we approximate the value of ξ by a very narrow rectangular distribution assuming $x_1 = \xi - \Delta x$, $x_2 = \xi + \Delta x$:

$$p_0(x) = \begin{cases} 1/2 \, \Delta x & \text{for} \quad x \in \langle x_1, x_2 \rangle \\ 0 & \qquad\quad \text{otherwise} \, , \end{cases} \tag{3}$$

we can the width of the interval, $x_2 - x_1 = 2\Delta x$, either consider as a small, in substance arbitrary constant, or set equal to the mass of the smallest particle (atom or molecule) of the determined component, namely $2 \, \Delta x = m/N_A$, where m denotes atomic or molecular mass of the determined component and N_A Avogadro's number ($6{\cdot}0225 \, . \, 10^{23} \, \text{mol}^{-1}$). An extreme case of an infinitely narrow rectangular distribution is the Dirac function $\delta(x)$ known from the quantum mechanics[6], which has the property that $\delta(x) = \infty$ for $x \neq 0$, is equal to zero for other x values, and $\int_{-\infty}^{+\infty} \delta(x) = 1$. Its use in the approximation of the true value of ξ, however, would be at variance with our concept that the true content of the determined component is a random quantity and moreover it would lead to an infinite value of $T(p_0, p)$ according to Eq. (*2b*). We could approximate the true value also by a very narrow normal distribution

$$p_0(x) = \frac{1}{\delta_0 \sqrt{(2\pi)}} \exp \left[- \frac{1}{2} \left(\frac{x - \xi}{\sigma_0} \right)^2 \right], \tag{4}$$

where we can set σ_0 equal either to a very small constant or to $m/2z_\alpha N_A$, where z_α is the critical value of the normal distribution on the significance level α. The best way is obviously to approximate the true value, ξ, by the rectangular distribution (3), which is justified in view of the fact that the concentration x does not strictly represent a continuum since mass is quantized.

In the case of the determination of higher contents of a component, where we assume a normal distribution of the results of the analyses

$$p(x) = \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad (5)$$

for which $\int_{-\infty}^{+\infty} p(x)\,\mathrm{d}x = 1$, the divergence characteristic of the reliability of the analyses is given as

$$T(p_0, p) = \ln\frac{\sqrt{(2\pi)}}{2\,\Delta x} + \ln\sigma + \frac{1}{2}\left(\frac{\xi-\mu}{\sigma}\right)^2 . \qquad (6)$$

Here we can set $\ln\left(\sqrt{(2\pi)}/2\,\Delta x\right) = T = \text{const.}$; this can be, *e.g.*, achieved by setting $2\,\Delta x = \text{const.} = \sqrt{(2\pi)}\,\mathrm{e}^{-n}$, so that $T = n$, or $2\,\Delta x = m/N_\mathrm{A}$, so that $T = \ln .$ $.\,(N_\mathrm{A}\sqrt{(2\pi)}/m)$. For small μ values, for example in the case of trace analyses, we assume rather a logarithmic-normal distribution

$$p(x) = \frac{1}{x\sigma_{\ln}\sqrt{(2\pi)}} \exp\left[-\frac{1}{2}\left(\frac{\ln x - \ln\mu}{\sigma_{\ln}}\right)^2 . \qquad (7)$$

where the standard deviation $\sigma_{\ln}$ has the character of a relative value. Here we have $\int_0^\infty p(x)\,\mathrm{d}x = 1$ and the divergence characteristic

$$T(p_0, p) = T + \ln(\xi\sigma_{\ln}) + \frac{1}{2}\left(\frac{\ln\xi - \ln\mu}{\sigma_{\ln}}\right)^2\right], \qquad (8)$$

where T can be determined as in the preceding case. The results of radiometric and X-ray spectral determinations are obtained in the form of whole-numbered values $(0, 1, 2, 3, \ldots)$ of the intensity of the analytical signal, and $\sum_{x=0}^{\infty} P(x) = 1$. Such results have the Poisson distribution

$$p(x) = \lambda^x\,\mathrm{e}^{-\lambda}/x!\,, \qquad (9)$$

which can be for $\lambda \geqq 20$ approximated by a normal one according to Eq. (5) with the mean value $\mu = \lambda$ and variance $\sigma^2 = \lambda$. Then $\int_{-\infty}^{+\infty} p(x)\,\mathrm{d}x = 1$ and the divergence characteristic of the reliability of results is

$$T(p_0, p) = T + \ln \sqrt{\mu} + \frac{(\xi - \mu)^2}{2\mu}. \qquad (10)$$

The value of T is determined analogously as in the preceding cases. Eqs (*6*) and (*8*) calculated with the use of the approximation (*2b*) are comparable with (*8*) and (*12*) in ref.[5] derived without any approximation.

RESULTS AND DISCUSSION

The divergence characteristic acquires the smaller value the more reliable are the results of analysis: the value of $T(p_0, p)$ decreases with increasing absolute accuracy, *i.e.*, with decreasing value of σ, $\xi\sigma_{\ln}$ or $\sqrt{\mu}$, and at the same time decreases with diminishing difference $|\xi - \mu|$, which is a measure of the correctness of results. If $\xi = \mu$, and σ for a normal, $\xi\sigma_{\ln}$ for a logarithmic-normal, or $\sqrt{\mu}$ for a Poisson distribution approximated by a normal one is equal to unity, then $T(p_0, p) = T$. If we choose T not too small (*e.g.*, $T = 50$), or if we set $T = N_A \sqrt{(2\pi)}/m$, then always $T(p_0, p) > 0$. The values of $T(p_0, p)$ for $T = 50$ and $T = N_A \sqrt{(2\pi)}/m$, for several values of the molecular mass m, various values of $a = |\xi - \mu|/\sigma$, $a = |\ln \xi - \ln \mu|/\sigma_{\ln}$, or $a = |\xi - \mu|/\sqrt{\mu}$, and for various values of $b = \sigma$, $b = \xi\sigma_{\ln}$ or $b = \sqrt{\mu}$ are given in Table I. It is seen that the value of $T(p_0, p)$ is influenced mainly by the variables a and b; as far as $T = N_A \sqrt{(2\pi)}/m$ is concerned, the molecular mass influences only the "shift" of the origin, *i.e.*, the value of $T(p_0, p)$ for $a = 0$, $b = 1$, and this in the range of only several units if m changes by three orders of magnitude.

In contrast to the so-called total error introduced by McFarren and coworkers[1] and in contrast to the modified total error[2,3], the divergence characteristic has an advantage in that it is based on the same measure as the most general equation used in expressing the information content of results of analyses[5], and hence it is related to the information properties of these results. In the value of the divergence characteristic, in contrast to the total error[2,3], the accuracy and correctness of results is manifested in the whole range of its applicability quite evenly. The value of the divergence characteristic is not dependent on the number of determinations carried out; it is, of course, necessary that the values of σ, $\sigma_{\ln}$ or $\sqrt{\mu}$ be determined from a sufficiently large number of analyses. In contrast to the total error[2], it does not show a discontinuity in dependence on a or b. It enables, unlike the total error[1-3], with which always only the normal distribution of results is taken into consideration, to characterize even the reliability of results having another (*e.g.*, a logarithmic-normal or Poisson) distribution. Since analytical expressions for the integral on the right-hand side of Eq. (*2b*) were for most of the common continuous distributions given by Peters[7], the use of Eq. (*2b*) is simplified to a mere substitution.

TABLE I

Values of $T(p_0, p)$ in Dependence on $a = |\xi - \mu|/\sigma$, $a = |\ln \xi - \ln \mu|/\sigma_{\ln}$ or $a = |\xi - \mu|/\sqrt{\mu}$ and on $b = \sigma$, $b = \xi\sigma_{\ln}$ or $b = \sqrt{\mu}$ and on Molecular Mass of Determined Component m

a			b		
	$1 . 10^{-9}$	$1 . 10^{-6}$	$1 . 10^{-3}$	$1 . 10^{-1}$	1
			$m = 50$		
0	31·039	37·946	44·854	49·459	51·762
1	31·539	38·446	45·354	49·959	52·262
2	33·039	39·946	46·854	51·459	53·762
3	35·539	42·446	49·354	53·959	56·262
5	43·539	50·446	57·354	61·959	64·262
10	81·039	87·946	94·854	99·459	101·762
			$m = 100$		
0	30·882	37·789	44·697	49·302	51·605
1	31·382	38·289	45·197	49·802	52·105
2	32·882	39·789	46·697	51·302	53·605
3	35·382	42·289	49·197	53·802	56·105
5	43·382	50·289	57·197	61·802	64·105
10	80·882	87·789	94·697	99·302	101·605
			$m = 500$		
0	28·736	35·634	42·551	47·156	49·459
1	29·236	36·143	43·051	47·656	49·959
2	30·736	37·643	44·551	49·156	51·459
3	33·236	40·143	47·051	51·656	53·959
5	41·236	48·143	55·051	59·656	61·959
10	78·736	85·643	92·551	97·156	99·459
			$m = 50000$		
0	24·131	31·029	37·946	42·551	44·854
1	24·631	31·529	38·446	43·051	45·354
2	26·131	33·029	39·946	44·551	46·854
3	28·631	35·529	42·446	47·054	49·354
5	36·631	43·529	50·446	55·051	57·354
10	74·131	81·029	87·946	92·551	94·854
			$T = 50$		
0	29·277	36·175	43·092	47·697	50·000
1	29·777	36·675	43·592	48·197	50·500
2	31·277	38·175	45·092	49·697	52·000
3	33·777	40·675	47·592	52·187	54·500
5	41·777	48·675	55·592	60·197	62·500
10	79·277	86·175	93·092	97·697	100·000

A quite new criterion can be, if considered suitable, the influence of the molecular or atomic mass of the determined component. Although the contemporary accuracy of analytical methods, characterized by the value of b referred to the mass of the determined component, is by many orders of magnitude larger than mass of a single particle, the number of particles corresponding to, *e.g.*, a unit value of b can vary by $4-5$ orders of magnitude if we take into account the limiting cases of lightest atoms and largest molecules of proteins. This circumstance is for the real accuracy of the determination not quite negligible and can be adequately expressed also by the value of $T(p_0, p)$.

The divergence characteristic $T(p_0, p)$ enables equally as the total error to determine a more advantageous procedure in a manner shown already earlier[3].

The author is indebted to Dr M. Šolc from the same Institute for stimulating discussions.

REFERENCES

1. McFarren E. A., Lishka R. J., Parker J. B.: Anal. Chem. *42*, 358 (1970).
2. Eckschlager K.: Anal. Chem. *44*, 878 (1972).
3. Eckschlager K.: This Journal *39*, 1426 (1974).
4. Kullback S.: *Information Theory and Statistics*, p. 6. Wiley, New York 1959.
5. Eckschlager K.: Fresenius' Z. Anal. Chem. *277*, 1 (1975).
6. Dirac P. A. M.: *The Principles of Quantum Mechanics*, 4th Ed., p. 216. Clarendon Press, Oxford 1958.
7. Peters J.: *Einführung in die Allgemeine Informationstheorie*, p. 195. Springer, Berlin—Heidelberg—New York 1967.

Translated by K. Micka.