LETTER TO THE EDITOR

ROUNDING OFF IN GEOCHEMISTRY*

SIR:

It is commendable that Dr. Shaw has attempted to provide rounding rules for the maximum number of figures to be retained in statistics such as the mean and the variance of a number of data (Shaw, 1962). However, it was found that Shaw's rules are open to some objections. New rules with a firmer basis in statistical theory are proposed in this letter.

Shaw implicitly assumes that all digits of the original data are truly significant. This is an ideal case which would imply that the error of the data is smaller than half of the unit of the last digit. In practice, the error is larger, since values are generally reported in such a way that the last digit is not truly significant. However, the assumption of exclusively significant digits in the original data is a useful working hypothesis under the circumstances. In this way, a theory can be developed for the calculation of the number of digits that may be retained in statistics based upon the original data. It will be seen that this theory leads to results which are considerably different from Shaw's results although the starting point is the same in both cases.

In the following, the concept of the *decimal error* which is the logical consequence of the working hypothesis will be introduced and discussed. Rules for the propagation of the decimal error to the calculated mean, variance, and standard deviation of the data will be developed. After the presentation of these new rules, Shaw's rules will be discussed and both sets of rules will be applied to some examples to demonstrate that there are significant differences between our rules and those of Shaw.

THE DECIMAL ERROR

Errors e are generally randomly distributed and described by the "error function" (International Dictionary of Physics and Electronics, p. 313) which means that the measurements describe a Gaussian curve about the true value. The amount of dispersion is indicated by the standard error $s(e)$. It will be seen that it is useful to define a maximum error Δe which is larger than almost all individual e_i values. Δe must be expressed in probability sense. In such cases, it is conventional to make use of the 95 per cent confidence interval. Thus, the maximum error is

*Published by permission of the Director, Geological Survey of Canada.

defined as being larger than 95 per cent of the individual errors. The relationship between Δe and $s(e)$ is:

$$\Delta e \cong 2.s(e) \tag{1}$$

In practice, numerical results are expressed in the decimal system, and the standard error is generally not reported. The number of digits is then the measure of the accuracy of the data. However, this way of expressing the accuracy is not very precise. The following working hypothesis has been framed as a foundation for developing rules for the number of significant digits of statistics. It is assumed that all digits of the original data are significant; half the unit of the last significant digit is called the maximum decimal error $\Delta e'$, which is equivalent to Eisenhart's $\frac{1}{2}w$ (Eisenhart, 1947) and related to Δe by a constant a so that:

$$\Delta e = a.\Delta e' \tag{2}$$

There are authors who use $a = 3$ or $a = 4$ (Eisenhart, p. 192) but it is often unknown which value of a has been used or even whether a could have been determined with certainty. In our case, we are not concerned with the value of a in the first place, but with the decimal error e' which is the known reduction of the often unknown error $e = a.e'$ and which is also normally distributed. However, formula (2) and the other formulas of this letter are also applicable when the random distribution of e departs from the normal model.

The decimal error is the basis for the following rules which have been developed in agreement with statistical theory. These rules also apply to the propagation of Δe and $s(e)$ instead of $\Delta e'$.

Rules for the Propagation of the Decimal Error to Statistics

Mean. When there are n original data X_i, the mean is:

$$\bar{X} = 1/n.(X_1 + X_2 + \ldots\ldots + X_n) \tag{3}$$

and the propagated maximum decimal error $\Delta(\bar{X})$ follows from the formula

$$\Delta^2(\bar{X}) = \left(\frac{\partial \bar{X}}{\partial X_1}\right)^2 \cdot \Delta^2 X_1 + \left(\frac{\partial \bar{X}}{\partial X_2}\right)^2 \cdot \Delta^2 X_2 + \ldots. \tag{4}$$

which has been derived from the formula for $S^2(R)$ with $R = f(X_1, X_2, \ldots, X_n)$, e.g. presented in table 26–2 of Laitinen (1960).

For each of the original data, the maximum decimal error is $\Delta e'$. Therefore:

$$\Delta(\bar{X}) = \pm \frac{\Delta e'}{\sqrt{n}} \tag{5}$$

(see also our formula 10).

Variance. The propagation of variance is according to the formula:

$$S^2(X_i) = s^2(x_i) + s^2(e_i') + 2.r.s(x_i).s(e_i') \qquad (6)$$

(Deming, 1943, p. 40; Shaw, 1962, p. 242).
$S^2(X_i)$ is the calculated sample variance; $s^2(x_i)$ is the true sample variance; $s^2(e_i')$ is the variance of the decimal errors of the sample; r is the correlation coefficient between x_i and e_i'.

The maximum decimal error $\pm\Delta e'$ represents the 95 per cent confidence limits of the normal distributions $X_i = x_i + e'$. The term $s^2(e_i') \cong (\frac{1}{2}.\Delta e')^2$ of equation (2) is a constant which may generally be neglected because $(\frac{1}{2}.\Delta e')^2 \ll s^2(x_i)$. The error of $S^2(X_i)$ is thus presented by $2.r.s(x_i).(\frac{1}{2}.\Delta e)$. Shaw has pointed out that this is a random variable. However, he does not consider the consequences of this fact, and in developing his rules abandons the right procedure in favour of an arbitrary one which lacks a firm foundation in the theory of errors.

In general, there is no relationship between x_i and its decimal error e_i'; i.e., the mathematical expectation of r is zero. Now let Δr indicate the 95 per cent confidence limits of r which is symmetrically distributed about zero. The probability that r is within $\pm\Delta r$ is then 95 per cent. The value Δr is a function of n. This relationship is presented in Table 1, which is based upon a chart by Pearson & Hartley (1958) (basic computations by David, 1938).

TABLE 1

n	Δr	n	Δr
3	0.99	15	0.52
4	0.95	20	0.45
5	0.88	25	0.40
6	0.81	50	0.28
7	0.75	100	0.20
8	0.71	200	0.14
10	0.63	400	0.10
12	0.58	∞	0.00

It is concluded that the error of the variance which corresponds to $\Delta e'$ of the original data is:

$$\pm S(X_i).\Delta e'.\Delta r \qquad (7)$$

because $s(x_i)$ is about equal to $S(X_i).\Delta r$ is related to n and follows from Table 1.

Standard deviation. There is a simple functional relationship between standard deviation S and variance S^2, namely $S = \sqrt{S^2}$.
In that case,

$$\Delta S \cong \frac{dS}{dS^2} \cdot \Delta S^2$$

(see also formula 9) or

$$\Delta S \cong \frac{\Delta S^2}{2S}, \tag{8}$$

where ΔS^2 is any error of the variance and ΔS is the corresponding error of the standard deviation.

Discussion of Shaw's Rules

Shaw considers three types of data: (1) "maximum rounding error data," (2) "exact number data," and (3) data with "variable error."

According to error theory, errors may be divided in *systematic* errors and *random* errors. Our discussion of systematic errors has been restricted to the next paragraph, because Shaw's "maximum rounding errors" are systematic errors. Shaw's "exact number data" form a separate case; his "data with variable error" should be equivalent to the general case of random errors, although Shaw himself does not use the term "random."[1]

(1) "*Maximum rounding error data.*" It is assumed that each value has a certain systematic error, which may amount to half the unit of the last digit. The problem is to calculate the propagation of the maximum systematic error.

It is obvious that the mean of a number of values, each with the same systematic error, will show this same systematic error. The variance and the standard deviation are free of it, because these statistics do not change when all the basic data are enlarged by a certain equal amount.

I agree with Shaw that, in case of a systematic error, the mean must be presented with as many figures as the original data. However, I disagree with Shaw when he states that the standard deviation should also have as many figures as the original data. When a solely systematic error is present in the basic data, and no random variable error, the error of the standard deviation is zero, and it would be correct to present the standard deviation with an infinite number of figures. However, an infinite number of figures is not allowed for the following reason: the last significant digit of the standard deviation cannot be determined from an assumed systematic error. In addition to the systematic error, an (unknown) random variable error will always be present, and this variable error of the original data determines the number of significant figures of S^2 and S according to the formulas (7) and (8).

(2) "*Exact number data.*" Shaw has suggested rules for "exact" number data which should be numbers without errors. The "exact" number 22.9975 may, for instance, be divided by 109. The quotient is

[1]Shaw implicity assumes randomness by using formula (6).

0.2109862385; this number should be rounded off to 0.210986, because $109 \times 0.210986 = 22.997474$; reporting the quotient in more figures would, after rounding off, also lead to the "exact" number 22.9975, so that reporting these extra figures does not make sense. On the other hand, reporting of fewer figures for the quotient would, after the multiplication check, result in another original "exact" number, so that reporting fewer figures results in loss of information.

Shaw's use of the term "exact" is somewhat misleading. If the number 22.9975 was really exact, the quotient might be reported with an infinite number of figures. Shaw implicitly assumes that his "exact" number has a decimal error. When it is assumed that the maximum decimal error is half of the unit of the last significant digit of the "exact" number, its propagation can be calculated by means of the procedure that has been used for calculating the error of the standard deviation (formula 8). If $f(x)$ is a function of x, and Δx and Δf are the errors of x and $f(x)$, then

$$\frac{df}{dx} \simeq \frac{\Delta f}{\Delta x}$$

or

$$\Delta f = f'(x) . \Delta x \tag{9}$$

This approximation is valid when the errors are small such as in the case of decimal errors. Otherwise, higher order terms of Δx in Taylor's series must be taken into consideration (Deming, 1943, p. 37).

For the above example, $f'(x) = 1/109$ and this results in the quotient 0.210986 ± 0.0000005 which is equal to Shaw's result. In this particular case, Shaw's rule for "exact" number data is thus more or less equivalent to the rule for the propagation of error.

(3) *Data with variable error.* Shaw proposes that his rules framed for "exact" number data also apply to cases in which data with variable error are combined to mean and variance. This approximation is not justified. Shaw himself uses the qualification "arbitrarily" (p. 242), when he discusses the circumstance that the error of the mean, which corresponds to the decimal error of the original data, is $\Sigma e'_i/n$ and assumes that this value may be neglected. However, this circumstance should not be disregarded. The mathematical expectation is that mean $(\Sigma e'/n)$ is zero, as Shaw noted, but he failed to consider that $\Sigma e'/n$ is a variable with

$$s\left(\frac{\sum e'}{n}\right) = \frac{s(e')}{\sqrt{n}} \tag{10}$$

The maximum decimal error propagation to the mean is therefore according to formula (5).

Formula (10) is met with in Deming (p. 40) and other statistical texts under circumstances which are different from those of the present case. Formula (10) is more general and solely identical to Deming's formula if x_i is constant for all i's.

Shaw's application of the "exact" number theory to the variance is not justified for reasons similar to those discussed when formula 7 was developed.

When there is a simple functional relationship—as between mean and sum of data, and between standard deviation and variance—Shaw's "exact" number theory may be applied and it may be expressed by formula (9). However, when the relation is not that simple and values with variable error are combined with each other (calculation of mean and variance from the original data) Shaw's rules may not be used.

Examples

The examples are the same as those given by Shaw so that direct comparison may be made.

Mean. Let the sum of 110 values be 7884.95. Then, according to formula (5): $\bar{X} = 71.6814 \pm 0.0005$, and according to Shaw: $\bar{X} = 71.6814 \pm 0.00005$. Shaw is over-precise by a factor 10.

Variance. Let the maximum decimal error for 110 original data be 0.005. $S^2 = 2.64139353648$ is to be rounded. From Table 1, after interpolation, it follows that $\Delta r = 0.19$; the standard deviation $S \pm 1.62536\ldots$ Then according to formula (7): $S^2 = 2.6414 \pm 0.0015$ and according to Shaw: $S^2 = 2.641394 \pm 0.0000005$. Shaw's rules again suggest too great a precision.

Standard deviation. Shaw's rule for the relationship between the numbers of significant figures of variance and standard deviation is: "one less than the variance for the standard deviation," while "the first figures in the variance and the standard deviation may be zeros" (Shaw, p. 236).

If the variance is: $S^2 = 0002.641394 \pm 0.0000005$ (Shaw's notation with zeros at the beginning is used; Shaw, p. 238), then according to (8): $S = 1.6252365 \pm 0.0000001$, and according to Shaw: $S = 01.6252365 \pm 0.00000005$.

These results are almost equal to each other. This is because formula (8) is a special case of formula (9) which is equivalent to Shaw's "exact" number theory. It is noted that Shaw's rule is less convenient in application than formula (8) because it may require that the first significant figures in the variance and standard deviation be zeros. Moreover, Shaw's present rule is not generally valid for variances smaller than one.

Finally, it has not been satisfactorily defined for variances ending in zeros (Shaw, personal communication).

Conclusion

Summarizing, it is stated that Shaw's rules for the mean and the variance are not valid for practical purposes because they suggest an accuracy which is not permitted by the accuracy of the original data. Shaw fails to consider that variable errors of the basic data combine to give errors in mean and variance which are larger than the errors obtained by the "exact" number theory.

In case of the standard deviation, Shaw's theory gives the same results as those provided by the theory of the propagation of error because there is a simple functional relationship between S and S^2. However, formula (9) is simpler and easier to use than Shaw's theory inasmuch as the latter requires extra manipulations of numbers because the rule based upon this theory is accompanied by several sub-rules.

Additional remarks

A problem of secondary importance arises if one wants to avoid mentioning the propagated maximum decimal error itself and express this error by means of the last digit of the statistics. It must then be changed into a maximum decimal error of the statistics. There are two possibilities: the statistic might be rounded off in such a way that the last digit is truly significant, or one might report an extra digit so that nothing of the reported precision of the original data is lost. In case of rounding off the original data, the second method is preferable; in case of rounding off the statistics, it must be considered that these values become more "over-precise" than the original data when the second method is used. This is because they are based upon the original data which have an assumed decimal error that is smaller than the true error.

Finally, the accuracy by which the calculated statistics (of the sample) approximate the true statistics (of the population from which the sample has been drawn) might be considered. This accuracy is generally much less than the accuracy obtained after application of the rules of this letter. The problem of accuracy of data in geochemistry will be considered more closely in a future paper by the writer.

Acknowledgments

Thanks are due to Dr. S. C. Robinson, Dr. K. R. Dawson, Dr. D. M. Shaw, and Dr. T. N. Irvine for critical reading of the manuscript.

References

DAVID, F. N. (1938): *Tables of the ordinates and probability integral of the distribution of the correlation coefficient*; Biometrika office, London.

DEMING, W. E. (1943): *Statistical adjustment of data*, John Wiley & Sons, New York.

EISENHART, C. in EISENHART, C., MASTAY, M. W., & WALLIS, W. A., editors (1947): *Selected techniques of statistical analysis*, McGraw-Hill, New York.

International dictionary of physics and electronics (1956): Van Nostrand, Princeton, New Jersey.

LAITINEN, H. A. (1960): *Chemical analysis*, McGraw-Hill, New York.

PEARSON, E. S. & HARTLEY, H. O., editors (1954): *Biometrika tables for statisticians*, **1**, Cambridge University Press.

SHAW, D. M. (1962): Rounding errors and Chayes' paradox, *Can. Mineral.*, **7**, 236–244.

F. P. AGTERBERG
Geological Survey
601 Booth Street, Ottawa, Canada

November 28, 1962