HOW TO MEASURE INFORMATION OBTAINED FROM BIASED QUANTITATIVE ANALYSES

Karel ECKSCHLAGER[a] and Vladimír ŠTĚPÁNEK[b]

[a] *Institute of Inorganic Chemistry,*
Czechoslovak Academy of Sciences, 160 00 Prague 6, and
[b] *Nuclear Research Institute, 250 68 Řež*

Received April, 28th, 1984

It is shown that the use of the divergence measure for expressing the information content of the results of a quantitative analysis is justified only then if we need not to expect the rise of a systematic, *i.e.*, statistically significant mean error. However, unless we can exclude the rise of a systematic error we have to adopt the information measure $I(r, p, p_0)$ proposed earlier; then of course every even statistically insignificant mean error has effect upon the information content. It is studied the relationship of this measure to another information measure used in a specific case for expressing the information content of biased results of a quantitative analysis.

The possibility of the rise of a systematic error represents the most reliable risk of distorting the results of a quantitative analysis. Therefore we have dealt already in one of preceding papers of this series[1] with the problem of the information content of the results of a quantitative determination subject to a systematic error and we have expressed it as the difference of the divergencies of the apriori and the aposteriori distributions and of the true and the aposteriori distributions respectively. Later on[2] we were concerned with the aposteriori uncertainty of results subject to a systematic error and we called its negative value "the measure of accuracy". Recently[3] we introduced an information measure based on the difference of Kerridge–Bongard measures for the true and the apriori distributions and for the true and the aposteriori distributions respectively and we showed its suitability to expresss the information content of the results subject to a systematic error. Because the concept of the systematic error itself has not been understood so far unambiguously and the measures from papers[1,3] do not yield equal results for different apriori distributions and for a normal aposteriori one, we are returning again in this paper to the problem of the information content of the results that can be subject to a systematic error and we will show the relationship between both measures of the information content introduced in papers[1,3] and we will judge their applicability in specific cases.

THEORETICAL

We will take for unbiased those results, for the mean of which it holds that the $100(1 - \alpha)\%$ confidence interval

$$\left\langle \bar{x} - \sigma \frac{z(\alpha)}{\sqrt{n_A}}, \quad \bar{x} + \sigma \frac{z(\alpha)}{\sqrt{n_A}} \right\rangle \tag{1}$$

covers the true value X, where n_A is the number of parallel determinations from which the mean is calculated, σ is the standard deviation (a parameter) and $z(\alpha)$ is the critical value of the normal distribution at the significance level α. Thus condition (1) is fulfilled also with the results subject to a mean error $\delta = |X - \bar{x}| \leqq \sigma[z(\alpha)/\sqrt{n_A}]$. We will take $\delta > \sigma[z(\alpha)/\sqrt{n_A}]$ for a systematic error, *i.e.*, a mean error δ significant at the level α.

The information content of the results, of which we can assume to be true, *i.e.*, that their contingent mean error will never be statistically significant at the significance level α, can be expressed by the use of the divergence measure[3-6] as

$$I(p, p_0) = H(p, p_0) - H(p) = \int_{x_1}^{x_2} p(x) \ln \frac{p(x)}{p_0(x)} \, dx \,, \tag{2}$$

where $p_0(x)$ is a continuous apriori distribution and $p(x)$ is a continuous aposteriori one, $H(p)$ is Shannon's entropy for the aposteriori distribution according to (4) of paper[3] and $H(p, p_0)$ is a Kerridge–Bongard measure according to (6) of the same paper. Here also the conditions are introduced, under which (2) is valid. For $p_0(x)$ being uniform $U(x_1, x_2)$ and $p(x)$ normal $N(X, \sigma^2)$, where X is the true value, the information content[4,5] will be for $x_1 + 3\sigma \leqq X \leqq x_2 - 3\sigma$

$$I(p, p_0) = \ln \frac{x_2 - x_1}{\sigma \sqrt{(2\pi e)}} \tag{3a}$$

and for $p_0(x)$ being normal $N(\mu_0, \sigma_0^2)$ and $p(x)$ also normal $N(X, \sigma^2)$, $\delta_0 = |X - \mu_0|$, the information content according to[4-6] yields

$$I(p, p_0) = \ln \frac{\sigma_0}{\sigma} + \frac{1}{2}\left[\left(\frac{\delta_0}{\sigma_0}\right)^2 + \frac{\sigma^2 - \sigma_0^2}{\sigma_0^2}\right]. \tag{3b}$$

In both last cases, *i.e.*, in (3a) and in (3b), the expected value of the aposteriori distribution is equal to the true value X, *i.e.*, the results are unbiased in the sense of condition (1).

If the results can be subject to a systematic error we can calculate their information content according to[1] as

$$I(p, p_0 \parallel r, p) = I(p, p_0) - I(r, p) = \\ = \int_{x_1}^{x_2} p(x) \ln \frac{p(x)}{p_0(x)} \, dx - \int_{x_1}^{x_2} r(x) \ln \frac{r(x)}{p(x)} \, dx \,, \qquad (4)$$

if we start from the maximum uncertainty prior to the analysis, *i.e.* if the apriori distribution $p_0(x)$ is uniform. For $p_0(x)$ being uniform $U(x_1, x_2)$, $p(x)$ normal $N(\mu, \sigma^2)$ and $r(x)$ also normal $N(X, \sigma^2)$, $\delta = |X - \mu|$, and for $x_1 + 3\sigma \leqq X \leqq x_2 - 3\sigma$ we obtain

$$I(p, p_0 \parallel r, p) = \ln \frac{x_2 - x_1}{\sigma \sqrt{(2\pi e)}} - \frac{1}{2} \left(\frac{\delta}{\sigma}\right)^2 . \qquad (5)$$

If we substituted in (*4*) $N(\mu_0, \sigma_0^2)$ for $p_0(x)$, $N(\mu, \sigma^2)$ for $p(x)$, and $N(X, \sigma^2)$ for $r(x)$, $\delta = |X - \mu|$, we would obtain

$$I(p, p_0 \parallel r, p) = \ln \frac{\sigma_0}{\sigma} + \frac{1}{2} \left[\left(\frac{\mu - \mu_0}{\sigma_0} \right)^2 - \left(\frac{\delta}{\sigma} \right)^2 + \frac{\sigma^2 - \sigma_0^2}{\sigma_0^2} \right] . \qquad (6a)$$

The rightfulness of this procedure will be discussed below.

Recently[3] we have introduced an information measure of results that can be subject to a systematic error as

$$I(r; p, p_0) = H(r, p_0) - H(r, p) = \int_{x_1}^{x_2} r(x) \ln \frac{p(x)}{p_0(x)} \, dx \, . \qquad (7)$$

Similarly as in the measure given in (*4*) the distribution of the same kind as the aposteriori one will be taken for true but with the expectation X. It is obvious that for $X = \mu$, *i.e.* for $r(x) \equiv p(x)$, $I(r; p, p_0)$ by (7) changes into the divergence measure $I(p, p_0)$ by (2). As it has been shown[3] the value of $I(r; p, p_0)$ for $p_0(x)$ uniform $U(x_1, x_2)$, $p(x)$ normal $N(\mu, \sigma^2)$ and $r(x)$ normal $N(X, \sigma^2)$, the mean error $\delta = |X - \mu|$, is given by formula (5). For $p_0(x)$ normal $N(\mu_0, \sigma_0^2)$, $p(x)$ normal $N(\mu, \sigma^2)$, and $r(x)$ normal $N(X, \sigma^2)$, the mean error $\delta = |X - \mu|$ and the bias of the apriori estimate $\delta_0 = |X - \mu_0|$, the information content by (7) turns out to be

$$I(r; p, p_0) = \ln \frac{\sigma_0}{\sigma} + \frac{1}{2} \left[\left(\frac{\delta_0}{\sigma_0} \right)^2 - \left(\frac{\delta}{\sigma} \right)^2 + \frac{\sigma^2 - \sigma_0^2}{\sigma_0^2} \right] , \qquad (6b)$$

which does not coincide with (*6a*) when $X \neq \mu$, because then $\delta_0 \neq |X - \mu_0|$. For unbiased results, *i.e.* for $\delta = 0$, it becomes the formula in (*3b*).

Now let us compare both measures introduced in (*4*) and (*7*). We have shown earlier[3,5] that $I(r; p, p_0)$ can be evaluated in terms of measures of inaccuracy as

$$I(r; p, p_0) = H(r \mid p_0) - H(r \mid p) \tag{8}$$

and since we know in addition that the relationship

$$H(q \mid p) = H(q) + I(q, p) \tag{9}$$

holds between the measure of inaccuracy $H(q \mid p)$, the entropy $H(q)$, and the information gain $I(q, p)$, we can write

$$I(r; p, p_0) = H(r) + I(r, p_0) - H(r) - I(r, p) = I(r, p_0) - I(r, p)\,. \tag{10}$$

This is, of course, a formula diverse from $I(p, p_0 \parallel r, p)$ given in (*4*) (it differs in the first term) and it shows that the measure in (*7*) is in fact information obtained in the transition from the apriori assumption to the real result diminished by" misinformation" caused by the statistical distribution of the results which is "better" than the apriori inaccuracy but biased from the true value of the content of the sample.

In the specific case of an apriori uniform distribution, *i.e.* for the maximum uncertainty before the experiment, it is true (which is valid only in this case, see[3,5]) that the information gain equals the decrease of uncertainty, *i.e.* the difference of entropies, and thus

$$I(r, p_0) = H(r) - H(p_0) \tag{11}$$

$$I(p, p_0) = H(p) - H(p_0)\,, \tag{12}$$

where p_0 is a uniform distribution. However, since p and r in our consideration are normal distributions with common variance and since it is well known that the entropy of a normal distribution does not depend on its mean value[5], it holds: $H(r) = H(p) = \ln \sigma \sqrt{(2\pi e)}$ so that both $I(r, p_0)$ and $I(p, p_0)$ are the same and, therefore, both measures considered in (*4*) and (*7*) provide identical results in this case.

Otherwise it is obvious that the use of the difference given in (*4*) outside the case of an apriori uniform distribution in such a way as it was presented in paper[1] would not be rightful.

RESULTS AND DISCUSSION

Formula (*5*), which takes the most frequent place in analytical practice, can be adjusted by substituting $w = (x_2 - x_1)/\sigma$ and $\delta = k\sigma$ so that

$$I(p, p_0 \parallel r, p) = I(r; p, p_0) = \ln \frac{w}{\sqrt{(2\pi e)}} - \frac{1}{2} k^2 \,. \tag{13}$$

Then it is possible to find, for a given ratio w, such a value $k = \sqrt{[\ln (w^2/2\pi e)]}$ for which the quantity in (*13*) takes on the zero value, or to find, for the given difference $(x_2 - x_1)$ and σ, such a mean error $\delta = k\sigma$ which completely depreciates the analytical results. If we set $k = z(\alpha)/\sqrt{n_A}$ we can look up the significance level α, at which this mean error is statistically significant, in the tables. Several values are shown in Table I. Obviously, for a small value of w a relatively small value of $k = \delta/\sigma$ is sufficient, *i.e.*, a δ statistically significant at a high level α, to completely depreciate the results. For a large value of w, *i.e.* for a great apriori and a small aposteriori uncertainty, the results can be deteriorated only with a large mean error. For values $w \gg 6$, $I(r; p, p_0) = I(p, p_0 \parallel r, p)$ is independent of the position of μ in the interval $\langle x_1 + 3\sigma, x_2 - 3\sigma \rangle$.

Formula (*6b*) for all three distributions normal can be analogously simplified by substituting $\sigma_0 = q\sigma (q \geqq 1)$, $\delta_0 = k_0\sigma_0$ and $\delta = k\sigma$; then

$$I(r; p, p_0) = \ln q + \tfrac{1}{2}\left(k_0^2 - k^2 + \frac{1 - q^2}{q^2}\right) \tag{14}$$

TABLE I

Values of k for different w and n_A resulting in null $I(r; p, p_0)$ given in (*13*) and corresponding confidence levels for mean errors $k \,.\, \sigma$ at a fixed σ

w	n_A	$k\sqrt{n_A}$	$(1 - \alpha)$
6·0	1	0·86	0·610
	2	1·22	0·778
	3	1·49	0·864
8·0	1	1·15	0·750
	2	1·63	0·897
	3	1·99	0·953
10·0	1	1·33	0·816
	2	1·88	0·940
	3	2·30	0·979
25·0	1	1·90	0·943
	2	2·69	0·993
	3	3·29	0·999

takes on its zero value for $\ln q^2 + (1 - q^2)/q^2 = k^2 - k_0^2$. Several corresponding values are shown in Table II. Similarly we could simplify $I(p, p_0 \parallel r, p)$ from (*6a*) by substituting $|\mu - \mu_0| = k_\mu \sigma_0$, $\sigma_0 = q\sigma$, $\delta = k\sigma$ and by tabulating values $k = z(\alpha)/\sqrt{n_A}$ in dependence on q and k_μ for zero $I(p, p_0 \parallel r, p)$. We would obtain a table similar to Table II, where values of k_μ would be shown instead of those of k_0. The meaning of k_μ is of course considerably enough distinct from k_0.

TABLE II

Values of k for different q and k_0 resulting in null $I(r; p, p_0)$ given in (*14*) and corresponding confidence levels for mean errors $k \cdot \sigma$ at a fixed σ

q	k_0	$k\sqrt{n_A}$	$(1 - \alpha)$
1·00	0·50	0·50	0·383
		0·71	0·522
		0·87	0·616
	1·00	1·00	0·683
		1·41	0·842
		1·73	0·916
	1·50	1·50	0·866
		2·12	0·967
		2·60	0·991
2·00	0·50	0·94	0·653
		1·33	0·816
		1·63	0·897
	1·00	1·28	0·800
		1·81	0·930
		2·22	0·974
	1·50	1·70	0·911
		2·40	0·984
		2·94	0·997
3·00	0·50	1·25	0·789
		1·77	0·924
		2·17	0·970
	1·00	1·52	0·871
		2·15	0·968
		2·63	0·991
	1·50	1·89	0·941
		2·67	0·992
		3·27	0·999

Next we will notice the behaviour of the information measure $I(r;\, p, p_0)$ by (*6b*) for various cases of the relationship between X, μ_0 and μ, in which we understand the mean result of analysis μ as experimental verification of the assumption μ_0. In the same time we will show what values the measure by (*6a*) would provide in given cases.

1) $X = \mu = \mu_0$: An unbiased result confirms a true estimate and the information gain $I(r;\, p, p_0) = \ln q + \frac{1}{2}(1 - q^2)/q^2 = \ln q + \frac{1}{2}Q$, where we put $Q = (1 - q^2)/q^2$, depends only on the ratio $q = \sigma_0/\sigma$, *i.e.* on the degree of making the analysis more precise. The measure in (*6a*) would yield the same result in this case.

2) $X = \mu$; $\mu \neq \mu_0$: An unbiased result states the inaccuracy of the apriori assumption. The information gain

$$I(r;\, p, p_0) = \ln q + \tfrac{1}{2}(k_0^2 + Q) \tag{15}$$

depends, besides on q, also on k_0 and it is thus dependent on giving more precision to the analysis and on correcting the original biased estimate. The measure in (*6a*) would provide the same value again.

3) $X = \mu_0$; $\mu \neq X$: Now the result of analysis is biased and causes the denial of a true estimate (analogy of the Type I error in statistical hypothesis testing). The information "gain"

$$I(r;\, p, p_0) = \ln q + \tfrac{1}{2}(-k^2 + Q) \tag{16a}$$

is negative and passes into positive values only in giving higher precision to the analysis. The information measure computed according to (*6a*)

$$I(p, p_0 \,\|\, r, p) = \ln q + \tfrac{1}{2}(k_\mu^2 - k^2 + Q) \tag{16b}$$

would not be negative.

4) $\mu_0 = \mu$; $X \neq \mu$: A biased result conduces to the acceptance of an untrue assumption (analogy of the Type II error in statistical hypothesis testing). Thus for $q = 1$ there exists no information gain which is fulfilled by $I(r;\, p, p_0)$, yet the measure in (*6a*)

$$I(p, p_0 \,\|\, r, p) = \ln q + \tfrac{1}{2}(-k^2 + Q) \tag{17}$$

would be negative.

CONCLUSION

The conducted discussion has shown the behaviour of the information measure $I(r;\, p, p_0)$ when all the three probability distributions are normal and when particular relations between them hold. In the same time it has revealed the discrepancies with reality in which the use of the measure from (*6a*) would result, whose incorrectness had been shown above (except for the transition from a uniform distribution). It is apparent that both formulae in (*6a*) and (*6b*) yield the same value provided the method is unbiased, *i.e.*, if $\mu = X$. If the adopted method is biased, the measure $I(r; p, p_0)$ decreases with increasing k, eventually to negative values, in agreement with reality, whereas the difference in (*6a*) could even grow large with deteriorating inaccuracy of the method (*e.g.*, in the situation when $\mu_0 < X < \mu$).

We can understand the measure $I(r; p, p_0)$ as a generalization of the divergence measure $I(p, p_0)$ (directed divergence[7]) into which it passes for $r(x) \equiv p(x)$. The measure $I(p, p_0 \parallel r, p)$ was used in[1] rather intuitively and only for the case of an apriori uniform distribution. For this distribution $I(r; p, p_0) = I(p, p_0 \parallel r, p)$ and it is independent, for $w \geqslant 6$, of the position of μ in the interval $\langle x_1 + 3\sigma, x_2 - 3\sigma \rangle$. Therefore the conclusions of paper[1] are valid also for the information measure $I(r; p, p_0)$.

If we use the divergence measure $I(p, p_0)$ we assume that the results are not subject to a systematic error, *i.e.*, to a statistically significant mean error δ but in adopting the measure $I(r; p, p_0)$ every non-zero mean error, also a statistically insignificant one, lowers the value of the information content, eventually till to its zero value or even to a negative value. Therefore we have to comprehend the use of the measures $I(p, p_0)$ or $I(r; p, p_0)$ as follows: If it is guaranteed that the results do not bear a systematic error in themselves we use the divergence measure $I(p, p_0)$; however, unless we can exclude the rise of such an error we adopt $I(r; p, p_0)$, yet we are aware that every non-zero mean error will lower the information content of the results even if they fulfil condition (*1*). This can be understood in such a way that already the possibility of the rise of a systematic error makes the aposteriori uncertainty of the results worse and thus lowers their information content. This stresses the importance of correct calibration in the use of instrumental analytical methods (compare[4,5,8]). The possibility to evaluate measurement results that can be subject to a systematic error from the information theoretic point of view represents a new contribution to the theory of "measurement information"[9].

REFERENCES

1. Eckschlager K.: This Journal *44*, 2373 (1979).
2. Eckschlager K., Štěpánek V.: This Journal *45*, 2146 (1980).
3. Eckschlager K.: This Journal *47*, 1580 (1982).
4. Eckschlager K., Štěpánek V.: *Information Theory as Applied to Chemical Analysis*. J. Wiley, New York 1979.
5. Eckschlager K., Štěpánek V.: *Analytical Measurement and Information*. Research Studies Press, Letchworth 1985.

6. Eckschlager K., Vajda I.: This Journal *39*, 3076 (1974).
7. Nath P.: J. Math. Sci. *3*, 1 (1968).
8. Eckschlager K., Štěpánek V.: Anal. Chem. *54*, 1115A (1982).
9. Vajda I., Eckschlager K.: Kybernetika *16*, 120 (1980).

Translated by the author (V. Š.)