INFORMATION CONTENT OF TRACE ANALYSES RESULTS*

K. ECKSCHLAGER

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

Received July 4th, 1974

The information content of the results of trace analyses is discussed on using the general Kulback measure; a normal or logarithmic-normal distribution of positive results of parallel determinations is assumed. The information content of negative results of trace analyses can be found with regard to the limit of the determination of the analytical method used. The sequential criterion is considered to distinguish reliably whether the result of the determination carried out on the limit of the determination is really positive or negative, and a graphical test is proposed to distinguish the normal distribution from the logarithmic-normal one.

We pointed out in our preceding work[1] that it is purposeful to express the information content of the results of analyses with the aid of the general Kulback measure[2], which leads to expressions specific for the distribution before the analysis characterized by the probability density $p_0(x)$ as well as for different continuous distributions of the results obtained by the analysis and characterized by the probability density $p(x)$. Prior to the trace analysis, we assume that the true content of the determined component, ξ, is in the interval $\langle x_0, x_1 \rangle$, where, as a rule, $x_0 = 0$ and x_1 is the maximum assumed content of the determined component. Then the limit of the determination, x_2 (*i.e.*, the content of the determined component corresponding to the minimum analytical signal distinguishable from the noise; $x_0 \leqq x_2 \leqq x_1$), divides the interval of the assumed content $\langle x_0, x_1 \rangle$ into two parts. In the first one, $\langle x_0, x_2 \rangle$, no resolution is possible and if we do not obtain an analytical signal differing from the noise we know that $x_0 \leqq \xi < x_2$. In the other part, $\langle x_2, x_1 \rangle$, a continuous resolution of the values of $\mu \in \langle x_2, x_1 \rangle$ is possible. If we do not obtain an analytical signal and $\xi \in \langle x_0, x_2 \rangle$, we speak about a negative result; in the opposite case, when the signal can be distinguished from the noise, $\xi \in \langle x_2, x_1 \rangle$, we speak about a positive result.

The choice of a specific measure of the information content appears very important in the case of trace analyses, where we must distinguish whether the result is positive or negative, and in the positive case also whether the results have a normal or, as usual in trace analyses[3], a logarithmic-normal distribution. Since it is not

* Part VII in the series Theory of Information as Applied to Analytical Chemistry; Part VI: This Journal *39*, 3076 (1974).

always easy to distinguish whether a result close to the limit of determination is positive or negative, a criterion is proposed, which is in principle the sequential method[4] of the signal detection on a noise background[5], and since it is important to distinguish between a normal and a logarithmic-normal distribution, we use here the previously mentioned[6] graphical test of the distribution of results.

THEORETICAL

The Kulback divergence measure[2], which was shown[1] to be more general than the commonly used Shannon measure[7], is given as

$$I(p, p_0) = \int_{-\infty}^{+\infty} p(x) \log_b [p(x)/p_0(x)] \, dx \,, \tag{1}$$

where the base of the logarithm, b, is set equal either to 2 (binary logarithm) or e (natural logarithm), and where $p_0(x)$ is the probability density of the continuous distribution before and $p(x)$ after the analysis. In practice, the distribution before the analysis (when we know only that the content of the component to be determined is $\xi \leqq x_1$) is always rectangular with the probability density

$$p_0(x) = 1/(x_1 - x_0) \tag{2a}$$

for $x \in \langle x_0, x_1 \rangle$ and $p_0(x) = 0$ for other x values. In trace analysis we have, as a rule, $x_0 = 0$, hence

$$p_0(x) = 1/x_1 \,. \tag{2b}$$

On the other hand, the probability density $p(x)$ will be different according to whether we obtain a positive or a negative result, and if it is positive, also according to whether the results of parallel determinations are distributed normally or logarithmic-normally. In the case of a negative result, *i.e.*, for $x \in \langle x_0, x_2)$ we have

$$p(x) = 1/(x_2 - x_0) \tag{3a}$$

and for $x_0 = 0$

$$p(x) = 1/x_2 \,, \tag{3b}$$

where x_2 is the limit of the determination of the analytical method used. For the normal (Gaussian) distribution is

$$p(x) = \frac{1}{\sigma \sqrt{(2\pi)}} \exp\left[-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right], \tag{4}$$

where μ is the most probable value found usually as the arithmetic mean $\bar{x}_a$ from n parallel determinations, and σ^2 is the mean square deviation of the results usually estimated as $\hat{\sigma}^2 \equiv s^2 = \sum(x_i - \bar{x})^2/(n-1)$. When the distribution is logarithmic-normal, *i.e.*, the logarithms of the results rather than the results themselves have a normal distribution, then

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

Trace analyses are carried out also radiometrically or with the aid of X-ray diagrams. The discontinuous, whole-numbered results of these methods are subject to the Poisson distribution

$$p(x) = (\lambda^x/x!)\, e^{-\lambda}, \qquad (6)$$

where $x = 0, 1, 2, \ldots$ and λ denotes the mean value. For $x \geqq 15$, which is always the case in practice, the Poisson distribution can be approximated by the normal one (*4*) with $\mu = \lambda$ and $\sigma = \sqrt{\lambda}$. The λ value can be estimated as $\hat{\lambda} \equiv \bar{x} = \sum_{i=1}^{\mu} n_i x_i/n$, where $n = \sum_{i=1}^{\mu} n_i$.

If we introduce the quantities $p_0(x)$ from Eq. (*2a*) or (*2b*) and $p(x)$ from (*3*)–(*5*) according to the actual distribution of results into Eq. (*1*), we obtain after integration an expression for the information content of the results of trace analyses providing that they have a distribution characterized by the probability density $p(x)$. To determine this quantity, it is important to find out whether the result is positive or negative, and in the positive case to find out whether the results of parallel determinations have a normal or a logarithmic-normal distribution. The decision whether the result is positive or negative is difficult with results on the limit of the determination, where the analytical signal in repeated individual analyses only more or less exceeds the noise level. It was shown[4] that the sequential analysis is especially suitable for this purpose since it enables to decide on a chosen significance level α whether the result is positive or negative or whether another determination is necessary to find out whether the result lies above or below the limit of the determination. The testing can be based on the dependence of the quantity $\sum_{i=1}^{n} x_i/\sigma$ on the number of determinations, n. In the sequential method of testing it is important how we define the determination limit x_2. If we choose the so-called 3-sigma criterion according to Kaiser[8], then $1 - \alpha = 0{\cdot}9986$ and we proceed so that for $A_0 = \frac{1}{3}\log_e[(1-0{\cdot}9986)/0{\cdot}9986] + (x_2 + 3\sigma)\, n/2\sigma < \sum x_i/\sigma < \frac{1}{3}\log_e[0{\cdot}998/(1 - 0{\cdot}9986)] + (x_2 + 3\sigma)\, n/2\sigma = A_1$ we must carry out one more determination, for $\sum x_i/\sigma < A_0$ the result is negative, hence $p(x) = 1/x_2$, and for $\sum x_i/\sigma > A_1$ the

result is positive. If the result is positive, we must find out whether the results of parallel determinations have a normal or a logarithmic-normal distribution, which is easily possible with the aid of the graphical test[6].

Information Content of Negative Result

The Kulback's divergence measure for $p_0(x) = 1/(x_1 - x_0)$, $p(x) = 1/(x_2 - x_0)$, $x_0 \leqq x_2 \leqq x_1$ leads to the result

$$P(p, p_0) = \log_e [(x_1 - x_0)/(x_2 - x_0)] \tag{7a}$$

or for $x_0 = 0$

$$I(p, p_0) = \log (x_1/x_2) . \tag{7b}$$

It is obvious that for a certain assumed maximum content of the determined component, x_1, the information content depends only on the limit of the determination, x_2, namely so that it increases with decreasing x_2. Eqs (*7a, b*) involve neither the most probable value μ (since we even do not find it by the analysis), nor a characteristic of the accuracy of the results or the number of parallel determinations, since these quantities do not influence the information content of a negative result, which does not enable a distinction between different values of $\mu \in \langle x_0, x_2)$. Several values of $I(p, p_0)$ for different maximum contents x_1 and determination limits x_2 are given in Table I; it is seen that even a negative result can have a considerable information content if it is obtained by a sufficiently sensitive analytical method, which has a low limit of the determination[8,9].

Information Content of Positive Result

For the case of normally distributed positive results of trace analysis, we can use the measure of the information content proposed earlier[10]

$$I(p, p_0) = \log_b [(x_1 - x_0) \sqrt{n}/\sigma \sqrt{(2\pi e)}] , \tag{8}$$

or for the case of the estimate $\hat{\sigma} \equiv s = [\sum (x_i - \bar{x})^2/(n - 1)]^{1/2}$

$$I(p, p_0) = \log_b [(x_1 - x_0) \sqrt{n}/2st_{\alpha,\nu}] , \tag{9}$$

where the critical value of the Student distribution $t_{\alpha,\nu}$ is found from tables affixed to ref.[11] for $\alpha = 0{\cdot}039$ and $\nu = n - 1$. The results of trace analyses made close to the limit of the determination are, however, distributed mostly logarithmic-normally[3], *i.e.*, the probability density is given by Eq. (*5*). Then it is necessary to determine the relation for the general measure of the information content according

to Eq. (*1*) by expressing $p_0(x)$ from (*2a*) and $p(x)$ from (*5*). On integrating $I(p, p_0) = \int_0^\infty p(x) \{\log [(x_1 - x_0)/\sigma\sqrt{(2\pi)}] - \log x - (\log x - \log \mu)^2/2\sigma_\infty^2\} \, dx$ and setting $\int_0^\infty p(x) \log x \, dx = \log \mu$, $\int_0^\infty (\log x - \log \mu)^2 \, p(x) \, dx = \sigma^2$, $\int_0^\infty p(x) \, dx = 1$ we obtain

$$I(p, p_0) = \log_b [(x_1 - x_0)/\sigma\sqrt{(2\pi e)}] - \log \mu \,. \qquad (10)$$

Since the result of the trace analysis is given usually as an average from n parallel determinations, the parameter σ in Eq. (*10*) must be set equal to $\sigma/\sqrt{n}$, hence

$$I(p, p_0) = \log_b [(x_1 - x_0)\sqrt{n}/\mu\sigma\sqrt{(2\pi e)}] \,. \qquad (11)$$

If we want to use the estimate $\hat{\sigma} \equiv a = [\sum(\log x_i - \log \bar{x}_g)^2/(n-1)]^{1/2}$, where the geometric mean $\bar{x}_g$ is given as $\log \bar{x}_g = \sum \log x_i/n$, it is necessary to use another estimate $\hat{\sigma}$ for the values $x_i < \bar{x}_g(s_-)$ and another one for $x_i > \bar{x}_g(s_+)$; naturally $s_+ > s_-$. Both these values have a relative character.
Then

$$I(p, p_0) = \log_b [(x_1 - x_0)\sqrt{n}/\bar{x}_g(s_+ + s_-)\, t_{\alpha,\nu}] \,, \qquad (12)$$

where the value of $t_{\alpha,\nu}$ is found in tables[11] for $\alpha = 0{\cdot}039$ and $\nu = n - 1$. It is obvious that in the case of the trace analysis, where $\bar{x}_g \geqq 10^{-2}$ and therefore $\log_e (1/\bar{x}_g) \geqq$ $\geqq 4{\cdot}6$, $\log_2 (1/\bar{x}_g) \geqq 6{\cdot}6$, the value of $I(p, p_0)$ according to (*10*) and (*12*) is not unconsiderably influenced also by the found content of the component, μ or $\bar{x}_g$. This means practically that $I(p, p_0)$ is the larger the smaller is $\bar{x}_g$, or that the methods with a lower determination limit can yield results with a higher information content. This dependence of the information content on the content of the determined com-

TABLE I
Values of $I(p, p_0) = \log_2 (x_1/x_2)$

x_1	x_2 = 10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}
10^{-2}	16·610	13·288	9·966	6·644	3·322	0·000
10^{-3}	13·288	9·966	6·644	3·322	0·000	—
10^{-4}	9·966	6·644	3·322	0·000	—	—
10^{-5}	6·644	3·322	0·000	—	—	—
10^{-6}	3·322	0·000	—	—	—	—
10^{-7}	0·000	—	—	—	—	—

TABLE II

Values of $I(p, p_0) = \log_2 [x_1 \sqrt{n/\mu\sigma} \sqrt{(2\pi e)}]$ for $x_1 = 0{\cdot}01$

n	$\mu . \sigma$ 10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}
1	4·597	7·919	11·241	14·563	17·884
2	5·097	8·419	11·741	15·063	18·384
3	5·389	8·711	12·033	15·355	18·677
4	5·597	8·919	12·241	15·563	18·884
5	5·758	9·080	12·402	15·724	19·045
6	5·889	9·211	12·533	15·855	19·177
7	6·000	9·322	12·644	15·966	19·288
8	6·097	9·419	12·741	16·063	19·384
9	6·182	9·504	12·826	16·148	19·469
10	6·258	9·580	12·902	16·224	19·545
15	6·550	9·872	13·194	16·516	19·838
20	6·758	10·080	13·402	16·724	20·045
25	6·919	10·241	13·563	16·884	20·206
50	7·419	10·741	14·063	17·384	20·706
100	7·919	11·241	14·563	17·884	21·206

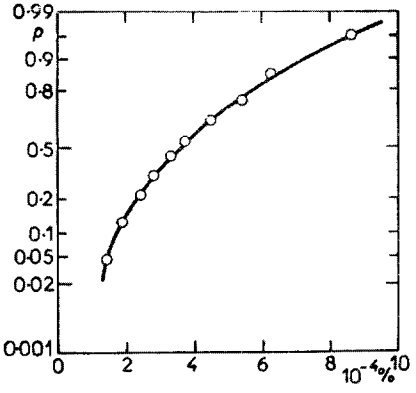

FIG. 1

Testing the Distribution of Results of Parallel Determinations by Plotting Cumulative Frequency against x_i

The nonlinear course suggests an asymmetrical distribution.

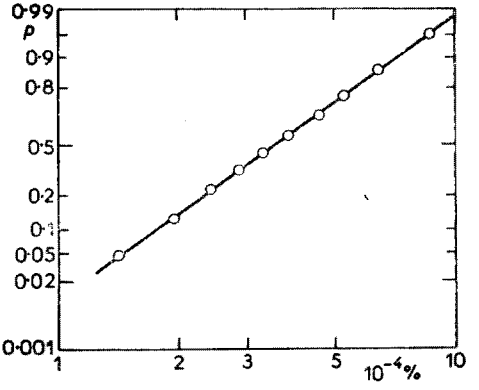

FIG. 2

Testing the Distribution of Results of Parallel Determinations by Plotting Cumulative Frequency against $\log x_i$

The linear course suggests a logarithmic-normal distribution.

ponent is not apparent from Eqs (*8*) and (*9*) derived under the assumption of a normal distribution of the results of parallel determinations. To obtain a really specific measure of the information content, we must therefore distinguish between the cases where the distribution of the trace analytical results in normal and where it is logarithmic-normal, and choose the measure of the information content accordingly either (*9*) (for a normal distribution) or (*12*) (for a logarithmic-normal distribution). Several values of $I(p, p_0)$ according to (*11*) for different numbers of parallel determinations, n, and for different values of $\mu\sigma$ are given in Table II.

An example of evaluating the information content of positive results of an emission-spectrographic determination of nickel in potassium hydroxide is shown in Table III; the calculations were done with the aid of the divergence measure $I(p, p_0)$. A graphical test[6] of the distribution is shown in Figs 1 and 2: By plotting the cumulative frequency against x_i and using a probability scale on the frequency axis we obtain a nonlinear dependence (Fig. 1), which suggests an unsymmetrical probability distribution, whereas by plotting the cumulative frequency against $\log x_i$ we obtain a linear dependence (Fig. 2) suggesting a logarithmic-normal distribution. From $\bar{x}_g = 3{\cdot}5 \,.\, 10^{-4}\%$ Ni, $s_+ = 6{\cdot}1 \,.\, 10^{-4}\%$, $s_- = 2{\cdot}0 \,.\, 10^{-4}\%$ and $x_1 = 0{\cdot}01\%$ follows according to Eq. (*12*) the information content in binary units as $I(p, p_0) = 15{\cdot}49$ bit.

TABLE III

Results of Emission-Spectrographic Determination of Ni in KOH

$x_i \,.\, 10^4$	$\log x_i \mid 4$	$(\log x_i - \log \bar{x}_g)^2$
1·4	0·146	0·159201
1·9	0·279	0·070756
2·4	0·382	0·026569
2·8	0·447	0·009604
3·3	0·518	0·000729
3·7	0·568	0·000529
4·5	0·653	0·011664
5·4	0·732	0·034969
6·2	0·792	0·061009
8·6	0·934	0·151321

$\Sigma \log x_i = 5{\cdot}451 - 40$ $\qquad$ $\Sigma(\log x_i - \log \bar{x}_g)^2 = 0{\cdot}526351$

$\log \bar{x}_g = 0{\cdot}545 - 4$ $\qquad$ $s = \sqrt{\tfrac{1}{9}0{\cdot}526351} = \pm 0{\cdot}2418$

$\bar{x}_g = 3{\cdot}5 \,.\, 10^{-4}$ $\qquad$ rel $s_+ = 1{\cdot}745$; rel $s_- = 0{\cdot}573$

$$s_+ = 1{\cdot}745 \,.\, 3{\cdot}5 \,.\, 10^{-4} = 6{\cdot}1 \,.\, 10^{-4}$$

$$s_- = 0{\cdot}573 \,.\, 3{\cdot}5 \,.\, 10^{-4} = 2{\cdot}0 \,.\, 10^{-4}$$

In the case where the result of the analysis is negative and the determination limit corresponds to the content of the determined component, $x_2 = 3{\cdot}5 \cdot 10^{-4}\%$, its assumed maximum content being again $x_1 = 0{\cdot}01\%$, the information content would be according to Eq. $(7b)$ $I(p, p_0) = 4{\cdot}84$ bit, *i.e.*, much smaller than in the preceding case of a positive result.

The author is indebted to Dr M. Šolc, Institute of Inorganic Chemistry, Czechoslovak Academy of Sciences, for a stimulating discussion of the problem of the information content of logarithmic -normally distributed results. The values in Tables I and II were calculated by Dr J. Fusek on a Gier type computer, Institute of Nuclear Research, Prague - Řež. The results in Table III are based on analyses of pure KOH *done by Mr O. Vahalík.*

REFERENCES

1. Eckschlager K., Vajda I.: This Journal *39*, 3076 (1974).
2. Kulback S.: *Information Theory and Statistics*, p. 27. Wiley, New York 1959.
3. Eckschlager K.: *Chyby chemických rozborů*, 2nd Ed., p. 130. Published by SNTL, Prague 1971.
4. Eckschlager K., Kodejš Z.: Chem. Prům. *24*, 400 (1974).
5. Hancock J. C., Wintz P. A.: *Signal Detection Theory*. McGraw-Hill, New York 1966.
6. Eckschlager K.: *Grafické metody v analytické chemii*. Published by SNTL, Prague 1966.
7. Shannon C. E.: Bell System Techn. J. *27*, 379, 623 (1948).
8. Kaiser H.: Fresenius' Z. Anal. Chem. *209*, 1 (1965).
9. Kaiser H.: Fresenius' Z. Anal. Chem. *216*, 80 (1966).
10. Eckschlager K.: This Journal *36*, 3016 (1971).
11. Eckschlager K.: Chem. Listy *69*, 810 (1975).

Translated by K. Micka.