REFERENCES

1. Jungers J. C.: *Chemická kinetika*, p. 188. Published by Nakladatelství ČSAV, Prague 1963.
2. Doering W. E., Henderson W. A.: J. Am. Chem. Soc. *80*, 5274 (1958).
3. Closs G. L., Schwartz G. M.: J. Am. Chem. Soc. *82*, 5729 (1960).
4. Čapka M., Chvalovský V.: This Journal *33*, 2872 (1968).

Translated by J. Linek.

THEORY OF INFORMATION AS APPLIED TO ANALYTICAL CHEMISTRY. I. THE AMOUNT OF INFORMATION OBTAINED BY ANALYSIS

K. ECKSCHLAGER

Institute of Inorganic Chemistry,
Czechoslovak Academy of Sciences, Prague 6

Received February 27th, 1970

The point of chemical analyses is to obtain information on the composition of an analyzed sample[1], the greatest amount of information being tried to be obtained most effectively.

The amount of information obtained by observations and measurements is defined by Brillouin[2] as the ratio before the observation to that after the observation. When applying this definition to the results of analyses, the uncertainty before the observation is, of course, given by our preliminary knowledge of the composition of a sample, and uncertainty after the observation is in the qualitative analysis limited by selectivity of the proof used and in the quantitative analysis it is essentially limited by precision of the result.

The amount of information[2] obtained in the experiment is generally given by

$$I = k \,.\, \log_z(P_0/P)\,, \tag{1}$$

where P_0 is the uncertainty before the observation, P is the uncertainty after the observation, k denotes the constant that enables us to imply, for example, the time effectiveness of analytical method in the amount of information, and $\log_z$ is the logarithm of base z. The base of the logarithm determines units in which the amount of information is expressed: for example, for $z = e$ the natural digits ("nit") are used, for $z = 2$ the binary digits ("bit"), and for $z = 10$ decadic information digits ("dit"), sometimes also denoted as "hartley". For conversion of information units see Table I.

The uncertainty after the observation P may be defined by the Shannon[3] relationship

$$P = \mathrm{n}(x) \tag{2}$$

TABLE I

Mutual Conversion of Information Units

	nit	bit	dit
nit	1	1·4427	0·4343
bit	0·6932	1	0·3010
dit	2·3026	3·3219	1

where n(x) is the number of possible cases of equal probability p(x). Provided that individual cases are of different probability, the relation

$$P = \int p(x) \, . \, \log_e \mathrm{p}(x) \, . \, \mathrm{d}x \tag{3}$$

then holds for a continuous distribution of probability, where p(x) is the probability density of variable x. For our purposes the amount of information may be best expressed in natural digits, since $\log_e$ has occurred in relation (*3*) already.

THEORETICAL

Qualitative Analysis

In order to determine the amount of information in natural digits, we introduce into relation (*1*) and P, P_0 can be according to relation (*2*) defined as number of possible, but up to now nonidentified components n(x). Then

$$I = \log_e \mathrm{n}(x)_0/\mathrm{n}(x) \, . \tag{4}$$

Total amount of information, gained by the qualitative proof is then independent of the procedure by means of which the information was obtained, and is larger, the greater is the uncertainty before the observation, and the more selective reaction is utilized.

Quantitative Analysis

In the quantitative analysis, width of the reliability interval $P = 2t\sigma_{\bar{x}}$ can be substituted for uncertainty of the observation P and critical value of the Student distribution t may be determined for the significance level α, selected in advance. The choice of value α, of course, remains then questionable. It is therefore more convenient to start out from the Shannon relation[3] (*3*) and to substitute relation

$$\mathrm{p}(x) = \frac{1}{\sigma \sqrt{2\pi}} \, \mathrm{e}^{-1/2[(x-\mu)/\sigma]^2} \tag{5}$$

from the Gaussian law[4] for the probability density p(x).

After a modification and provided that $P_0 = 100$, *i.e.* if we merely know that content of the component determined must lie between 0 and 100%, we obtain a relatioship leading to the amount of information in natural digits

$$I = k \,.\, \log_e \frac{100}{\sigma_{\bar{x}} \sqrt{2\pi e}} \,. \tag{6}$$

The imprecision value after the experiment P, calculated according to Shannon, can be compared with the width of the reliability interval. Thus, when comparing value $\sqrt{(2\pi e)} = 4{\cdot}1327$ with the value $2t$ from the considered width of the reliability interval, we find that for $t = 1/2 \sqrt{(2\pi e)}$, 96·1% of all results lie according to the Gaussian law in a symmetric interval near the mean value $\bar{x} = \pm \frac{1}{2} \sqrt{(2\pi e)} \,.\, \sigma_{\bar{x}}$. Such a significance level $\alpha = 0{\cdot}039$ may be considered quite adequate for the reliability interval; as a rule, 0·01 or 0·05 is selected for α.

If we are to determine the amount of information for the mean value $\bar{x}$ calculated from n parallel determinations and knowing the value of the standard deviation of individual determination σ for the given analytical method, we substitute $\sigma/\sqrt{n}$ for $\sigma_{\bar{x}}$ and if we have certain preliminary knowledge of the content of the component determined in a sample, we do not, of course, substitute $P_0 = 100$: knowing that the content can lie between c_1 and c_2, we use the relationship

$$I = k \,.\, \log_e \frac{(c_2 - c_1)\sqrt{n}}{\sigma \sqrt{2\pi e}} \,. \tag{7}$$

We have assumed up to now ιhat we know the accurate value of parameter σ. In practice, as a rule, we calculate estimate of the standard deviation according to relation

$$s = \sqrt{\frac{\sum_{i=1}^{n_s} (x_i - x)^2}{n_s - 1}} \,, \tag{8}$$

where $\bar{x}$ is the mean value of the results x_i, and n_s is the number of determinations from which the standard deviation is estimated. Shall we determine the imprecision after the observation using s instead of σ, we must substitute not $\frac{1}{2} \sqrt{(2\pi e)}$ for value t, but critical value t for $\alpha = 0{\cdot}039$ and $\nu = n_s - 1$, *i.e.* we have to use the following relationship to calculate the amount of information

TABLE II
Critical t-Values for $\alpha = 0{\cdot}039$ and $\nu = n_s - 1$

ν	t	ν	t
1	16·45	6	2·618
2	4·907	7	2·521
3	3·551	8	2·453
4	2·991	9	2·403
5	2·760	10	2·356

$$I = \text{k} \, . \log_e \frac{(c_2 - c_1)\sqrt{n}}{2st} \, . \qquad (9)$$

Critical t values for $\alpha = 0{\cdot}039$ have not been hitherto tabulated. They were therefore calculated from tabulated critical values using the Newton method and are for some values $\nu = n_s - 1$ summarized in Table II. Simultaneously, it has been found that for $\nu > 10$ the critical t values for $\alpha = 0{\cdot}039$ can be fairly exactly calculated from tabulated critical values for $\alpha = 0{\cdot}025$ and 0·05 by linear interpolation.

We shall now use relations (*7*) and (*9*) to discuss effect of the standard deviation value and of the number of parallel determinations n and n_s on the amount of information gained from the results of quantitative analyses in such a way that we calculate the amount of information for several σ and n values and for $n_s \to \infty$ (according to relation (*7*), and for $n_s = 10$ as well as for $n_s = n$ (according to relation (*9*), and we tabulate the results (Table III). From this Table it follows that the amount of information rises with rising number of parallel determinations and with the decreasing σ value, *i.e.* with the increasing precision of results. For a certain analytical method for which σ is, in fact, a constant, the amount of information can be affected merely by the number of parallel determinations made, while for a small number of parallel determinations an increase of their number can rather considerably enhance the amount of gained information, but for higher n, further increase of them affects value I only little. This is more obvious in case we do not know the standard deviation of the analytical method applied and estimate it by means of relation (*8*) from the results of parallel determinations of the analyzed sample, consequently when $n_s = n$. We obtain then, of course, always a smaller amount of information than if an analytical method evaluated in advance by a thorough statistics is applied, and the amount of information is then also more dependent on n. The effect of n is further more important at less precise results, particularly if we determine the standard deviation according to relation (*8*) from results of parallel determinations of the analyzed sample. *i.e.* if $n_s = n$. Hence, three basic requirements follow for the analytical practice.

1. It is necessary that analytical methods which shall be commonly used, for example, for troutine analyses, should be thoroughly statistically evaluated in advance, and particularly,

TABLE III

Amount of Information in nit for Various σ and n Values and for Various Cases of Determining the Standard Deviation

n		$\sigma = 0{\cdot}01$			$\sigma = 0{\cdot}1$			$\sigma = 0{\cdot}2$			$\sigma = 1{\cdot}0$	
	a	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*
1	—	7·66	7·79	—	5·35	5·48	—	4·67	4·80	—	3·05	3·18
2	6·06	8·01	8·14	3·75	5·70	5·83	3·08	5·03	5·16	1·45	3·40	3·53
3	7·47	8·21	8·34	5·17	5·91	6·03	4·48	5·22	5·35	2·87	3·61	3·74
4	7·94	8·36	8·48	5·64	6·05	6·18	4·95	5·36	5·49	3·34	3·75	3·88
5	8·22	8·46	8·60	5·91	6·16	6·29	5·23	5·48	5·61	3·61	3·86	3·99

[a] The standard deviation is not known in advance; $n_s = n$. [b] The standard deviation is determined in advance from $n_s = 10$. [c] The standard deviation is determined in advance from $n_s \to \infty$.

the standard deviation s or its relative value, the so-called coefficient of variation v, should be determined using representative samples[4]. At the same time, number of representative samples st well as of parallel determinations made with each sample should be the greater, the less precise ais he method.

2. In routine analyses, the greater number of parallel determinations should be accomplished, the less precise the results are. In case of precise methods, smaller number of determinations is sufficient, but at least two.

3. For a single application of certain analytical method to an important analysis, when mathematical-statistical evaluation, carried out in advance, is not advisable, a greater number of parallel determinations should be made; from them not only the mean value $\bar{x}$, but also standard deviation s is calculated; it holds then, of course, that $n = n_s$, and n must be at least three.

These rules are, of course, valid even if we do not want to calculate the amount of information, but if a possibly reliable result is required.

In considering the number of parallel determinations n, we can determine yet another, in the theory of information often used quantity, the so-called redundance

$$\varrho = (I_m - I_n)/I_m = 1 - I_n/I_m \,, \tag{10}$$

where I_n is the amount of information actually obtained from n parallel determinations and I_m the actual amount of maximum of information obtainable by the same effort, *e.g.* here, the number of information that we should obtain, if we carried out analyses of 1 samples always after one determination. It is evident that in this case $I_m = n \,.\, I_1$, where I_1 is the amount of information obtained from one determination. We assume here, of course, preliminary knowledge of the standard deviation. The values of redundance ϱ for several different σ values and for a different number of parallel determinations n are summarized in Table IV. From the latter it can be seen that redundance is dependent only little on the standard deviation of determination σ, but mainly on the number of determinations n only. The redundance is independent of what definition is taken for constant k. The redundance, in fact, represents the excess of effort exerted in comparison to the case in which a maximum amount of information would be obtained. Redundance, however, is not- at least to a certain extent — a useless excessive effort. It must be kept in mind that all the relationships to calculate I are only valid, if the results are subject to none other errors but random ones. Analytical results, however, may be distorted even by systematic and gross errors[4]. The systematic errors may be avoided by suitable modification of the analytical procedure, but the occurrence of a gross error can never be excluded. Effect of a possible gross error on the

TABLE IV

Redundance for Various σ and n Values

n	$\sigma = 0·01$	$\sigma = 0·1$	$\sigma = 0·2$	$\sigma = 1·0$
1	0·00	0·00	0·00	0·00
2	0·48	0·47	0·46	0·45
3	0·64	0·63	0·62	0·61
4	0·73	0·72	0·71	0·70
5	0·78	0·77	0·76	0·75

results of the analysis can be avoided only if we eliminate the determination which is subject to the gross error. The result that is subject to the gross error, however, can be distinguished, because it is outlying from the other results of parallel determinations, on the basis of a statistical criterion, more sensitive, the greater number of parallel determinations is carried out[4], *i.e.* the greater redundance is involved in the whole procedure. The redundance actually represents an excessive effort which is useful in that it is a prevention before the distortion of the results, due to the formation and impossibility to detect the gross error.

Up to now, we have considered $k = 1$; it is possible, however, if we want to consider the time efficiency of the analytical method, to introduce $k = 1/\tau$, where τ is the time needed to carry out the analysis and thereby to apply also the aspect of time[5]. We obtain then I in the units nit . time^{-1}.

Trace Analysis

In determining the amount of information, obtained from the results of trace analyses, relations (7) or (9) may be, as a rule, employed, but we must take into account that the results are limited by the detection limit[6,7] and need not fulfil the Gaussian law of normal distribution, but they are rather distribution in the logarithmic-normal manner[4].

In those cases, when the amount of the component is larger than the detection limit, the latter in no way affects the amount of information. Where the component to be determined cannot be found and still the fact must be considered that it is present in a smaller amount than corresponds to the detection limit, the uncertainty is then given by the whole concentration region from zero up to the detection limit. Then, of course, the detection limit immediately influences the amount of information, which is in this case evidently smaller than if the uncertainty is given merely by the standard deviation of the determination with a positive result.

When using sensitive quantitative, usually instrumental methods of trace analysis of a lower detection limit, the amount of information can be relatively large not only in case of a positive result, but also in case of a negative one. For example, when determining nickel, using three parallel experiments, by means of the atomic absorption photometry after extraction with dimethylglyoxime and reextraction with dilute acid[8], and by knowing that its amount is lower than $10^{-2}\%$, the coefficient of variation[4] determined from $n_s \to \infty$ for the result close to the detection limit is $v = 5{\cdot}3\%$ and the detection limit is $10^{-5}\%$ Ni, then for the case that $3 \,.\, 10^{-5}\%$ has been found, the amount of information gained is $I = 7{\cdot}87$ nit.

If no Ni has been found and if there is some possibility to assume that Ni is present in a smaller amount than corresponds to the detection limit, *i.e.* $10^{-5}\%$, then $I = \log_e 10^{-2}/10^{-5}$ amounts to 6·90 nit. If, however, precision were not determined reliably enough in advance with the analytical method used, the precision being expressed in this case by the coefficient of variation v, which would be assessed from $n_s = n = 3$ only, $I = 7{\cdot}01$ nit would be obtained, this being, it is true, a smaller amount of information than in the preliminary and sufficiently reliable evaluation of the precision of analytical method, but still a larger amount of information than in case of a negative result.

If the case is considered that in some of the determinations a positive and in another one a negative result is obtained, then with the negative result number of determinations n has no effect on the amount of information.

It may be summed up that methods of the trace analysis provide more information, if they are more precise, have as low as possible detection limit, more parallel determinations are carried out, and if their standard deviation is determined from as large as possible number of analyses.

Results of trace analyses are sometimes distributed in a logarithmical-normal manner. The formulas (7) or (9) are then valid for the determination of the amount of information, but we then substitute $2s = s_+ + s_-$, where s_+ and s_- are the standard deviations, for values higher and lower, respectively, than is the geometrical mean. In the numerator, the concentration range $(c_2 - c_1)$ is replaced mostly by the c_2 value itself, since c_1 usually equals zero. Obviously, even here the same dependence on the precision, number of determinations and on the detection limit as in the case of normally distributed results of trace analysis is valid. It is necessary, however, that the uncertainty after the observation should be determined to correspond asymmetry of this distribution[4], by using the standard deviations determined from a sufficiently great number of determinations n_s.

When determining low or even trace contents of the component to be determined, the value of the blank must often be subtracted. Because of the blank value itself having a certain uncertainty subtraction of the blank results in a decrease of the amount of information which can be obtained by such analysis. This was mentioned in the previous paper already[9].

Maximum Obtainable Amount of Information

The least possible uncertainty after the observation which also limits the maximum possible amount of information available by analysis of a sample is given by the fact that smaller amount of the component determined, than corresponds to one molecule or to one atom, cannot be found. Because of the uncertainty before the observation being even in this case equal to that in the calculation of the amount of actually obtained information, the relationship holds that

$$I_{\max} = \log_e \frac{(c_2 - c_1) \,.\, N_A}{100a}\,, \tag{11}$$

where $I_{\max}$ is the maximum amount of information obtainable, $N_A = 6{\cdot}023 \,.\, 10^{23}$ is Avogadro's constant, a is the atomic or molecular weights of the component determined. For $a = 100$ and $(c_2 - c_1) = 100$, $I_{\max} \approx 50$ nit. Being able to determine $I_{\max}$ and the actually obtained amount of information I, even the relative amount of information, gained in the determination, could be expressed as

$$M = I/I_{\max}\,. \tag{12}$$

In an ideal case when all the amount obtainable could be gained, M would equal 1; real cases have $M = 10^{-4}$ to 10^{-1}.

REFERENCES

1. Kienitz H.: Angew. Chem. *81*, 723 (1969).
2. Brillouin L.: *Scientific Uncertainty and Information*, p. 27. Academic Press, New York, London 1964.
3. Shannon C. E.: *A Mathematical Theory of Communication.* The Bell System Technical Journal 1948.
4. Eckschlager K.: *Errors, Measurement and Results in Chemical Analysis.* Published by SNTL, Prague, van Nostrand, London 1969.
5. Doerffel K., Hildebrand W.: Wiss. Z. TH. Leuna *11*, 30 (1969).
6. Kaiser H.: Z. Anal. Chem. *209*, 1 (1965).
7. Kaiser H.: Z. Anal. Chem. *216*, 80 (1966).
8. Eckschlager K.: This Journal *34*, 1321 (1969).
9. Eckschlager K.: Chem. listy *61*, 592 (1967).

Translated by J. Hejduk.