# INFORMATION CONTENT OF THE DETERMINATIONS OF THE AMOUNT OF AN ANALYTE*

Karel ECKSCHLAGER and Jiří FUSEK

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

Received September 15th, 1987
Accepted December 22nd, 1987

*Dedicated to Prof. Z. Holzbecher on the occasion of his 70th birthday.*

It is shown how the information content expressed in terms of the extended divergence measure $I(r; p, p_0)$ depends on the amount of the component to be determined when homoscedastic and heteroscedastic dependences of the analytical signal on the amount of the analyte are distinguished. The importance of the accuracy is pointed out, with which the amount of the analyte is known in a reference material employed for testing a particular analytical method. Rules are set, the maintaining of which is aimed at avoiding the origin of null information content of analytical results.

In the preceding paper[1] the extended divergence measure $I(r; p, p_0)$ was introduced to evaluate the information content of the results of quantitative analyses that can be subject to a systematic error and some properties of this measure were shown. The information content $I(r; p, p_0)$ depends on the accuracy of the results, which is characterized by the value of the standard deviation $\sigma$ and on their bias given by the value of a systematic error $\delta$ and, in the case of fairly great $\delta$, it can assume the null value or a negative value. The dependence of $I(r; p, p_0)$ on the amount of the analyte $x \in \langle x_1, x_2 \rangle$, where $x_1$ and $x_2$ are the lower and the upper boundary, respectively, of the interval, in which the particular analytical method is used, is not explicitly expressed in refs[1,2]. Yet it is known that the values of the quantities $\sigma$ and $\delta$ entering $I(r; p, p_0)$ often depend on the amount $x$ of the compound to be determined. In this paper it is shown that the dependence of the information content on the amount of the analyte can be considerable also then when the values of $\sigma$ and $\delta$ vary only little with a change of $x$. In conclusion some rules are derived emerging from these findings and the maintaining of them aims at excluding the origin of a low or even null information content of the results of quantitative analyses.

* Part XXI in the series: Theory of Information as Applied in Analytical Chemistry; Part XX: Collect. Czech. Chem. Commun. *53*, 1647 (1988).

## THEORETICAL

An extended divergence measure $I(r; p, p_0)$ for evaluating the information content has been introduced earlier [1]. Let us assume an apriori uniform distribution $U(x_1, x_2)$, an aposteriori normal distribution $N(\mu, \sigma^2)$ found statistically, and a "true" also normal distribution $N(X^*, \sigma_r^2)$; $X^*$ is the true amount of the analyte in a standard used for testing a particular analytical method and $\sigma_r$ is the standard deviation featuring the accuracy of our knowledge of the value $X^*$. Then the information content appears as (compare with the Appendix)

$$I(r; p, p_0) = \ln \frac{x_2 - x_1}{\sigma \sqrt{(2\pi e^k)}} - \frac{1}{2}\left(\frac{\delta}{\sigma}\right)^2 . \qquad (1)$$

Here $k = \sigma_r^2/\sigma^2$ and the mean (systematic) error $\delta = |\mu - X^*|$. Until now[1,2] only the case $\sigma^2 = \sigma_r^2$ has been considered, so that $k = 1$. Yet if we understand $\sigma_r$ as a quantity characterizing the accuracy with which we know the "true" content $X^*$ of the analyte in the reference material used for testing the particular analytical method, we have also to admit $\sigma_r < \sigma$ and thus $0 \leqq k \leqq 1$. The number $n$ of parallel determinations of which the value of $\sigma$ has been calculated is known as we choose it ourselves and usually $n \geqq 12$; the number $N$ of determinations employed for finding $X^*$ is either indicated in the certificate of the reference material or we can estimate it as $N \in \langle 15, 20 \rangle$ (ref.[3]).

Unlike the "classic" information theory, which fails to define negative information gain, the inclusion of semantic and pragmatic points of view has enabled us to understand $I(r; p, p_0) = 0$ as a case when incorrect results misinform us (refs[2,4]).We shall emerge from the scheme

$$I = \begin{cases} I(r; p, p_0) & \text{for } I(r; p, p_0) > 0 \\ 0 & \text{for } I(r; p, p_0) \leqq 0 \end{cases}$$

in which the case $I = 0$ will be called "null information content" regardless of $I(r; p, p_0)$ being able to assume even a large negative value. In practice we will always perform such metrological measures so that the information content evaluated in terms of the extended divergence measure $I(r; p, p_0)$ may always be positive and possibly high enough. In Eq. (1) any value $\delta > 0$ should be substituted regardless of its statistical significance. However, in practice, when testing by means of reference material, we can find out that:

*a*) $\delta \leqq \sigma t(m, \alpha)/\sqrt{n}$, where $t(m, \alpha)$ is the critical value of the Student distribution with $m = n - 1$ degrees of freedom, i.e., the systematic error is not statistically significant on the level $\alpha$; then we only know about the true information content

of the results that

$$\ln \frac{x_2 - x_1}{\sigma \sqrt{(2\pi e^k)}} - \tfrac{1}{2} t^2(m, \alpha)/n \leqq I(r; p, p_0) \leqq \ln \frac{x_2 - x_1}{\sigma \sqrt{(2\pi e^k)}} . \qquad (2)$$

*b*) $\delta > \sigma t(m, \alpha)/\sqrt{n}$, i.e., it is statistically significant and the value $\delta = |\mu - X^*|$ will be substituted in Eq. (*1*).

Case *b*) should not occur in practice; most frequently we can really avoid it with adequate calibration but also in case *a*) null information content can arise when

$$\frac{x_2 - x_1}{\sigma} = \sqrt{2\pi} \exp\left[\frac{1}{2}\left(\frac{t^2(m, \alpha)}{n} + k\right)\right] = A(k, n, \alpha) . \qquad (3)$$

When calibrational is properly carried out, statistical insignificance of the systematic error is verified on the level $\alpha = 0{\cdot}05$, and the interval $\langle x_1, x_2 \rangle$ is wide enough, null information content should not be encountered as far as the values of $\sigma$ and $\delta$ do not vary with the content of the analyte $x \in \langle x_1, x_2 \rangle$. However, this is fulfilled only exceptionally; therefore we will screen the information field for $x \in \langle x_1, x_2 \rangle$ under assumptions of various kinds of dependence of $\sigma$ and $\delta$ on $x$. In utilizing Eq. (*1*) we will adopt several values of $k$. Consider the following cases:

1. The ratio of the variances $k = \sigma_r^2/\sigma^2$ will be most frequently within $\langle 0, 1 \rangle$ because $\sigma_r \geqq 0$, $\sigma > 0$ and, as a rule, we can expect $\sigma_r < \sigma$. However, consequences of case $k > 1$ will be pointed out as well.

2.1. The dependence of the signal on the content of the analyte is homoscedastic, i.e., $\sigma = \text{const}$ for $x \in \langle x_1, x_2 \rangle$ (ref.[5]).

2.2. The dependence of the signal on the content of the analyte is heteroscedastic, i.e., $\sigma = \sigma(x)$ for $x \in \langle x_1, x_2 \rangle$, ref.[5]; only linear dependence will be involved.

3.1. The systematic error $\delta$ is constant for $x \in \langle x_1, x_2 \rangle$.

3.2. The systematic error depends on the content of the analyte; only the most simple case $\delta = \delta_0 + bx$, $b \neq 0$, $\delta > 0$ for $x \in \langle x_1, x_2 \rangle$ will be considered.

3.3. The systematic error depends on the content of the analyte and it assumes the null value for some $x \in \langle x_1, x_2 \rangle$; i.e., the difference $(\mu - X^*)$ takes on both positive and negative values. Important are mutual combinations of individual cases: regarding the effect of the ratio $\delta/\sigma$ on the value of $I(r; p, p_0)$, the behaviour of $\delta/\sigma$ will be determining. If $\sigma$ decreases and $\delta$ increases (cases 2.2. and 3.2.), the ratio $\delta/\sigma$ and the statistical significance of the systematic error $\delta$ rapidly increase. Simultaneous increase or decrease of both quantities causes that $\delta/\sigma$ almost does not vary and $I(r; p, p_0)$ is varied only by the change of $\sigma$. Actual changes of $I(r; p, p_0)$ with varying $x \in \langle x_1, x_2 \rangle$ can be detected by numerical investigation under various conditions.

## RESULTS AND DISCUSSION

The information field related to Eq. (*1*) was examined with use of a Hewlett–Packard 9825 computer for all introduced cases 1.–3.3. and for some combinations of them. A few results selected from a voluminous computer output are depicted in Figs 1–3.

*1*) The effect of the inaccuracy, with which we know the true value of the content of the analyte $X^*$ in the employed reference material and which we evaluate by the value $\sigma_r^2$, upon the information content in Eq. (*1*) is given by the ratio $k = \sigma_r^2/\sigma^2$. This effect can be practically neglected for $\sigma_r/\sigma \leqq 0{\cdot}25$, as can be seen from Fig. 1. This assessment casts light on the importance of the preparation and certification of reference materials and of their use for verifying analytical methods, in which the accuracy of our knowledge of $X^*$ has always to exceed that of results provided by the verified analytical method. It should hold: $\sigma_r \leqq 0{\cdot}25\sigma$; then the effect of the inaccuracy of the value $X^*$ can be omitted.

*2*) In seeking for the calibration function, it is not too important in practice, whether the relationship between the signal $y$ and the content of the analyte $x$ is homoscedastic or heteroscedastic because of the least squares method being robust enough against the change of $\sigma$ with varying $x$; only the shapes of confidence levels along the calibration curve differ. However, the dependence of the information content on the amount of the analyte $x \in \langle x_1, x_2 \rangle$ is given by the possibility of both $\sigma$ and $\delta$ to vary with $x$ or of the ratio $\delta/\sigma$ to vary with $\sigma$ when the mean error $\delta > 0$ is constant. The changes of $I(r; p, p_0)$ with varying $\delta$ in dependence on $x$ are shown in Figs 2*a*, 2*b* for various cases.

*3*) In the case of a heteroscedastic dependence of the signal on $x$, the content of the analyte $x \in \langle x_1, x_2 \rangle$ is relevant for testing the statistical significance of the error $\delta$;

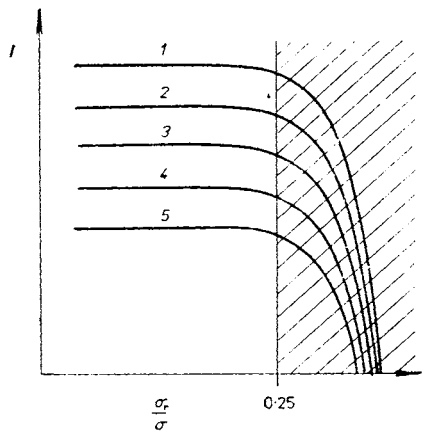


FIG. 1

The dependence of $I(r; p, p_0)$ on the ratio $\sigma_r/\sigma$ for $x_2 - x_1 = 50$, $\delta = 0$. Curve: *1* $\sigma = 0{\cdot}003$, *2* $\sigma = 0{\cdot}01$, *3* $\sigma = 0{\cdot}03$, *4* $\sigma = 0{\cdot}1$, *5* $\sigma = 0{\cdot}3$. The domain for $\sigma_r/\sigma \geqq 0{\cdot}25$ (infavourable effect of $k$ upon $I(r; p, p_0)$) is hatched

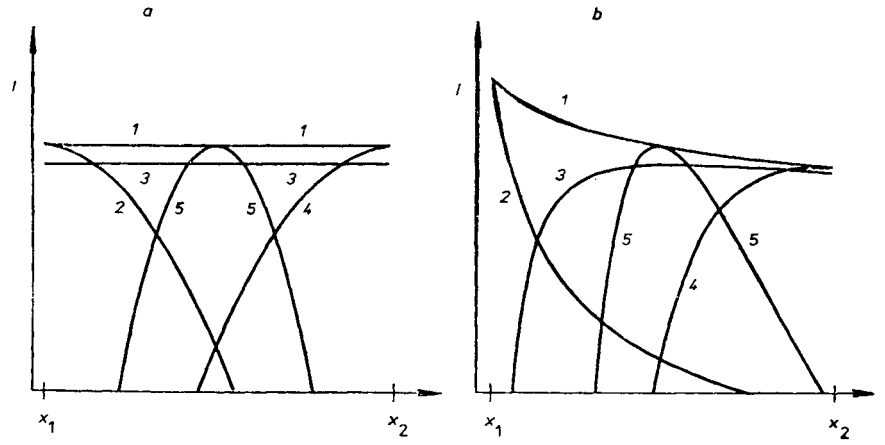


FIG. 2

The dependence of $I(r; p, p_0)$ on $\delta$ for $k = 0{\cdot}25$, $x \in \langle 2, 52 \rangle$: *a* homoscedastic dependence $\sigma = 0{\cdot}015$; *b* heteroscedastic dependence $\sigma = 0{\cdot}001 + 0{\cdot}005x$. Curve: *1* $\delta = 0$, *2* $\delta = -0{\cdot}003 + 0{\cdot}002x$, *3* $\delta = 0{\cdot}015$, *4* $\delta = 0{\cdot}105 - 0{\cdot}002x$, *5* $\delta = -0{\cdot}108 + 0{\cdot}004x$, i.e., the difference $(\mu - X^*)$ assumes both positive and negative values and for $x = 27$ it is zero

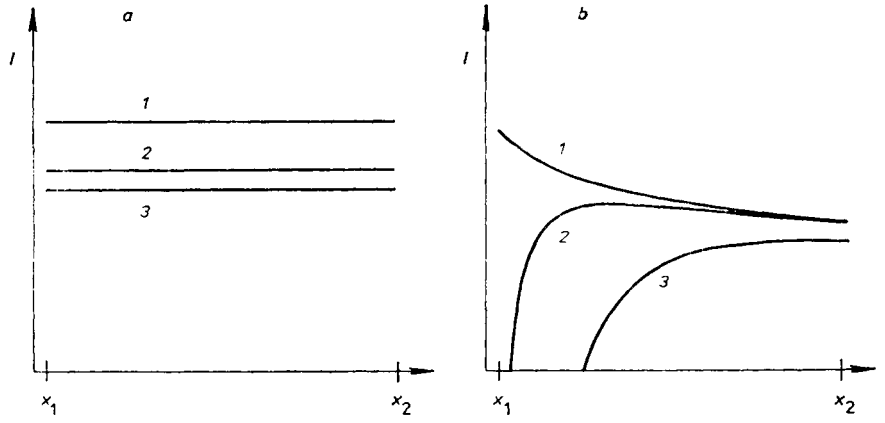


FIG. 3

The dependence of $I(r; p, p_0)$ on the statistical significance of $\delta$: *a* homoscedastic dependence $\sigma = 0{\cdot}015$, curve: *1* $\delta = 0$, *2* $1 - \alpha = 0{\cdot}90$, $\delta = 0{\cdot}0247$, *3* $1 - \alpha = 0{\cdot}95$, $\delta = 0{\cdot}0294$; $I(r; p, p_0)$ is independent of $x$; *b* heteroscedastic dependence $\sigma = 0{\cdot}001 + 0{\cdot}005x$, curve: *1* $\delta = 0$, *2* $1 - \alpha = 0{\cdot}95$ for $x = 10$, i.e., $\delta = 0{\cdot}09996$, *3* $1 - \alpha = 0{\cdot}90$ for $x = 32$, i.e., $\delta = 0{\cdot}26485$

namely, here it is not sufficient to find out that $\delta \leqq \sigma(x)\, t(m, \alpha)/\sqrt{n}$ but it is necessary to take into account the content $x$ corresponding to $\sigma(x)$ and to realize that the same value appearing as unimportant for some $x$ can be for another lower value of $\sigma(x)$ statistically significant. It is apparent from Fig. 3*a* that in the case of homoscedastic dependence only the level $\alpha$ is effective; yet Fig. 3*b* depicts the change of $I(r; p, p_0)$ with $\sigma(x)$ when $\delta = \text{const}$. It follows from it that, when verifying an analytical method with heteroscedastic dependence of $y$ on $x$, we do not manage with one reference material. Since the dependence $\sigma = \sigma(x)$ is quite frequent, the production and certification of sets of reference materials are to be preferred to inconvenient practice of one reference material, for any material to be analyzed.

As far as the prevention of the rise of null information content is concerned, it is always expedient, after discovering the heteroscedasticity, to evaluate or graphically illustrate (Fig. 3*b*) the dependence of

$$\ln \frac{x_2 - x_1}{\sigma \sqrt{(2\pi e^k)}} - \frac{1}{2} \frac{t^2(m, \alpha)}{n}$$

on $x \in \langle x_1, x_2 \rangle$ for $\alpha$ chosen in advance and to check if it is always nonzero and greater than the information content needed for the solution of a given analytical task. If, in addition, we find out that $\delta$ also varies with $x \in \langle x_1, x_2 \rangle$, we have to carry out the calibration in such a way that it never exceed the value $\sigma(x)\, t(m, \alpha)/\sqrt{n}$. This is usually not easy and we cannot do it without a set of reference materials covering almost uniformly the entire domain of the contents $x$. Then also the value of $k = \sigma_r^2/\sigma^2$ plays an important role; as we have already mentioned, it is desirable for the ratio $\sigma_r/\sigma$ not to exceed 0·25. All these conditions will not be always maintained in practice, particularly in determining low contents of analytes; then we have to be at least aware of our using results with low information contents in decision making.

## CONCLUSION

When generalizing the results of the performed investigation of the information field of an analytical method in the domain $\langle x_1, x_2 \rangle$ of the analyte contents, it is obvious how important an assessment is if the value of $\sigma$ depends on the analyte content $x$, if a systematic error is present, or if this error is constant or depends on $x \in \langle x_1, x_2 \rangle$. For calibration, such reference materials should be employed whose certificates are so equipped that it is possible to determine from them the accuracy of the contents of individual analytes and the number $N$ of parallel determinations from which the "true" contents $X^*$ were derived. If the dependence of the signal $y$ on the analyte content $x$ is heteroscedastic, it is, in addition, purposeful to use, in order to verify a method, such a set of reference material in which the analyte content almost uniformly covers the whole interval $\langle x_1, x_2 \rangle$. The effect of the number of calibration

samples has been studied earlier (ref.[2]), 8–10 samples are sufficient and for more than 10 samples this effect can be completely omitted. Rather important is to reliably know the value $X^*$, i.e., $\sigma_r$ is to be small.

The situation when $I(r; p, p_0) \leqq 0$ is to be considered for most infavourable and, therefore, in practice we state or estimate at least approximately the interval of null information. It can lie anywhere within $\langle x_1, x_2 \rangle$ if the error depends on the content of the analyte and the relationship between $y$ and $x$ is heteroscedastic; for $\delta = \text{const}$ it is sufficient to determine the upper boundary $x_u$ of the interval. It can be found from Eq. (2) if we set $x_1 = x_d$ (the detection limit) and $x_2 = x_u$. From the relation (4)

$$(x_u - x_d)/\sigma = A(k, n, \alpha) \tag{4}$$

it follows

$$x_u = x_d + A(k, n, \alpha)\,\sigma\,; \tag{5}$$

moreover, estimating $x_d$ with $3\sigma$,

$$x_u = 3\sigma + A(k, n, \alpha)\,\sigma\,. \tag{6}$$

## APPENDIX

Formula (1), employed above for the investigation of the information field for analyte contents $x \in \langle x_1, x_2 \rangle$, has been derived from extended information measure[1]

$$I(r; p, p_0) = \int_{-\infty}^{\infty} r(x) \ln \frac{p(x)}{p_0(x)}\,\mathrm{d}x \tag{A1}$$

in the following way.

The apriori probability distribution is uniform:

$$p_0(x) = \begin{cases} 1/(x_2 - x_1) & \text{for} \quad x \in \langle x_1, x_2 \rangle \\ 0 & \text{otherwise}\,. \end{cases}$$

The aposteriori distribution is the normal distribution found by the analysis:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right], \tag{A2}$$

where $\sigma$ is determined from an independent verification series of $n$ determinations.

The "true" distribution escapes from the analysis of $N$ determinations is also normal with parameters $(X^*, \sigma_r^2)$:

$$r(x) = \frac{1}{\sigma_r\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - X^*}{\sigma_r}\right)^2\right]. \tag{A3}$$

The information measure appears as

$$I = \int_{-\infty}^{\infty} r(x) \left\{ \ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 + \ln(x_2 - x_1) \right\} \mathrm{d}x =$$

$$= \ln \frac{x_2 - x_1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} (x^2 - 2\mu x + \mu^2)\, r(x)\, \mathrm{d}x =$$

$$= \ln \frac{x_2 - x_1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2} (\sigma_r^2 + X^{*2} - 2\mu X^* + \mu^2) =$$

$$= \ln \frac{x_2 - x_1}{\sigma\sqrt{2\pi}} - \frac{1}{2}\left[\frac{\sigma_r^2}{\sigma^2} + \frac{(X^* - \mu)^2}{\sigma^2}\right]. \qquad (A4)$$

After setting $\delta = |X^* - \mu|$, $k = \sigma_r^2/\sigma^2$, we obtain

$$I = \ln \frac{x_2 - x_1}{\sigma\sqrt{2\pi}} - \frac{1}{2}\left[k + \left(\frac{\delta}{\sigma}\right)^2\right] = \ln \frac{x_2 - x_1}{\sigma\sqrt{(2\pi \mathrm{e}^k)}} - \frac{1}{2}\left(\frac{\delta}{\sigma}\right)^2 \qquad (A5)$$

which is Eq. (*1*).

REFERENCES

1. Eckschlager K., Štěpánek V.: Collect. Czech. Chem. Commun. *50*, 1359 (1985).
2. Danzer K., Eckschlager K., Wienke D.: Fresenius Z. Anal. Chem. *327*, 312 (1987).
3. Musil J.: Chem. Listy *80*, 1233 (1986).
4. Eckschlager K., Štěpánek V.: *Analytical Measurement and Information*, p. 46. Research Studies Press, Letchworth 1985.
5. Schwartz L. M.: Anal. Chem. *51*, 723 (1979).

Translated by V. Štěpánek.