INFORMATION CONTENT OF ANALYTICAL RESULTS SUBJECT TO SYSTEMATIC ERROR*

Karel ECKSCHLAGER

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

Received March 2nd, 1978

The information content of results of analyses, subject to a systematic error, is expressed with the use of the divergence measure. It is shown how this error lowers the information content of the results. The significance of methods enabling to eliminate the systematic error or to diminish it (calibration, blank experiment, *etc.*), mostly at the expense of the accuracy of the results, is discussed.

In our earlier communications[1-7] we discussed the information content of results of quantitative analyses, expressed with the use of the Kullback's[8] divergence measure, for the cases of more accurate analyses[1,7], trace analyses[2,3], instrumental analyses[4], and analyses for the purpose of the quality control[5]. Various measures of the information content of quantitative analytical results were compared recently[6]. Up to now, we assumed in all cases that the results were unbiased, *i.e.*, they were not subject to a systematic error. In practice, however, we must be sometimes satisfied with results that are not quite correct, or we must use some steps, such as subtracting the blank experiment[9], to eliminate or at least diminish the systematic error.

The influence of the latter, or of the steps serving for its elimination, on the information content of the results forms the subject of the present work, where the information content is expressed with the use of the divergence measure.

THEORETICAL

Quantitative analysis is a process of obtaining information about a certain, although unknown content, ξ, of a component to be determined in an analysed sample. The information which we obtain, *i.e.*, the result is a continuous random variable, φ. In an ideal case, the process of obtaining information should proceed so that $E[\varphi] = \xi$, *i.e.*, the most probable value of $E[\varphi]$ found by analysis should not differ from the true content, ξ, of the determined component, or the results should be

* Part XIII in the series Theory of Information as Applied to Analytical Chemistry; Part XII: This Journal *43*, 231 (1978).

correct. In practice, however, this condition is sometimes not fulfilled and the results are subject to a systematic error $\delta = \xi - E[\varphi]$.

The information content of the analytical results can be expressed with the use of the divergence measure[8] as

$$I(p_i, p_{i-1}) = \int_{-\infty}^{+\infty} p_i(x) \ln \frac{p_i(x)}{p_{i-1}(x)} \, dx \,, \tag{1}$$

where $i = 1, 2, \ldots$, $p_{i-1}(x)$ is the *a priori* probability density, which characterizes our preliminary knowledge or assumptions about the value of ξ, and $p_i(x)$ is the *a posteriori* probability density, which characterizes the probability distribution of the analytical results. Prior to the analysis, we assume usually a rectangular *a priori* distribution:

$$p_0(x) = \begin{cases} 1/(x_2 - x_1) & \text{for} \quad x \in \langle x_1, x_2 \rangle \\ 0 & \text{for other values of } x \,. \end{cases} \tag{2}$$

The distribution of the results of analyses is, as a rule, normal. If these results are unbiased, *i.e.*, if $\xi = E[\varphi]$, we can write

$$p_1(x) = \frac{1}{\sigma_1 \sqrt{(2\pi)}} \exp\left[-\frac{1}{2}\left(\frac{x - \xi}{\sigma_1}\right)^2\right] \tag{3a}$$

and the information content according to the Kullback's measure

$$I(p_1, p_0) = \int_{-\infty}^{+\infty} p_1(x) \ln \frac{p_1(x)}{p_0(x)} \, dx = \ln \frac{x_2 - x_1}{\sigma_1 \sqrt{(2\pi e)}} \,. \tag{4a}$$

If, however, the results are subject to a certain systematic error, *i.e.*, if they are biased, then the *a posteriori* probability distribution is given as

$$p_2(x) = \frac{1}{\sigma_1 \sqrt{(2\pi)}} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma_1}\right)^2\right], \tag{3b}$$

where the systematic error $\delta = \xi - \mu$ is different from zero. The information content according to the Kullback's measure must then be calculated as

$$I(p_2, p_0) = \int_{-\infty}^{+\infty} p_1(x) \ln \frac{p_1(x)}{p_0(x)} \, dx - \int_{-\infty}^{+\infty} p_2(x) \ln \frac{p_2(x)}{p_1(x)} \, dx \,, \tag{4b}$$

since the true information content is smaller than $I(p_1, p_0)$, namely by $I(p_2, p_1)$ according to (*4a*). Eq. (*4b*) can be rewritten by expressing the logarithms of fractions as differences of logarithms; by further calculation we obtain

$$I(p_2, p_0) = \ln \frac{x_2 - x_1}{\sigma_1 \sqrt{(2\pi)}} - \frac{1}{2\sigma_1^2} \int_{-\infty}^{+\infty} (x - \xi)^2 \, p_2(x) \, dx \, .$$

Since the latter integral is equal to $\sigma_1^2 + (\xi - \mu)^2 = \sigma_1^2 + \delta^2$, we have

$$I(p_2, p_0) = \ln \frac{x_2 - x_1}{\sigma_1 \sqrt{(2\pi e)}} - \frac{1}{2} \left(\frac{\delta}{\sigma_1} \right)^2 . \qquad (4c)$$

The information content of results subject to a systematic error δ is hence given by Eq. (*4c*), which for $\delta = 0$ (unbiased results) takes the form of (*4a*), derived under the assumption of unbiased results,

DISCUSSION

It is apparent that the information content of results subject to a systematic error decreases markedly with increasing value of δ; for example, for $x_2 - x_1 = 100$ and $\sigma_1 = 0{\cdot}005$ the dependence of $I(p_2, p_0)$ on δ is according to Eq. (*4c*) as follows:

δ	0	0·025	0·050	0·075	0·100	0·125	0·150
$I(p_2, p_0)$	6·182	6.057	5·682	5·057	4·182	3·057	1·682

TABLE I
Values of $I(p_2, p_0)$ for Different Values of σ_1 and δ/σ_1

δ/σ_1	σ_1 0·05	0·10	0·20
0·00	6·182	5·489	4·796
0·25	6·151	5·458	4·765
0·50	6·057	5·364	4·671
0·75	5·901	5·208	4·515
1·00	5·682	4·989	4·296
1·50	5·057	4·364	3·671
2·00	4·182	3·489	2·796

The dependence of the information content on the ratio of δ/σ_1 for various values of σ_1 is given in Table I. Similarly as, *e.g.*, in ref.[1] we can illustrate the dependence of $I(p_2, p_0)$ on σ_1 and δ by curves joining the points corresponding to equal values of $I(p_2, p_0)$ for different pairs of σ_1 and δ. These "isoinforms" are for several different information contents shown in Fig. 1, whence the points for $\delta = 0$ and the pronounced maximums of these curves are well apparent.

In the analytical work, we mostly want to eliminate or at least diminish the systematic error either by subtracting the blank experiment, or by a suitable calibration, *etc.* Let us consider a systematic error $\delta < 0\,(\mu > \xi)$ that can be eliminated or at least diminished by subtracting the blank experiment, which is however possible only at the expense of the accuracy of the results[9]. The result of the analysis is then found from the difference $x_D - x_B$, where x_D is the result of the determination proper carried out with the sample, and x_B is the result of the blank experiment. Since the determinations of x_D and x_B are mutually independent, we have $\sigma = \sqrt{(\sigma_D^2 + \sigma_B^2)}$ and for the case where $\sigma_D = \sigma_B$ we have $\sigma = \sigma_D\sqrt{2}$. To attain the highest possible information content of the results, it is recommended to subtract the blank experiment (thereby diminishing the systematic error but increasing the original value of σ_D to $\sigma_D\sqrt{2}$) only if

$$\ln \frac{x_2 - x_1}{\sigma_1\sqrt{2\pi e}} - \frac{1}{2}\left(\frac{\delta}{\sigma_1}\right)^2 \leqq \ln \frac{x_2 - x_1}{2\sigma_1\sqrt{\pi e}}\,;$$

i.e. if $-\delta \geqq \sigma_D\sqrt{\ln 2} \approx 0{\cdot}83\sigma_D$ and $x_B \approx -\delta$.

Every experimentally determinable value of the blank experiment is larger than the determination limit; in the case where we accept the Kaiser's definition of the determination limit[10], we have always $x_B \geqq 3\sigma_B$. Therefore, it is preferable to subtract every experimentally detectable nonzero value of the blank experiment even in

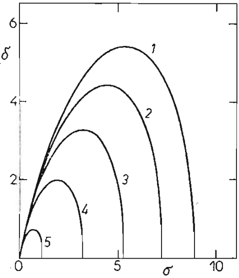

FIG. 1

Isoinforms

1 Information content 1·0; *2* 1·2; *3* 1·5; *4* 2·0; *5* 3·0.

cases where the blank experiment did not completely remove the systematic error, *i.e.* when $3\sigma_B \leqq x_B < -\delta$. The decrease of the information content of the results by the increase in σ_D due to subtracting the blank experiment will be always negligible against the gain of the information content caused by diminishing or even eliminating the systematic error.

We can proceed analogously also in the case of choosing the method of calibration; since different methods of calibration are differently time-consuming, we shall choose such one that has for the given procedure the highest information performance[11]

$$L(p_2, p_0) = \frac{1}{t_A} I(p_2, p_0), \qquad (5)$$

where t_A denotes duration of the analysis including calibration and calculations necessary to evaluate the results, and the information content $I(p_2, p_0)$ is given by Eq. (*4c*).

REFERENCES

1. Eckschlager K., Vajda I.: This Journal *39*, 3076 (1974).
2. Eckschlager K.: This Journal *40*, 3627 (1975).
3. Eckschlager K., Štěpánek V.: Mikrochim. Acta *1978*, I, 107.
4. Eckschlager K.: This Journal *41*, 1875 (1976).
5. Eckschlager K.: This Journal *43*, 231 (1978).
6. Eckschlager K.: This Journal *42*, 225 (1977).
7. Eckschlager K.: Z. Anal. Chem. *227*, 1 (1975).
8. Kullback S.: *Information Theory and Statistics*. Wiley, New York 1959.
9. Eckschlager K.: Chem. Listy *61*, 592 (1967).
10. Kaiser H.: Z. Anal. Chem. *203*, 1 (1965); *216*, 80 (1966).
11. Eckschlager K.: Anal. Chem. *49*, 1265 (1977).

Translated by K. Micka.