INFORMATION CONTENT OF QUANTITATIVE ANALYSIS*

K. ECKSCHLAGER

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

Received July 30th, 1975

The possibility of the determination of the information content of quantitative analysis is discussed in the case where the result of the analysis does not substantiate the original assumption about the composition of the analysed sample.

In our previous work the divergence measure has been made use of in the determination of the information content of more accurate analyses[1], trace analyses[2], and results of instrumental analyses[3]. This measure, which is characterized by the equation

$$I(p, p_0) = \int p(x) \ln [p(x)/p_0(x)] \, dx \,, \qquad (1)$$

where $p_0(x)$ denotes the probability density of the distribution of the content of the determined component assumed prior to the analysis and $p(x)$ probability density of the distribution of the analytical results, has been assumed to represent a generally valid measure enabling to derive expressions which are specific for the particular cases of analyses. Moreover, this measure enables, in contrast to the Wiener's measure[4,5], to determine the information content even when the analytical results do not confirm the original assumption.

In the present article we shall discuss the latter case and delimit the range of utilizable results which are at variance with the expected composition of an analysed sample.

THEORETICAL

In a quantitative determination of higher contents of a certain component in an analysed sample, the distribution of the results of parallel determinations is, as a rule, normal, *i.e.*

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

where μ denotes the most probable result and σ standard deviation. Since prior to the analysis we do not know more about the content of the component to be

* Part IX in the series Theory of Information as Applied to Analytical Chemistry, Part VIII: This Journal *41*, 1875 (1976).

determined than that it is in an assumed range from x_0 to x_1, we can set $p_0(x)$ equal to the probability density of a rectangular distribution

$$p_0(x) = 1/(x_1 - x_0) \quad \text{for} \quad x \in \langle x_0, x_1 \rangle ,$$
$$p_0(x) = 0 \quad \text{for other } x \text{ values} . \tag{3}$$

The information content for the case where the result lies in the midle of the assumed interval, *i.e.*, that the preliminary assumption about the content of the determined component was substantiated, *eg.*, $x_0 + 3\sigma \leqq \mu \leqq x_1 - 3\sigma$, can be derived by introducing Eqs (2) and (3) into (1):

$$I(p, p_0) = \int_{x_0}^{x_1} p(x) \left[\ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2 \right] \mathrm{d}x + \ln (x_1 - x_0) \int_{x_0}^{x_1} p(x)\, \mathrm{d}x .$$

Since for $(x_1 - x_0) \geqq 6\sigma$ and $(x_0 + 3\sigma) \leqq \mu \leqq (x_1 - 3\sigma)$ we have $\int_{x_0}^{x_1} p(x)\, \mathrm{d}x \approx 1$ and $\int_{x_0}^{x_1} (x - \mu)^2\, p(x)\, \mathrm{d}x \approx \sigma^2$, and since $-\frac{1}{2} = \ln (1/\sqrt{\mathrm{e}})$, we obtain

$$I(p, p_0) = \ln \frac{x_1 - x_0}{\sigma\sqrt{2\pi \mathrm{e}}} , \tag{4}$$

a relation which could be obtained for the case under discussion by using the Wiener's measure as well. If, however, the result either does not substantiate the original assumption or it lies on the upper limit of the assumed interval, *i.e.*, $x_1 \geqq \mu \geqq x_1 - 3\sigma$, we can find the equation for the divergence measure as follows. We replace the zero value of $p_0(x)$ for $x > x_1$ by a small constant $p_1(x)$ different from zero, which we obtain by separating a quantity Δx from the original interval of width $(x_1 - x_0)$ and distributing it equally in the interval $\langle (x_1 - \Delta x), (\mu + 3\sigma) \rangle$ (Fig. 1). We then can write

$$I(p, p_0) = \int_{x_0}^{\mu + 3\sigma} p(x) \ln p(x)\, \mathrm{d}x - \left[\int_{x_0}^{x_1 - \Delta x} p(x) \ln p_0(x)\, \mathrm{d}x + \right.$$
$$\left. + \int_{x_1 - \Delta x}^{\mu + 3\sigma} p(x) \ln p_1(x)\, \mathrm{d}x \right] .$$

Since the probability density of the rectangular distribution $p_1(x)$ is given as $\mathrm{p}_1(x) = \Delta x/(x_1 - x_0)\,[(a + 3)\,\sigma + \Delta x]$, where $a = (\mu - x_1)/\sigma$, we have

$$I(p, p_0) = \ln \frac{1}{\sigma \sqrt{2\pi e}} - \left[\ln \frac{1}{x_1 - \Delta x - x_0} \int_{x_0}^{x_1 - \Delta x} p(x)\, dx + \right.$$

$$\left. + \ln \frac{\Delta x}{(x_1 - x_0)\left[(a + 3)\sigma + \Delta x\right]} \int_{x_1 - \Delta x}^{\mu + 3\sigma} p(x)\, dx \right].$$

The values of these integrals are tabulated[6]; if we write

$$\int_{x_0}^{x_1 - \Delta x} p(x)\, dx = \Phi\left(-a - \frac{\Delta x}{\sigma}\right) - \Phi\left(\frac{x_0 - \mu}{\sigma}\right),$$

$$\int_{x_1 - \Delta x}^{\mu + 3\sigma} p(x)\, dx = 1 - \Phi\left(-a - \frac{\Delta x}{\sigma}\right),$$

where we set $\Phi(\mu + 3\sigma) \approx 1$, the information content is given by

$$I(o, p_0) = \ln \frac{1}{\sigma \sqrt{2\pi e}} + \ln (x_1 - x_0 - \Delta x)\left[\Phi\left(-a - \frac{\Delta x}{\sigma}\right) - \Phi\left(\frac{x_0 - \mu}{\sigma}\right)\right] +$$

$$+ \ln \frac{(x_1 - x_0)\left[(a + 3)\sigma + \Delta x\right]}{\Delta x}\left[1 - \Phi\left(-a - \frac{\Delta x}{\sigma}\right)\right]. \qquad (5)$$

This equation seems to be complicated for practical purposes, but it can be in particular cases substantially simplified. For example, if $x_1 > \mu + 3\sigma$, then $\Phi[(x_1 - \Delta x)/\sigma] \approx 1$, and if at the same time $\Delta x \ll (x_1 - x_5)$, Eq. (5) takes the form of (4); in the opposite case, if $\mu > x_1$ and $a \geqq 3$, then $\Phi[(x_1 - \Delta x)/\sigma] \approx 0$ and Eq. (5) takes the form

$$I(p, p_0) = \ln \frac{1}{\sigma \sqrt{2\pi e}} + \ln \frac{(x_1 - x_0)\left[(a + 3)\sigma + \Delta x\right]}{\Delta x}.$$

Then it is essential how large is the portion Δx which we separate from the originally assumed rectangular interval $\langle x_5, x_1 \rangle$. If this portion is large, Eq. (5) does not take the form of (4) and will not be therefore its more general case; if Δx is chosen very

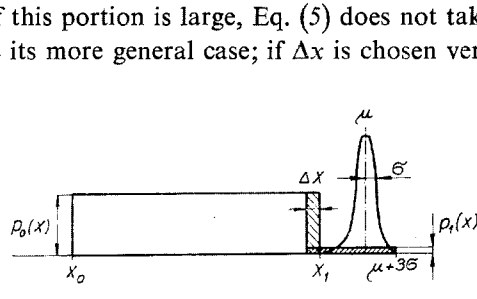

Fig. 1
Normal and Rectangular Distribution for $\mu > x_1$

small then the last term in Eq. (5) will grow very markedly with increasing $a = (\mu - x_1)/\sigma$.

If we choose $\Delta x = \sigma$, then, since almost always $\sigma \ll (x_1 - x_5)$, we have

$$\Phi\left(\frac{x_0 - \mu}{\sigma}\right) \ll \Phi\left(\frac{x_1 - \mu - \Delta x}{\sigma}\right) \approx \Phi\left(\frac{x_1 - \mu}{\sigma}\right)$$

and Eq. (5) takes the form

$$I(p, p_0) = \ln\left[(x_1 - x_0)/\sigma\sqrt{2\pi e}\right] + \left[1 - \Phi(-a - 1)\right] \ln(a + 4). \qquad (6)$$

For $\mu \leqq (x_1 - 3\sigma)$ this result takes the form of (4), and for $\mu > x_1$ and $a \geqq 3$ it is simplified to $I(p, p_0) = \ln\left[(x_1 - x_0)/\sigma\sqrt{2\pi e}\right] + \ln(a + 4)$. Since for $a > -3$ is $\ln(a + 4) < 0$, the information content is higher in the case where the result of the analysis does not substantiate the preliminary assumption than in the positive case, and this in a direct proportionality with $a = (\mu - x_1)/\sigma$.

In the case where the results have a Poisson distribution, which we approximate by a normal one with $\mu = \lambda$, $\sigma = \sqrt{\mu}$, Eqs (4) and (6) take the form

$$I(p, p_0) = \ln\left[(x_1 - x_0)/\sqrt{2\pi\mu e}\right] \qquad (7)$$

and

$$I(p, p_0) = \ln\left[(x_1 - x_0)/\sqrt{2\pi\mu e}\right] + \left[1 - \Phi(-a - 1)\right] \ln(b + 4), \qquad (8)$$

where $b = (\sqrt{\mu} - x_1)/\sqrt{\mu}$. It should be noted, however, that the Poisson distribution can be approximated by a normal one only when λ is large enough, *e.g.*, $\lambda \geqq 20$.

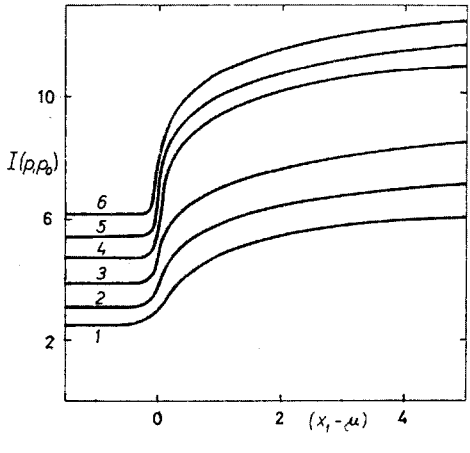

Fig. 2

Dependence of $I(p, p_0)$ on $\mu - x_1$ for Various Values of $x_1 - x_0$ and σ

1 $x_1 - x_0 = 10$, $\sigma = 0{\cdot}02$; *2* 10, 0·10; *3* 10, 0·05; *4* 5, 0·01; *5* 10, 0·01; *6* 20, 0·01.

DISCUSSION

The dependence of the information content on $\mu - x_1$ for different values of $x_1 - x_0$ and σ is shown in Fig. 2. It is obvious that in the vicinity of the point $\mu = x_1$ the value of $I(p, p_0)$ shows a most pronounced dependence on the difference $\mu - x_1$ and this the more accurate are the results of the analysis, hence the smaller is the value of σ.

As in ref.[1], we can define the units in which we shall express the information content according to Eq. (6) so that we determine the "isoinform" $I(p, p_0) = 1$ for various values of $\mu - x_1$ and σ and for the purpose of the definition of units we choose $x_1 - x_0 = 1$. Couples of the values of $\mu - x_1$ and σ corresponding to $I(p, p_0) = 1$ for $x_1 - x_0 = 1$ are given in Table I.

The information content is defined by Eq. (6) for cases where the found value either substantiates or exceeds the presumed interval. Analogously it would be possible to derive a similar measure even for $\mu < x_0$, but with respect to what was stated about the information content of trace analyses results[2] this case is not too probable in practice.

The use of Eq. (6) is, however, practically restricted. In making the analysis, we choose namely the conditions (*e.g.*, sample weight, calibration standards, sensitivity of the apparatus) according to the presumed content of the determined components. Hence, if the case $\mu \gg x_1$ occurs then the determination was done under unsuitable conditions, the results are not necessarily reliable, the determination must be repeated, and the information content must be determined in a manner described earlier[1]. It can be estimated that Eq. (6) may be used for $a \leqq 200$, *i.e.*, with $\sigma \leqq 0{\cdot}05$ up to a difference of $(\mu - x_1) \leqq 10\%$ and with $\sigma \leqq 0{\cdot}1$ to a difference $(\mu - x_1) \leqq 20\%$.

TABLE I

Couples of Values of $\mu - x_1$ and σ for which $I(p, p_0) = 1$ for $x_1 - x_0 = 1$

$\mu - x_1$	$\leqq -3{\cdot}0$	0·000	0·036	0·238
σ	0·089	0·179	0·200	0·300

The author is indebted to Dr J. Fusek from this Institute for calculating the values in Table I and those used in constructing the curves in Fig. 2.

REFERENCES

1. Eckschlager K., Vajda I.: This Journal *39*, 3076 (1974).
2. Eckschlager K.: This Journal *40*, 3627 (1975).
3. Eckschlager K.: This Journal, in press.
4. Wiener N.: *Kybernetika*. Published by SNTL, Prague 1960.
5. Eckschlager K.: Chem. Listy *69*, 810 (1975).
6. Janko J.: *Statistické tabulky*. Academia, Prague 1958.

Translated by K. Micka.