HOW TO ACHIEVE A HIGH INFORMATION GAIN OF RESULTS OF QUANTITATIVE ANALYSES*

Karel ECKSCHLAGER

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

Received February 2, 1990
Accepted April 28, 1990

Dedicated to Dr František Vláčil on the occasion of his 65th birthday.

The effect of procedure optimization, good laboratory practice and quality data control on the information gain of quantitative analyses and on its time stability is discussed.

The information gain of results of quantitative analysis is determined by their precision and accuracy as well as by how reliably these characteristics of the analytical procedure are verified[1-3] and tested during the application of the analytical method in question. In the present paper, based on the extended divergence measure[1] and on expressing the information gain in terms of the variance reduction[3,4] attained by carrying out the analysis, it is shown which principles must be adhered to in order to achieve the maximum information gain of results of quantitative analysis.

In quantitative chemical analysis, an analytical signal with information content $I(p, p_0)$ is created and converted, via the stoichiometric equivalent or by calibraton, to the result[2], which represents information gain $I(r, p, p_0)$, or $I(p, p, p_0)$ if $p(x) = r(x)$. This gain depends on the information content of the signal $I(p, p_0)$ and on how the information contained in the signal is transformed into analytical information[2,5-7], as well as on how the uncertainty of the analytical results is controlled. The object of this paper is the effect of what is called Good Laboratory Practice[8] and Quality Data Control[9-11] on the information gain of results of quantitative analysis and on its stability during the long-term use of the analytical procedure in question.

THEORETICAL

If it can be assumed that the major component in the uncertainty of results of quantitative analysis is the random component, characterized by variance σ^2, the a poste-

* Part XXV in the series Theory of Information as Applied in Analytical Chemistry; Part XXIV: Collect. Czech. Chem. Commun. *55*, 91 (1990).

riori distribution is regarded as correct[1,2], i.e. we put $r(x) = p(x)$, where $r(x)$ is the probability density of the true distribution and $p(x)$ is the probability density of the a posteriori (experimentally established) distribution. Then the information gain for the a posteriori normal distribution $N(\mu, \sigma^2)$ and for the a priori uniform distribution $U(x_1, x_2)$, $x_1 + 3\sigma \leqq \mu \leqq x_2 - 3\sigma$, is

$$I(p, p, p_0) = \ln\left[(x_2 - x_1)/(\sigma\sqrt{2\pi\,\mathrm{e}})\right]. \tag{1a}$$

Since variance of the a priori uniform distribution is $\sigma_0^2 = (x_2 - x_1)/12$, the information gain is

$$I(p, p, p_0) = -(1/2)\ln R + (1/2)\ln\left[12/(2\pi\,\mathrm{e})\right] =$$
$$= -(1/2)\ln R - 0{\cdot}176, \tag{1b}$$

where the variance reduction[3,4] is $R = (\sigma/\sigma_0)^2$, $R \in (0, 1\rangle$. Information gain according to Eq. (*1a*) is $I(p, p, p_0) \geqq 0{\cdot}373$, and it is positive also in case the condition $x_1 + 3\sigma \leqq \mu \leqq x_2 - 3\sigma$ is not met, i.e. if $(x_2 - x_1) \geqq 6\sigma$ (ref.[5], p. 115). Then Eq. (*1a*) does not hold true and has to be replaced by Eq. (*5.6*) in ref.[5].

If the results may be biassed, which must be admitted in the majority of cases where calibration is performed, the information gain is[1,2]

$$I(r, p, p_0) = \ln\left[(x_2 - x_1)/(\sigma\sqrt{2\pi\,\mathrm{e}^k})\right] - (1/2)(\delta/\sigma)^2, \tag{2a}$$

where $k = (\sigma_r/\sigma)^2$. The mean error can be established as $\delta = (\mu^* - X^*)$, where μ^* is the analyte content found by analysis in a reference material (RM) with a certified content X^* known with a precision characterized by the σ_r^2 value. The true distribution $r(x)$ then can be regarded as normal, $N(X^*, \sigma_r^2)$, and Eq. (*2a*) can be written as

$$I(r, p, p_0) = -(1/2)\ln R + (1/2)\ln\left[12/(2\pi\,\mathrm{e}^k)\right] - (1/2)(\delta/\sigma)^2 \tag{2b}$$

or

$$I(r, p, p_0) \approx -(1/2)\left[\ln R + (\delta/\sigma)^2\right] + 0{\cdot}3 \tag{2c}$$

because according to ref.[1], the relation $\sigma_r \leqq (1/4)\,\sigma$ should invariably hold true, i.e. the exponent $k = (\sigma_r/\sigma)^2$ lies within the limits of $0 \leqq k \leqq 0{\cdot}065$, and $0{\cdot}29 \leqq (1/2)\,.\,\ln\left[12/(2\pi\,\mathrm{e}^k)\right] \leqq 0{\cdot}32$. Thus the information gain $I(r, p, p_0)_0$ for the case where the equality $\delta = 0$ has been verified experimentally is higher than $I(p, p, p_0)$ where we only assume that the results are unbiassed. The difference $I(p, p, p_0) - I(r, p, p_0)_0$ depends on the value of $k \in \langle 0, 1\rangle$ and is highest for $k = 0$, when its value is 0·5 natural units ("nit"), hence it is by no means negligible.

For the case of the a priori normal distribution $N(\mu_0, \sigma_0^2)$, the information gain of accurate results is

$$I(p, p, p_0) = \ln(\sigma_0/\sigma) + (1/2)\{[(\mu - \mu_0)/\sigma_0]^2 + \\ + (\sigma^2 - \sigma_0^2)/\sigma_0^2\} \tag{3a}$$

which can be written as

$$I(p, p, p_0) = -(1/2)\ln R + (1/2)\{[(\mu - \mu_0)/\sigma_0]^2 + (R - 1)\}\,. \tag{3b}$$

The information gain according to Eqs (*3a*, *3b*) for accurate results is always non-negative. The quantity $(\mu - \mu_0)/\sigma_0$, however, is not metrological; it rather characterizes the "measure of surprise" from the actual result.

If the results involve a mean error $\delta \neq 0$, the information gain is

$$I(r, p, p_0) = \ln(\sigma_0/\sigma) + (1/2)\{[(\mu - \mu_0)/\sigma_0]^2 + \\ + k(\sigma^2 - \sigma_0^2)/\sigma_0^2\} - (1/2)(\delta/\sigma)^2\,; \tag{4a}$$

if the variance reduction $R = (\sigma/\sigma_0)^2$, which should always lie within the limits $0 < R \leqq 0.0625$, is used, this gain can be written as

$$I(r, p, p_0) = -(1/2)\ln R + (1/2)\{[(\mu - \mu_0)/\sigma_0]^2 + \\ + k(R - 1)\} - (1/2)(\delta/\sigma)^2\,. \tag{4b}$$

Since for $k \leqq 0.0625$ we have $(1/2)\,k(R - 1) \ll -(1/2)\ln R$, the following holds true:

$$I(r, p, p_0) \approx -(1/2)\ln R + (1/2)[(\mu - \mu_0)/\sigma_0]^2 - (1/2)(\delta/\sigma)^2\,. \tag{4c}$$

It can be demonstrated that for $k = 0$ and for a low R the $I(p, p, p_0) - I(r, p, p_0)_0$ difference equals 0·5 natural units, hence it is the same as in the case of the uniform a priori distribution.

It is clear that the information gain according to Eqs (*2*) and (*4*) depends on the standard deviation σ and on the mean error δ, which, however, are somewhat different for the individual results of particular analyses. Low fluctuations of these metrological characteristics have no substantial effect on the information gain; it is only important that they do not increase systematically during the further long-term use of the analytical method in question. Thus, it is necessary not only to attain a high information gain for a method but also to take steps to ensure its stability during the routine application of the method.

Relations (*2*) and (*4*) can be written in the general form

$$I(r, p, p_0) = I(r, p, p_0)_0 - (\delta/\sigma)^2/2\,. \tag{5}$$

Each value $\delta \neq 0$ must be inserted in Eq. (*5*) irrespective of whether it is statistically significant at the chosen $(1 - \alpha)$ significance level or not. In practice we usually only know that $\delta \leqq \sigma z(\alpha)$ where $z(\alpha)$ is the critical value of the normal distribution for the significance level chosen, the true δ value (which, moreover, can vary quite appreciably) being unknown.

Then we actually only know that

$$I(r, p, p_0)_0 - (1/2)\, z(\alpha) \leqq I(r, p, p_0) \leqq I(r, p, p_0)_0\,. \qquad (6)$$

While the upper limit in this relation is always positive, the lower limit can be zero or even negative. The case where $I(r, p, p_0) \leqq 0$ can be interpreted as a situation where incorrect results misinform us[1,2,6]. If the condition is satisfied that the results involve a mean error statistically significant no more than at the conventional level of $(1 - \alpha) = 0{\cdot}95$, for which $z(\alpha) = 1{\cdot}96$, then for $I(r, p, p_0)_0 \geqq 1{\cdot}92$, which corresponds to a variance reduction $R \leqq 0{\cdot}021$, i.e. $\sigma \leqq 0{\cdot}145\sigma_0$, the information gain is always positive.

DISCUSSION AND CONCLUSIONS

The following conclusions, useful in practice, can be drawn from the above relations:

1) Information gain according to Eqs (*1a*) and (*3b*), where the results are assumed to be accurate, is primarily determined by the variance reduction and is affected by how the obtained result μ agrees with the expected result. For instance, if we do not known in advance anything more than that $\mu \in \langle x_1, x_2 \rangle$, the information gain according to Eq. (*1b*) is lower than $-(1/2) \ln R$ by a small constant which is determined by the different forms of the a priori and a posteriori distributions. If the former distribution is normal, i.e. the same as the latter, this constant is zero[3]; the difference between the obtained (μ) and expected (μ_0) values, however, plays a role according to Eq. (*3b*).

2) Variance reduction for a given a priori variance σ_0^2 is determined by the precision of the results. This depends on the procedure and on care exercised in the analytical work. Therefore, the analytical procedure has to be optimized so that the variance σ^2 be as low as possible, and the latter should be checked periodically to make sure that it does not increase during the long-term use of the method; if necessary, appropriate steps should be taken to prevent this[8-11].

3) Information gain of biased results also depends on the value (statistical significance) of the mean error δ. The dependence of the information gain on $z(\alpha) = (\delta/\sigma)$ is the same for the a priori uniform and for the normal distributions. The gain according to Eqs (*2a*), (*2b*), (*4a*) and (*4b*) also depends on the ratio $k = (\sigma_r/\sigma)^3$, which characterizes the reliability of the reference material employed for testing the method in question. This implies that it is necessary to have available high-quality reference materials that provide metrological backup for an always the same, i.e. stable in time and sufficiently high, information gain from routine analyses. This is imperative particularly where the dependence of the signal on the analyte concentration is heteroskedastic[1].

4) If we have proved experimentally that the results are unbiassed, the information gain is higher than if the absence of a bias is only assumed. The difference between $I(p, p, p_0)$ and $I(r, p, p_0)_0$, e.g. according to Eqs (*1b*) and (*2b*) or according to Eqs (*3b*) and (*4b*), is the higher the lower is $k = (\sigma_r/\sigma)^2$, and it can be as high as 0·5 natural units, regardless of the a priori distribution. This demonstrates the importance of testing the accuracy of results of analytical methods, particularly if they are used routinely.

5) Information gain of low-precise results involving a highly statistically significant mean error can be zero or even negative, i.e. the results misinform us instead of providing us with relevant information. Zero or negative information gain never emerges from accurate, even though low precise, results. It is therefore the principal aim of good laboratory practice[8] and quality data control[9-11] to prevent results from being biassed.

In conclusion, high and stable information gain can only be obtained from results of optimized procedures[7] under conditions of good laboratory practice[8] and using efficient methods of quality data assurance[9-11]. The most marked loss of information gain is brought about by statistically significant mean errors of results of quantitative analysis.

REFERENCES

1. Eckschlager K., Fusek J.: Collect. Czech. Chem. Commun. *53*, 3021 (1988).
2. Eckschlager K.: Collect. Czech. Chem. Commun. *54*, 3031 (1989).
3. Eckschlager K.: Collect. Czech. Chem. Commun. *55*, 91 (1990).
4. Kateman G.: Anal. Chim. Acta *191*, 125 (1986).
5. Eckschlager K., Štěpánek V.: *Information Theory as Applied to Chemical Analysis.* Wiley, New York 1979.
6. Eckschlager K., Štěpánek V., Danzer K.: J. Chemometrics *4*, 195 (1990).
7. Doerffel K., Eckschlager K.: *Optimální postup chemické analýzy*, 2nd ed. SNTL — Nakladatelství technické literatury, Prague 1988.
8. Horwitz W.: Anal. Chem. *50*, 521A (1978).
9. Kateman G., Pijpers F. W.: *Quality Control in Analytical Chemistry.* Wiley, New York 1981
10. Taylor J. K.: Anal. Chem. *53*, 1588 (1981).
11. Taylor J. K.: Anal. Chem. *55*, 600A (1983).

Translated by P. Adámek.