AMOUNT OF INFORMATION OF REPEATED HIGHER PRECISION ANALYSES*

K. ECKSCHLAGER and I. VAJDA

Institute of Inorganic Chemistry,
Czechoslovak Academy of Sciences, 250 68 Prague - Řež, and
Institute of Information and Automation Theory,
Czechoslovak Academy of Sciences, 120 00 Prague 2

Received August 10th, 1973

A characteristic is proposed, based on Kullback's divergence measure which allows to determine the amount of information of repeated higher precision analyses.

The aim of chemical analyses is to obtain information on the composition of an analyzed sample or of the material from which the sample was taken. Gottschalk[1] even extended this conception to the basis of modern aspects of analytical chemistry. Recently, several papers have been published[2-7] dealing with the problem of the amount of information of the results of analyses.

The amount of information of analyses which were carried out without any *a priori* knowledge on the composition of the analyzed sample, or under the assumption that the content of the component to be determined lies within the interval (c_1, c_2), whereby the probability of the correct content is the same over the whole interval (the so-called rectangular distribution), was discussed in some papers[2-7]. In the first paper of the present series[3], the amount of information is expressed by the relation, based on Shannon's measure of information

$$I = \log [(c_2 - c_1)/[\sigma \sqrt{(2\pi e)}]], \qquad (1)$$

corresponding to the model shown in Fig. 1. The initial uncertainty before experiment, given by the rectangular distribution in the interval c_1 to c_2 is diminished by carrying out n parallel determinations to the uncertainty after experiment which is given by normal distribution of value x_i with the reliability interval $\mu \pm \frac{1}{2}\sigma \sqrt{(2\pi e)}$, where the value $\frac{1}{2}\sqrt{(2\pi e)} = 2{\cdot}066$ corresponds, for the number of degrees of freedom $\nu \to \infty$, to the significance level $\alpha = 0{\cdot}039$. Some consequences of this relation valid for practical performing of analyses were discussed in papers[4-6]. Since we assume, in the model shown in Fig. 1, that the found mean value μ lies within the initial uncertainty given by the interval (c_1, c_2) (*e.g.* 0—100%) it never appears in relation (*1*), as for $c_1 \leqq \mu \leqq c_2$ any value of μ represents the same information. However, the total amount of information obtained depends on the decrease of the uncertainty which is given by narrowing of the

* Part VI in the series Theory of Information as Applied to Analytical Chemistry; Part V: This Journal *39*, 1426 (1974).

interval $(c_2 - c_1)$ to the width given by the value of the expression $\sigma\sqrt{(2\pi e)}$. In analytical practice, however, we often have to do with a case when we have a certain knowledge about the composition of the analyzed sample obtained in preliminary analyses. In this case the amount of information obtained by means of a higher precision analysis should not be given only by the uncertainty before and after the experiment, but also by the fact whether the higher precision analysis verified or modified the initially found value. In this case the point in judging the amount of information is that the uncertainty before the experiment characterized by distribution p_0 of the preliminary results, is lowered so that its value after the experiment, *i.e.* after carrying out the higher precision analysis, is characterized by distribution p (Fig. 2). In this the following two cases must be further distinguished:

1) $\sigma < \sigma_0$, but $\mu = \mu_0$, *i.e.* the results of the new determination are more precise but the result remains unchanged, only the reliability interval of the result changes; *2)* $\sigma \leqq \sigma_0$, but $\mu \neq \mu_0$, *i.e.* the results of the new determination are more or equally precise, the reliability interval is narrower or it does not change, but at the same time the result changes, too. The case of $\sigma > \sigma_0$ need not be taken into account since it is excluded even by the character and purpose of the higher precision analyses. The amount of information of higher precision analyses cannot be judged using Shannon's relation[8] because it operates only with uncertainty based on precision and not with mean values.

The aim of the present paper is to choose from the possible criteria the one which would make it possible, in a manner most suitable for practical use, to determine the amount of information of higher precision analyses, under the assumption that the criterion will attain the higher values the greater the precision increase will be, *i.e.* the less the value σ/σ_0 and the greater the difference $|\mu_0 - \mu|$ will be found. It would be advantageous that the new criterion be a more general case of relation (*1*) and that its calculation be not complicated.

THEORETICAL

The aim of analyses is to obtain certain "information" as a basis for further decision. The more precisely the respective decision model is defined, the more adequate

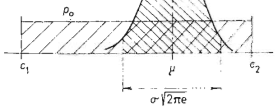

FIG. 1

Model of Analysis

Rectangular distribution p_0, normal p.

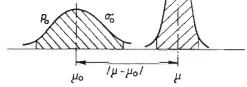

FIG. 2

Model of the Higher Precision Analysis

Normal distribution p_0; p; $\sigma_0 < \sigma$.

and more specific measure of the amount of information can be expected. In papers[9,10] we considered the most universal measure and therefore, as far as possible, no assumptions were made on the decision models. This, however, means that the adequacy of the respective measures has to be verified in each actual case anew. Since the present paper aims to express the amount of information, obtained in the higher precision analysis, we shall further limit ourselves to the prerequisite of normal distribution of the results of both the preliminary and the higher precision series of parallel determinations, *i.e.* to the case when it holds that the densities of both the initial and the higher precision distribution are

$$p_0(x) = [1/\sigma_0 \sqrt{(2\pi)}] \exp - [(x - \mu_0)^2/(2\sigma_0]\,, \tag{2a}$$

$$p(x) = [1/\sigma \sqrt{(2\pi)}] \exp - [(x - \mu)^2/2\sigma^2]\,. \tag{2b}$$

In paper[9] we have demonstrated that it is useful to look for the measure of information in the class of expressions

$$H(p, p_0) = \int_{-\infty}^{+\infty} \mathrm{f}\left(\frac{p(x)}{p_0(x)}\right) p(x) \,.\, \mathrm{d}x\,, \tag{3}$$

where f is a function convex from above. If we choose a logarithmic function, which is convex from above over the whole definition range, then the expression

$$I(p, p_0) = \int_{-\infty}^{+\infty} p(x) \,.\, \log_e \frac{p(x)}{p_0(x)}\, \mathrm{d}x \tag{4}$$

is the divergence measure introduced in paper[11] and discussed in detail in paper[12]. This measure is at the same time a more general case of the Shannon's relation

$$P = \int_{-\infty}^{+\infty} p(x) \,.\, \log_e p(x) \,.\, \mathrm{d}x \tag{5}$$

and for the case that $p(x)$ is a normal and $p_0(x)$ is a rectangular distribution it leads to relation (*1*). If we substitute into (*4*) for our model of the higher precision analyses for $p_0(x)$ from (*2a*) and for $p(x)$ from (*2b*), we obtain for the amount of information of the precisified analyses the following expression

$$I(p, p_0) = \frac{(\mu - \mu_0)^2 + \sigma^2 - \sigma_0^2}{2\sigma_0^2} + \log_e (\sigma_0/\sigma)\,. \tag{6}$$

In order to be able to judge the dependence of the new measure of the amount of information, $I(p, p_0)$, on the precision increase of the results and on the found

TABLE I

Values of $I(p, p_0)$ for Different $a = (\mu_0 - \mu)/\sigma_0$ and $b = \sigma/\sigma_0$

a	b				
	1·0	0·8	0·5	0·3	0·1
0·0	0·000	0·043	0·318	0·749	1·808
1·0	0·500	0·543	0·848	1·249	2·308
2·5	3·125	3·168	3·443	3·874	4·933
5·0	12·500	12·543	12·818	13·249	14·308
7·5	28·125	28·168	28·443	28·874	29·933
10·0	50·000	50·043	50·318	50·749	51·808

TABLE II

Values of $b = \sigma/\sigma_0$ and the Corresponding Values of $a = (\mu - \mu_0)/\sigma_0$ for $I(p, p_0) = 1$

a	0·00	0·71	1·00	1·17	1·27	1·38	1·41
b	0·23	0·30	0·40	0·50	0·60	0·80	1·00

difference of the expected value, we relate both distribution parameters $p(x)$, determined by higher precision analysis, to the scattering of the initial distribution $p_0(x)$, *i.e.* to σ_0^2 or σ_0 by substituting $b = \sigma/\sigma_0$ and $(\mu_0 - \mu)/\sigma_0 = a$ into relation (*6*). This is thus transformed into the form

$$I(p, p_0) = \frac{a^2 + b^2 - 1}{2} + \log_e (1/b) . \qquad (7)$$

The dependence of $I(p, p_0)$ on $0 < b = \sigma/\sigma_0 \leqq 1$ and for some values of $a = |(\mu_0 - \mu)/\sigma_0| \geqq 0$ is presented in Table I.

DISCUSSION

The proposed measure of the amount of information of higher precision analyses $I(p, p_0)$, as defined by relations (*6*) and (*7*) is based on Kullback's divergence measure (*4*). It depends on the precision increase of the results, characterized by the value $b = \sigma/\sigma_0$ as well as on the extent to which the higher precision analysis corrects the found value, *i.e.* on the value $a = (\mu_0 - \mu)/\sigma$. At the same time $I(p, p_0)$ increases with increasing value of a, at an increasing rate, and also increases with increasing

value of b, at the same rate for all a. This is obvious, if we rewrite expression (7) into the form $I(p, p_0) = \frac{1}{2}a^2 + [\log(1/b) + (b^2 - 1)/2]$, in which two terms appear of which one depends only on $a = (\mu_0 - \mu)/\sigma_0$ and the other only on $b = \sigma/\sigma_0$. Fig. 3 presents several curves connecting points of the same value of $I(p, p_0)$ for different values of a and b. These "isoinforms" illustrate how the amount of information depends on the difference $(\mu_0 - \mu)$ and on the ratio σ/σ_0; for great values of $a = (\mu_0 - \mu)/\sigma_0$ the change of b has no considerable effect. Only for small b, *i.e.* for a great precision increase of the results, a large amount of information can be obtained even for a small value of a. The values of a, b for $I(p, p_0) = 1$ are presented in Table II. From this table it is also evident that a unit of the amount of information of higher precision analyses is defined as the case when either no increase in the precision of the results was obtained, but the initial value found μ_0 was corrected by $\sigma\sqrt{2}$, or as the case when the higher precision analysis entirely verifies the initial value μ, but the precision of the new result changes so that the ratio $(\sigma/\sigma_0) = 0{\cdot}23$. It is useful to distinguish cases when the difference between the expected and the found value is or is not statistically significant. Therefore in Fig. 3 several "isoinforms" are compared with the region where the difference $\mu_0 - \mu$ is characterized by the value a and for different values of $b = \sigma/\sigma_0$ is of different statistical significance on the significance level $\alpha = 0{\cdot}05$, *i.e.* on the level on which most of the significances of the difference between two means are statistically tested. It is obvious that if the higher precision analysis yields a new value, which, however, is not statistical significantly different, *i.e.* if the higher precision analysis only verifies the result of the preliminary analysis, a relatively small amount of information is obtained (usually $I(p, p_0) \leqq 8$). If, however, the result of the higher precision analysis differs in its statistical significance from the preliminary result, the amount of information gained is considerably higher. The found properties of the quantity $I(p, p_0)$ according to expression (6)

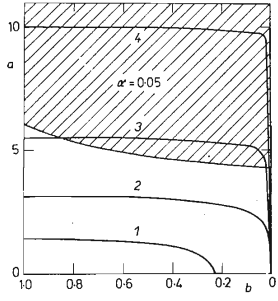

FIG. 3

Values of a, b for $I(p, p_0) = 1$, 5, 15 and 50
Dashed area: difference $|\mu - \mu_0|$ is for σ; σ_0 statistically significant on the significance level $\alpha = 0{\cdot}05$.

correspond to those which are intuitively expected from an adequate measure in Fig. 2. At the same time expression (*4*) from which (*6*) is derived, results, when applied to the model shown in Fig. 1, in relation (*1*), generally used for the mentioned case, as has been previously demonstrated[9]. Thus the divergence measure (*4*) represents a generalization of the Shannon's relation for the measure of information.

The values, given in Tables I and II and plotted in Fig. 3, were computed using a computer Gier in the computer center of the Institute of Nuclear Research, Prague - Řež. The programs were set up by Dr J. Fusek and Mr A. Petřina, to whom our thanks are due.

REFERENCES

1. Gottschalk G.: Z. Anal. Chem. *258*, 1 (1972).
2. Doerffel K., Hildebrand W.: Wiss. Z. TH "Carl Schorlemmer" Leuna *11*, 30 (1969).
3. Eckschlager K.: This Journal *36*, 3016 (1971).
4. Eckschlager K.: This Journal *37*, 137 (1972).
5. Eckschlager K.: This Journal *37*, 1486 (1972).
6. Eckschlager K.: This Journal *38*, 1330 (1973).
7. Doerffel K.: Chem. Tech. (Berlin) *25*, 94 (1973).
8. Shannon C. E.: The Bell System Technical Journal *27*, 379, 623 (1968).
9. Vajda I., Eckschlager K.: Kybernetika, in press.
10. Vajda I.: Period. Math. Hungar. *2*, 223 (1972).
11. Kullback S., Leibler R.: Ann. Math. Stat. *21*, 79 (1951).
12. Kullback S.: *Information Theory and Statistics*, p. 27. Wiley, New York 1959.

Translated by V. Čermáková.