ROUNDING ERRORS AND CHAYES' PARADOX

D. M. SHAW

Department of Geology, McMaster University, Hamilton, Ontario

> *"Significant numbers are far from ideal as a means of expressing the results of fundamental operations."* (*Dwyer*, 1951, *p.* 14.)

ABSTRACT

The rounding rule in computation is based on the principle that no result of computation can have more significant figures than the original data. Use of this principle in statistics is ill-advised and leads to loss of information by rounding.

Rules for determining the maximum number of figures to be retained in simple statistical calculations are developed. Let m be the number of figures in the maximum plus the minimum values of the sample of n observations, where $n = f \times 10^k$ (f is a fractional number between 1 and 10 and k is an integer). Then:

(A) for samples of at least 100 the maximum number of figures will be,
- (a) $m + k$ or $m + k + 1$ for the sum and the mean,
- (b) $2m + k$, $2m + k + 1$ or $2m + k + 2$ for the variance,
- (c) one less than the variance for the standard deviation;

(B) for smaller samples the maximum number of figures will be,
- (a) m or $m + 1$ for the sum and the mean,
- (b) $2m$, $2m + 1$ or $2m + 2$ for the variance,
- (c) one less than the variance for the standard deviation;

(C) the first figures in the variance and standard deviation may be zeros;

(D) when a calculated statistic is to be used in further calculations it is preferable not to round it first;

(E) the sum and sum of squares should be always included with summary statistics, to prevent ambiguity.

Analogous methods can be readily applied to calculations of other statistics.

INTRODUCTION

Any scientist who has to submit data to statistical procedures has to make decisions about significant figures and rounding errors. As far as can be ascertained from a rather cursory survey of recent literature, there are few guiding rules and those are mostly relevant to general computation problems. The situation in statistical work is not necessarily the same.

It may be that rounding has received little attention because most workers find it too rudimentary, although Eisenhart (1947) has made a penetrating study.* However, it is likely that the geochemical journals will be inundated with ever-increasing numbers of analytical data in the

*Eisenhart's paper is concerned chiefly with the effects of rounding the original data, and, although fundamental to the topics here discussed, deals with them only implicitly.

years to come, and as Chayes (1953) has advisedly pointed out, the presentation of summary statistics such as the sum or the sum of squares should neither discard information contained in the original data, nor attempt to add to it.

With Chayes, I believe that information is often discarded by overzealous rounding, but would go further than Chayes in retaining figures.

Standard Practice

Two texts on numerical analysis (Dwyer, 1951; Nielsen, 1956) give some error theory and standard practice for approximate numbers.

Nielsen discusses briefly the manner in which rounding errors may affect the results of computations. He gives the following (familiar) rule (p. 3). "During the computation retain one more figure than that given in the data and round off after the last operation has been performed." He subsequently (p. 4) notes that no rule is infallible and that the procedure adopted should be dictated by the type of problem in hand.

Dwyer devotes a chapter of his book to a useful discussion of significant numbers and to the consequence in computation of various types of error symbolism and rounding. Two rules are proved:

(i) The product or quotient of two numbers, each containing m significant figures (at least two of which are not zero), is a significant number of at least $(m - 2)$ figures. If the leading digits of these numbers are both equal to or greater than 2, then the product (or quotient) has at least $(m - 1)$ significant figures.

(ii) The square root of an m-place number is significant to $(m - 1)$ places if the first digit is $\leqslant 5$ and to m places if >5.

The first rule, giving the minimum number of significant figures, is not much help. Thus, if we evaluate 1.11^2, where $m = 3$, then the rule merely tells us that in the product, 1.2321, at least the integer is significant.

Another general rule which is commonly upheld states that *no result of computation can have more significant figures than the original data.* This is equivalent to the first rule cited from Nielsen (1956). I do not believe this to be valid in statistical operations, and will attempt to prove so. It will be referred to as the *rounding rule* in the discussion which follows.

General Considerations

Let us first note that rounding in computation can be made at four main levels: (a) on the raw data; (b) at one or more intermediate steps in the computation; (c) on the result; (d) in the tabulation and publication of any of the preceding three. Item (*d*) is not of course part of the computation, but Chayes (1953) rightly stressed the obligation of authors to publish data in a form such that the reader can verify the computations, even if this means supplying more figures than would customarily be regarded as significant.

The last sentence illustrates the paradox we are faced with: *i.e.*, the figures are necessary but not significant. This paradox results from application of the rounding rule.

Numerical data fall in the categories of exact or approximate. If there are 233 zircons in a sample of 100,000 sand grains then the percentage of zircon is 0.233 exactly. On the other hand, if a chemist finds 0.6899 gm. SiO_2 in a sample of rock weighing 1.0035 gm., the content of 68.74937 *etc.* per cent is rounded off to the same number of significant figures as the weighed SiO_2, *i.e.*, to give 68.75. The weight of SiO_2 was not of course exactly 0.6899 gm., but might have been anywhere between 0.68985 and 0.68995. The maximum error attached to the weighing thus is 5 in the fifth place: similarly, the maximum error in the percentage 68.75 is ± 0.005. This last term is the ***rounding error***.

In the following discussion we will first consider data which all possess the maximum possible rounding error, then exact-number data. The usual situation will be somewhere in between and will be treated last. Mean and standard deviation calculations will be used as examples.

It will be convenient to refer to the total number of figures rather than significant figures. The only difference will be that zeros at the beginning will be included. Thus the difference between two 8 total figure numbers with the same number of decimals will always be another 8-figure number:

$$e.g.\ 7619.3425 - 7613.2122 = 0006.1303$$

Customary usage would describe the difference as a five-figure significant number, ignoring the first three zeros.

Also the number of figures m in a set of data will include all the figures in the largest plus the smallest observation. Thus if these are 12.11 and 0.002 respectively then m is equal to 5.

These two conventions will facilitate subsequent discussion.

Maximum Rounding Error Data

Consider a set of n analyses for silica, each to 4 significant figures (*e.g.* 68.75 wt. per cent): the maximum error is ± 0.005. Let any actual silica content x_i be measured as X_i with the maximum positive error $+\epsilon$. Then

$$X_i = x_i + \epsilon$$

$$\sum_n X_i = \sum_n x_i + n\epsilon$$

$$\bar{X} = \bar{x} + \epsilon$$

This implies that the sum and the mean will each have the same number of significant figures as an original observation. This is the rounding rule. A similar conclusion arises of course if we take the maximum negative error. For example, 110 chemical analyses of SiO_2 in the Mortagne granite (Shaw, in press) give the following statistics:

$$\sum X_i = 7884.95, \qquad \bar{X} = 71.681364 \text{ etc.}$$

Assuming that every individual analysis (*e.g.* 75.12) is in error by +0.005 then,

$$\sum x_i = 7884.95 - 0.55,$$

and

$$\bar{x} = 71.681364 - 0.005.$$

We have therefore to round these numbers to

$$\sum x_i = 7885,$$

and

$$\bar{x} = 71.68$$

Next consider the variance*:

$$(n - 1)\, var\, X_i = \sum_n (x_i + \epsilon)^2 - \frac{1}{n}\left[\sum_n (x_i + \epsilon)\right]^2$$

$$= \sum_n x_i^2 - \frac{1}{n}\left[\sum_n x_i\right]^2$$

That is, the error term has dropped out and this implies that we can treat X_i as if it were x_i and the right-hand side of the equation will be evaulated to the last place in the square of any of the terms:

e.g. square of 76.18 is 5803.3924, and $\sum X_i^2 = 565491.8801$, so the right-hand side of the equation above, which for the same example works out to 287.911895455, is rounded off to 287.9119: the variance is obtained by dividing by 109 to obtain 2.6414 after rounding.

It should be noticed that figures have been lost as a result of the subtraction of two similar numbers. The variance is more correctly written as

*The usual computing form for $(n - 1)$ *var* x, as used here, is to be preferred over the sum of the squares of the deviations from the mean, unless calculations are to be evaluated to many decimal places.

0002.6414, conforming to the 8 figures of the square of an analysis, but customary usage forbids calling the zeros significant figures.

The standard deviation was computed as 01.62523645550 but should presumably be rounded to 01.63, to have the same number of figures as the data. Note however, that the square of 1.63 is 2.6569, so the rounded standard deviation is of less value than the rounded variance.

It is clear that the foregoing treatment amounts to analysing the situation where every reading is biassed by the constant amount ϵ. The results could have been obtained more simply if we assume that x is $N(\mu, \sigma^2)$, from the relationship that X must be $N(\mu + \epsilon, \sigma^2)$.

This relationship is not likely to conform to most runs of experimental values, where ϵ will vary in magnitude and sign.

Exact Number Data

An observation has m figures with no error. This situation could arise in enumeration statistics, or alternatively with a continuous data-variable in the case where the rounding error ϵ is always zero: *e.g.* $68.75 = X_i = x_i$.

If we have n values then write

$$n = f \times 10^k$$

where f is a fractional number between 1 and 10 and k is an integer. The sum $\sum_n x_i$ will have a maximum of $(m + k)$ or $(m + k + 1)$ figures, all significant: the mean value $\bar{x}$ will have the same number.

e.g. in the previous example $m = 4$; $n = 110 = 1.10 \times 10^2$
then $\sum x_i = 7884.95$ and has $4 + 2 = 6$ significant figures.

Moreover $\bar{x} = 71.6814$ and also has 6 significant figures.
This should be clear from the principle that the sum has no rounding errors and the mean must permit the recalculation of all the figures in the sum (110 × 71.6814 equals 7884.954).

Consider now the variance. The square of an m-figure observation will have a maximum of $2m$ figures. The sum $\sum_n x_i^2$ will therefore have a maximum of $(2m + k)$ or $(2m + k + 1)$ significant figures, as also will $(\sum_n x_i)^2/n$. Thus $(n - 1)$ *var* x_i will be the difference between these two numbers, and will have a maximum of $(2m + k)$ or $(2m + k + 1)$ figures, *the first few of which may be zeros*: the variance will have the same number of figures, except that the operation of dividing by $(n - 1)$ allows the additional possibility of $(2m + k + 2)$ as a maximum.

Note that uncertainty only arises where division (or taking a root) has to be carried out (rounding enters). The alternatives of $(2m + k)$ or

$(2m + k + 1)$ figures for the sum of squares means simply that all the figures are retained: similarly for the quantity $(n - 1)\,var\,x_i$. The possibility of $(2m + k + 2)$ for the variance is seen clearly in the following two examples:

A. First digit in $n - 1$ less than first digit in numerator:

$$\begin{aligned} (n-1)\,var\,x_i &= 0022.9975 && \text{8 figures (2 zeros)} \\ n - 1 &= 109 && \\ \text{quotient} &= 00.2109862385 && \\ var\,x_i &= 00.210986 && \text{8 figures (2 zeros)} \end{aligned}$$

Check: $109 \times 00.210986 = 0022.997474$

B. First digit in $n - 1$ greater than first digit in numerator:

$$\begin{aligned} (n-1)\,var\,x_i &= 0022.9975 && \text{8 figures (2 zeros)} \\ n - 1 &= 912 && \\ \text{quotient} &= 00.02521655701 && \\ var\,x_i &= 00.0252166 && \text{9 figures (3 zeros)} \end{aligned}$$

Check: $912 \times 00.0252166 = 0022.9975392$,

but $912 \times 00.025217 = 0022.997904$ which is in error in the eighth figure.

Thus for the data of the previous example we have

$$\begin{aligned} m &= 4 \qquad n = 110 \qquad k = 2 \\ \sum x_i^2 &= 565491.8801 \qquad \text{with } 10 = 2m + k \text{ figures} \\ (n-1)\,var\,x_i &= 000287.911895455 \text{ significant to 4 decimals* (10 figures)} \\ n - 1 &= 109 \\ \text{quotient} &= 0002.64139353648 \\ var\,x_i &= 0002.641394 = s^2 \end{aligned}$$

In order to calculate the standard deviation s we note that s^2 must be correct in the sixth decimal. The root of the variance must be taken to the seventh decimal place and is 01.6252365. Two zeros have, however, been lost from the beginning, thus reducing the total by one, to $2m + k - 1$.

We may summarise the rules for exact numbers as follows:

(*a*) the sum and the mean will have a maximum of $(m + k)$ or $(m + k + 1)$ figures;

(*b*) the variance will have a maximum of $(2m + k)$, $(2m + k + 1)$ or $(2m + k + 2)$ figures;

*The additional figures come from using an unrounded value for $\bar{x}$.

(*c*) the standard deviation will have one figure less than the variance;
(*d*) the first figures in the variance and standard deviation may be zeros.

These conclusions of course contravene the rounding rule.

Variable Error

The two previous sections outline the rules for the limiting cases of maximum rounding error and no rounding error. In practice we usually have to deal with approximate number data where the rounding error varies from zero to the maximum. That is

$$X_i = x_i + \epsilon_i$$

where ϵ_i may be positive or negative. Let us assume that ϵ_i is $N(0, \sigma^2_\epsilon)$. Then

$$\sum_n X_i = \sum_n x_i + \sum_n \epsilon_i$$

and

$$\bar{X} = \bar{x} + \bar{\epsilon}$$

Now

$$E(\bar{\epsilon}) = 0,$$

but in general $\bar{\epsilon}$ will have a small finite value and will decrease as n increases. In any case $\bar{\epsilon}$ will be considerably less than the maximum rounding error and the mean value $\bar{X}$ may legitimately be allowed more significant figures than a single observation.

I propose arbitrarily, that for a large sample ($n \geqslant 100$) $\bar{\epsilon}$ be disregarded, so that the sum and the mean be allowed the same number of figures as for exact data; *i.e.*, $(m + k)$ or $(m + k + 1)$. For smaller samples reduce each by one, to obtain m and $m + 1$ respectively (since $k = 1$). This conforms with the use of the rounding rule for small samples or small numbers of operations. In presenting summary statistics however the sum should always be included, since it cannot be accurately recalculated from the rounded small sample mean.

For the variance, we have the relationship

$$var\ X = var\ x + var\ \epsilon + 2\ cov_{x\epsilon}$$

It is usually reasonable to assume that x and ϵ are independent. The expected value of the covariance will then be zero although the sample value will not. Thus

$$S^2 = s^2 + s^2_\epsilon + 2s_{x\epsilon}$$

We do not know s^2_ϵ, but if the maximum rounding error is ±0.005 then we know that $s^2_\epsilon < 0.000025$. The two-decimal data which have this error will square to four-decimal numbers, so the rounding error variance will always be negligible.

The covariance can be put in the following form:

$$s_{x\epsilon} = r\sqrt{var\,x\,.\,var\,\epsilon} = r\,.\,s\,.\,s_\epsilon,$$

where r is the correlation coefficient of x and ϵ.
Thus

$$S^2 = s^2 + 2r\,.\,s\,.\,s_\epsilon$$

The population correlation coefficient ρ is zero: r however may have a small finite value (within ±0.3 for a sample of size 50 for 95% probability).

We may, therefore, approximate again and write

$$S^2 = s^2 \pm \frac{s\,.\,s_\epsilon}{2}$$

The significance of the covariance term will thus depend on the magnitude of s and no general rounding rule can be established. We know, however, that the expected value of this term is zero and it will be very small for any large sample. We also know from the previous discussion that the maximum permissible number of figures must lie between $2m$ (maximum error) and $2m + k$, $2m + k + 1$ or $2m + k + 2$ (no error).

It appears reasonable to accept an arbitrary rule similar to the one for the sum and the mean. That is, for large samples ($n \geqslant 100$: $k \geqslant 2$) the variance can be allowed $2m + k$, $2m + k + 1$ or $2m + k + 2$ figures, as for exact number data. For smaller samples reduce by one to obtain $2m$, $2m + 1$, $2m + 2$ respectively. These alternatives all include any beginning zeros.

As discussed above the standard deviation should be allowed one figure less than the variance, this to include beginning zeros. We may summarise as follows:

(*A*) for samples of at least 100 the maximum number of figures will be,
(*a*) $m + k$ or $m + k + 1$ for the sum and the mean,
(*b*) $2m + k$, $2m + k + 1$ or $2m + k + 2$ for the variance,
(*c*) one less than the variance for the standard deviation;
(*B*) for smaller samples the maximum number of figures will be,
(*a*) m or $m + 1$ for the sum and the mean,
(*b*) $2m$, $2m + 1$ or $2m + 2$ for the variance,
(*c*) one less than the variance for the standard deviation;
(*C*) the first figures in the variance and standard deviation may be zeros;

(*D*) when a calculated statistic is to be used in further calculations it is preferable not to round it first;

(*E*) the sum and sum of squares should be always included with summary statistics, to prevent ambiguity.

Analogous methods can be readily applied to calculations of other statistics.

It should be stressed that the foregoing discussion is solely concerned with the manipulation and presentation of statistical data. The physical meaning of the results must always provide the argument for any further rounding.

Acknowledgments

This note has arisen out of geochemical research supported by grants from the Geological Survey of Canada and the National Research Council. I am indebted to Dr. G. V. Middleton and Dr. D. J. Kenworthy, of the Departments of Geology and of Mathematics, McMaster University, for helpful discussion and criticism.

References

Chayes, F. (1953): In defence of the second decimal. *Am. Mineral.*, **38**, 784–793.

Dwyer, P. S. (1951): *Linear computations.* 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.

Nielsen, K. L. (1956): *Methods in numerical analysis.* MacMillan, New York.

Shaw, D. M.: The variance of oxides in a granite. An attempt to measure some geochemical parameters (in press).

Manuscript received November 10, 1961