# ON THE MATHEMATICALLY SIGNIFICANT FIGURES IN THE SOLUTION OF SIMULTANEOUS LINEAR EQUATIONS

By L. B. Tuckerman

*The National Bureau of Standards*

1. **Introduction.** The number of mathematically significant figures in the solution of simultaneous linear equations has received attention from a number of writers [1–6]. It is an important subject, not only in least squares and correlations, but in many other problems of science where simultaneous equations arise: it may not be amiss, therefore, to examine it from a fresh start, particularly since (as will be shown) some of the rules that have been published on it fail in certain frequently occurring circumstances.

2. **Definitions.** Before proceeding into the subject it will be necessary to distinguish between the computer's terms "significant figures" and "determinate significant figures." The former are the figures that compose a number, without the consecutive ciphers that precede or follow them, merely to locate the decimal point. "Determinate significant figures," on the other hand, are figures that are justifiable on computational grounds. From the computer's point of view, the number of significant figures remains independent of what is statistically significant. To avoid confusion in what follows, the term "significant figures" will be used in the computer's sense, and the adjective "determinate" will be supplied where mathematical determinacy is implied.

To avoid prolixity the term "observational error" will include any uncertainty arising either from errors in the observations or from the statistical nature of the problem (e.g. sampling errors, grouping errors, etc.). *The observational error of the result is independent of the particular sequence of computation followed and the accuracy with which it is carried out.*

The term "computational error" will include all the additional uncertainties arising from the approximations occurring in the particular sequence of computation used, including the "rounding off" of the final result. *The computational errors, unlike the observational errors, depend in general upon the sequence of the intermediate steps used in the computation as well as on the number of significant figures to which they are carried.*

3. **Criterion of an adequate computation.** If the number written down at the end of a computation is to serve its purpose the maximum possible computational error must be suitably limited.

A decimal representation of a number containing $f$ significant figures is subject

to an uncertainty (upper limit of absolute error) of 5 in the $(f + 1)$th place. It has, therefore, a possible relative (not absolute) error of representation somewhere between $5 \times 10^{-(f+1)}$ and $5 \times 10^{-f}$, in magnitude. This relative computational error sets the limit to any valid final rounding off. Regardless of the accuracy to which the intermediate steps of the computation have been carried, this relative computational error introduced by the final rounding off alone must be suitably limited.

In case all of the accuracy obtainable from the data is not needed in the result, the sum of the maximum possible computational error (including the error of the final rounding off) and the maximum possible observational error must be kept below the error which can be tolerated in the result.

In case all of the accuracy obtainable from the data is needed in the result, the maximum possible computational error in the result (including the error of the final rounding off) must be negligible in comparison with the uncertainty (observational error) in the result arising from uncertainty in the data. *Just how small a fraction of the observational error is "negligible" is necessarily a matter of judgment, and will depend upon the nature of the problem.* A computational error that would be wholly negligible in some ordinary computations might be intolerably large in the adjustment of an accurate geodetic survey. In any case the only basis for a valid judgment of the adequacy of the computation lies in a comparison of (i) the maximum possible computational error that can arise in the sequence of computations including the final "rounding off," with (ii) the observational error of the result arising from the observational errors inherent in the data.

4. **Propagation of error in a system of linear equations.** Assume that

$$\sum_t a_{st} x_t = b_s, \qquad s = 1, 2, \cdots, n, \tag{1}$$

is a set of simultaneous linear equations derived in some way from observations and in which the coefficients $a_{st}$ and the absolute terms $b_s$ may all be subject to observational error. If the relative (not absolute) observational error of a quantity $q$ be represented by $\delta_q$ it may readily be seen that

$$\begin{cases} \delta x_j = -\sum_h \sum_k (x_k/x_j) A_{hj} a_{hk} \delta a_{hk} + \sum_s (b_s/x_j) A_{sj} \delta b_s \\ \delta \Delta = \sum_h \sum_k A_{hk} a_{hk} \delta a_{hk} \end{cases} \tag{2}$$

where $\Delta$ is the determinant of the coefficients $a_{hk}$, and $A_{hk}$ is the term corresponding to $a_{hk}$ in the reciprocal (not the adjoint) determinant.

5. **Upper limits to observational errors.** The sign and magnitude of the relative errors $\delta a_{hk}$ and $\delta b_s$ are unknown, but we shall assume that it is possible

in any problem to assign to them *upper limits*

$$|\delta a_{hk}| \quad \text{and} \quad |\delta b_s|$$

which in magnitude they cannot exceed. If the problem is such that the values of each of the $\delta a_{hk}$ and the $\delta b_s$ are wholly independent of each other, it is then possible that their magnitudes may all reach their upper limits $|\delta a_{hk}|$ and $|\delta b_s|$ simultaneously, in which case *upper bounds* of $\delta x_j$ and $\delta\Delta$ may be placed at

$$(3) \qquad \begin{cases} |\delta x_j| = \sum_h \sum_k |(x_k/x_j)A_{hj}a_{hk}|\,|\delta a_{hk}| + \sum_s |(b_s/x_j)A_{sj}|\,|\delta b_s| \\ |\delta\Delta| = \sum_h \sum_k |A_{hk}a_{hk}|\,|\delta a_{hk}| \end{cases}$$

6. **Indefiniteness of the problem in the general case.** The values of the $\delta a_{hk}$ and $\delta b_s$ may not be independent of each other, in which circumstance knowledge of the law of their dependence would make it possible to assign upper limits to the magnitudes of $\delta x_j$ and $\delta\Delta$. These upper limits can not be larger than the upper bounds shown in equation (3), and in special cases they will be much smaller. Since the dependence of $\delta a_{hk}$ and $\delta b_s$ may in general have any form whatever, cases can and will occur in which the upper limits of the relative errors of $\delta x_j$ and $\delta\Delta$ may have any ratio whatever.

7. **Case of independent errors.** Any general discussion of the errors that can occur in $x_j$ and $\Delta$ must be based either on some special assumption or on the limiting assumption that the errors are independent. It is this latter assumption that underlies the usual discussion, and will be the basis of what follows. Equation (3) gives the upper limit to the $\delta x_j$ and $\delta\Delta$ under these assumptions.

8. **The ratios of $|\delta x_j|$ and $|\delta\Delta|$ are still indefinite in spite of the assumption of independent errors in the coefficients.** However, equation (3) does not determine any definite ratio or inequality between the upper bounds $|\delta x_j|$ and $|\delta\Delta|$. The nature of the observations may be such that some of the errors in the $a_{hk}$ and $b_s$ are very small and some relatively large. Not infrequently it is safe to assume that some of them are free from appreciable error and to ascribe all the error of the $x_j$ to the error in one or two of the $a_{hk}$ or $b_s$. If any statement of a definite relationship, either as an equality or an inequality between $|\delta\Delta|$ and the $|\delta x_j|$ is valid for all possible sets of linear equations, it must at least hold in the special case in which the errors of all the $b_s$ and the errors of all except one of the $a_{hk}$ are negligible.

If such a statement of a definite general relationship between these upper limits of errors can be made, it must be possible to write down an equation or an inequality between any one of the expressions $|A_{hk}|$ and some or all of the corresponding expressions $|(x_k/x_j)A_{hj}|$, $j = 1, 2, \cdots, n$, that will remain true no matter what be the values of the $a_{hk}$ and the $b_s$ in the original set of simultaneous equations. It is obvious that the ratio of $|A_{hk}|$ and $|(x_k/x_j)A_{hj}|$, $(j \neq k)$, depends upon the values of the $a_{hk}$, and sets of equations can be found

to give any assigned value to that ratio. It is therefore impossible to state any rule that will restrict the ratio of the relative error of $\Delta$ and the relative error of any one of the $x_j$, valid for all possible sets of linear equations.

9. **Definite statement about the sum of the relative errors in the unknowns.** However, in the summation $\sum_j |\delta x_j|$ there occurs the term corresponding to $j = k$, for which $|(x_k/x_j)A_{hj}| = |A_{hk}|$, so that under the assumption that the $a_{hk}$ and $b_s$ are independent sources of error, we may write the inequality

$$\sum_j |\delta x_j| \not< |\delta\Delta| \tag{4}$$

which states that the sum of the upper bounds to the relative errors of all the $x_j$ cannot be less than the upper bound to the relative error of the determinant $\Delta$. A corresponding statement can easily be proved for the standard deviations.

A limiting case can be constructed in which the inequality (4) reduces to

$$\sum_j |\delta x_j| = |\delta\Delta| \tag{5}$$

and in which all of the $|\delta x_j|$ are equal. For this case,

$$|\delta\Delta| = n|\delta x_j| \text{ for all values of } j. \tag{6}$$

If $n \not< 10$ it is obvious that there will be at least one more determinate significant figure in each of the $x_j$ than in the determinant $\Delta$ of the coefficients.

It is frequently assumed that the number of determinate significant figures in the solution for any unknown cannot exceed the number of determinate significant figures in the determinant $\Delta$ of the coefficients. We see now that this statement can not be generally valid, even under the assumption that the $a_{hk}$ and $b_s$ are independent sources of error. As a matter of fact, it is necessary in some cases to compute some or even all of the unknowns to more significant figures than are determinate in the determinant $\Delta$ of the coefficients, if one would retain in the result all the accuracy that is obtainable from the data.

Cases in which the relative observational error of every one of the unknowns is less than the relative error of the determinant $\Delta$ probably occur rarely in practice; in fact the only ones that I have seen are those that I constructed purposely to show that such a thing is possible. However, cases in which the relative errors of one or several but not all of the unknowns are much smaller than the relative error of the determinant $\Delta$, occur fairly frequently.

10. **Remarks on the case of "near indeterminacy."** The major interest in curve fitting centers around the condition of "near indeterminacy," i.e., of a small or near vanishing determinant $\Delta$. Even in the circumstance where the relative error of the determinant is much greater than the relative error of some or all of the coefficients and absolute terms, the relative error of one or more of the unknowns may be much smaller than the relative error of the determinant, as may be seen from what follows.

In accurate experimentation the endeavor is, wherever possible, to arrange the experiment so that the quantity sought comes directly from the measurement as represented by an equation such as

$$x = p. \tag{7}$$

However, so ideal an experimental arrangement is rarely if ever possible, and it is a common experience to find that the measurements are represented by an equation such as

$$x + qy + rz + su + \cdots = p, \tag{8}$$

where $qy$, $rz$, $su$, etc., are small corrections that must somehow be evaluated. For simplicity, the discussion will be confined to the almost trivial case

$$x + qy = p. \tag{9}$$

Not infrequently the only way the correction can be evaluated is to rearrange the conditions of the experiment so that another equation is obtained in the form

$$x + q'y = p'. \tag{10}$$

Sometimes the nature of the experiment is such that it is not possible to change the coefficient of $y$ by more than a small amount, under which conditions

$$q' = q(1 + \beta), \tag{11}$$

and

$$p' = p(1 + \alpha), \tag{12}$$

where $\beta$ and $\alpha$ are small in comparison with 1. The solution of equations (9) and (10) now gives

$$x = \frac{\begin{vmatrix} p & q \\ p' & q' \end{vmatrix}}{\begin{vmatrix} 1 & q \\ 1 & q' \end{vmatrix}} = \frac{pq' - p'q}{q' - q} = p(1 - \alpha/\beta). \tag{13}$$

The quantity $q' - q$ seen in the denominator of this equation is the determinant $\Delta$ of the coefficients, and by equation (11) its value is $\beta q$. Since $\beta q$ is assumed to be small here, the solution for $x$ encounters a near vanishing denominator. It would, however, be wrong to assume that the number of determinate significant figures in $x$ that can be obtained by solving the equations is necessarily limited to the number of determinate significant figures in the denominator $\Delta$.

If the experimenter has been fortunate in finding suitable experimental conditions, the denominator $\Delta = \beta q$, although small in comparison with either $q'$ or $q$, will still not cause difficulty. It will be observed that the coefficients of $q'$ and $q$ in the denominator are equal (both being unity). Now if the coefficients $p$ and $p'$ in the numerator are nearly enough equal, so that $q'$ and $q$ occur in both

numerator and denominator so nearly proportionally that the uncertainties in $q$ and $q'$ produce nearly compensating errors in both numerator and denominator, then $x$ will be given to more determinate significant figures than are found in the denominator $\Delta$. It can then be said that the experiment is successful in evaluating the correction term $qy$ in equation (9).

On the other hand, in less fortunate circumstances, to the exasperation of the experimenter, the denominator $\Delta = q' - q = \beta q$ is not only small, but $p'$ and $p$, although still nearly equal, differ enough so that the errors in $q'$ and $q$ are not compensated by the nearly equal coefficients in the numerator. The experiment will then fail to improve the approximation $p$ for $x$ by failing to evaluate the small correction $qy$ in equation (9). This would be an inherent defect in the experiment and could not be removed by any manner of computation.

The same conclusion would of course be drawn from the coefficient of $p$ (viz., $1 - \alpha/\beta$) at the extreme right of equation (13). It is not the size of $\beta$ that alone determines the number of determinate significant figures in $x$, it is rather the ratio between $\alpha$ and $\beta$. In the fortunate experimental circumstances described above, the near equality of $p'$ and $p$ offsets the near equality of $q'$ and $q$ by reducing the term $\alpha/\beta$ to a value small compared with unity; the term $\alpha/\beta$, being small, acts to reduce the effect of the uncertainties in $q$ and $q'$ (i.e., in $q$ and $\beta$) in the evaluation of $x$. On the other hand, in less fortunate circumstances, the correction term $\alpha/\beta$ can not now shield $x$ from the uncertainties in $q$ and $q'$ since the relative difference $\alpha$ between $p$ and $p'$ is not small enough to reduce $\alpha/\beta$ to innocuity.

11. **Numerical illustration of compensating errors.** As a "horrible example" especially constructed to emphasize the theoretical possibilities, take the following special case—

$$\begin{cases} 1000.10000x + 10.00000y = 1010.10000 \\ 1000.00000x + 10.00000y = 1010.00000 \end{cases} \tag{14}$$

wherein it is assumed that the coefficients and the absolute terms (assumed to be derived from the observational data) are all correct to the fifth decimal place as given, and no closer estimate of their errors is possible. So far as known, the upper limit to the absolute observational error of each is then the same, i.e. $5 \times 10^{-6}$, but the coefficients of $x$ ($a_{11}$ and $a_{21}$), and the absolute terms ($b_1$ and $b_2$), all have nine determinate significant figures, while the coefficients of $y$ ($a_{12}$ and $a_{22}$), have only seven. Thus,

$$|\delta a_{11}| \not> 5 \times 10^{-9}, \quad |\delta a_{21}| \not> 5 \times 10^{-9}, \quad |\delta b_1| \not> 5 \times 10^{-9}, \quad |\delta b_2| \not> 5 \times 10^{-9},$$

but

$$|\delta a_{12}| \not> 5 \times 10^{-7}, \quad |\delta a_{22}| \not> 5 \times 10^{-7}, \tag{15}$$

and $x = 1$, $y = 1$, $\Delta = 1$, whereupon a substitution of values from (15) into (3) gives the inequalities

$$(16) \quad |\delta x| \not> 3 \times 10^{-4}, \qquad |\delta y| \not> 3 \times 10^{-2}, \qquad |\delta \Delta| \not> 1.01 \times 10^{-2}.$$

So far as known, the determinant $\Delta$ may thus be in error by as much as 1 per cent, and $y$ by as much as 3 per cent, yet $x$ is known closer than 1/30th per cent. Here the value of the unknown $x$ cannot be adequately represented by less than four significant figures, and might even require five, in spite of the fact that neither $\Delta$ nor $y$ requires more than three significant figures to represent all that is certainly known about them.

The reason for this disparity in relative errors can be more easily seen by substituting numerical values for all the coefficients in the expression for $x$ except $a_{12}$ and $a_{22}$. The possible relative errors of $a_{12}$ and $a_{22}$ are, as noted above, about 100 times as great as the possible relative errors of $a_{11}$, $a_{21}$, $b_1$, and $b_2$, and are the controlling errors in $\Delta$. In the solution

$$(17) \qquad x = \frac{1010.10000a_{22} - 1010.00000a_{12}}{1000.10000a_{22} - 1000.00000a_{12}},$$

however, both $a_{12}$ and $a_{22}$ occur in both numerator and denominator, and moreover the coefficient of each in the numerator is nearly equal to its coefficient in the denominator, so that a change in either $a_{12}$ or $a_{22}$ changes both numerator and denominator nearly proportionally, with the result that their ratio $x$ is known much more accurately than either the numerator or the denominator $\Delta$.

This kind of compensation of errors in a computation is not confined to the solution of simultaneous equations (and it is not an infrequent occurrence in other computations). This is one of the many reasons why it is impossible to give general rules for the retention of significant figures that will be valid for all types of computations.

12. **Geometrical analogy.** Moulton [4] illustrated his reasoning by the following geometrical analogy. The solution of three linear equations is equivalent to finding the point of intersection of three planes. When the determinant of the coefficients is small in comparison with the coefficients themselves, these planes are either nearly parallel, or the line of intersection of any two of them is nearly parallel to the third. In these cases small uncertainties in the location of any one of the planes correspond to large uncertainties in the position of their point of intersection.

In the first circumstance the planes might all be nearly parallel to one of the three coordinate planes, with the result that large uncertainty would afflict the value of the determinant and two of the unknowns, the third being much more accurately determined.

In the second circumstance, the line of intersection of two of the planes might be nearly parallel to one of the coordinate axes. When that happens, large un-

certainty will afflict the value of the determinant, but only one of the unknowns, the other two being much more accurately determined.

This geometrical analogy can be extended to cover simultaneous equations with any number of unknowns. Near-vanishing of the determinant $\Delta$ of the coefficients necessarily implies relatively large uncertainties in the determinant and also in at least one of the unknowns, but not necessarily in all of them. These are, of course, very special cases, but, as noted above, they are of frequent occurrence in actual problems.

13. **Evaluation of computational error.** The relative computational error in $x_j$ must be kept within certain definite limits which depend upon the particular problem to be solved (section 3). To do this it is necessary to be able to calculate an upper bound to the relative computational error inherent in any particular sequence of computations.

In many computations it is easy to write down a simple formula that will set an upper bound to the relative computational error involved in that particular sequence. This formula contains numbers $f_1, f_2, f_3$, etc., each representing the number of significant figures accurately computed at some particular step. Once a simple formula for relative computational error is written down, it is easy to choose values of $f_1, f_2, f_3$, etc. that will give an upper bound to the relative computational error not larger than the permissible limit of maximum possible computational error outlined in section 3. This method of determining an upper bound of the relative computational error should be used whenever such a simple formula can be found. For example, to compute $x$ from equation (13) we may use the following sequence: $r_1 = q' - q$, $r_2 = r_1/q = \beta$, $r_3 = p' - p$, $r_4 = r_3/p = \alpha$, $r_5 = r_4/r_2 = \alpha/\beta$, $r_6 = 1 - r_5 = 1 - \alpha/\beta$, $r_7 = pr_6 = p(1 - \alpha/\beta) = x$. $x$ may then be written as a function of these partial results, viz.:

$$x = r_7 = pr_6 = p(1 - r_5) = p(1 - r_4/r_2) = p(1 - r_3/pr_2) = p(1 - r_4q/r_1). \tag{18}$$

Applying first order error theory we find

$$|\epsilon(x)| \leqq \left|\frac{\alpha/\beta}{1 - \alpha/\beta}\right| \{|\epsilon(r_1)| + |\epsilon(r_2)| + |\epsilon(r_3)| + |\epsilon(r_4)| + |\epsilon(r_5)|\} + |\epsilon(r_6)| + |\epsilon(r_7)| \tag{19}$$

where $\epsilon(r_i)$ represents the relative error in $r_i$ arising from the computation by which $r_i$ was determined from the preceding partial results, $r_1, r_2, \cdots, r_{i-1}$, and $\epsilon(x)$ is the total relative computational error in $x$ when so computed. It is easy to keep $\epsilon(x)$ within any desired limits by suitably limiting each error term of (19). Since a computation accurate to $f$ significant figures involves a relative computational error not greater than $5 \times 10^{-f}$, any desired limits can then be set to each error term of (19) by a proper choice of the number of significant figures that should be carried in that step.

Unfortunately there seem to be no reasonably simple formulae for determining upper bounds of the relative computational errors that arise in the solution of simultaneous linear equations in more than two variables. This does not absolve the computer from the necessity of ensuring that his computational errors are suitably limited.

The method I have found most economical is to carry the solution of simultaneous linear equations to the capacity of the machine, and as each partial result $r_i$ is obtained, write it as

$$r_i(1 \pm \epsilon_i),$$

where $r_i$ is the value actually found and $\epsilon_i$ is a positive number representing the accumulation of uncertainty introduced by all preceding steps in the computation. At the end of the computation each of the unknowns is found in the form

$$x_j(1 \pm \epsilon_j), \tag{20}$$

where $x_j$ represents the value found and $\epsilon_j$ is the upper bound of the relative computational error in $x_j$.

A comparison of $\epsilon_j$ with the upper bound of the observational error $|\delta x_j|$ of equation (3) will then indicate whether the computation is adequate. If the comparison shows that the computation was inadequate, it will show in which steps the number of significant figures $f_i$ was too small, and by how much. The computer can recompute, carrying these steps to the requisite number of figures with the assurance that his recomputation will then be adequate. The comparison will further indicate in which steps if any the number of significant figures $f_i$ was larger than necessary.

When a computer has thus set suitable upper bounds to the relative computational error in the solution of a set of linear equations, he is in a position to plan solutions of future similar sets so as to perform his computations more economically and yet safely. This is especially true when the solution of simultaneous linear equations arises week after week in routine testing.

14. **Conclusions.** Summary rules have been published, purporting to be safe guides to computers in avoiding needless work, and ensuring that the computations are carried to a sufficient degree of accuracy. Many of them are useful guides for certain types of computation and for limited ranges of the numerical values entering into the computation, but none of those that I have seen can be used generally. The only safe rule, where the matter is of importance, is to calculate the maximum possible computational error that can enter in the particular sequence of computation followed, and make sure that it is kept within the necessary limits.

It is sometimes necessary to carry the intermediate steps of a computation to many significant figures beyond the significant figures given in the data, or kept in the result. The relative error of one of the unknowns may be very much smaller than the relative errors of the data from which it is computed, while the

relative error of another of the unknowns may be larger. The methods of ensuring that the computations are adequate are outlined in section 13.

For the best sequence to follow in the elimination of the unknowns, I shall pass along a suggestion of Dr. W. Edwards Deming which he gave in one of our discussions of this subject. I venture to pass it along, because it has worked in every special case that I have constructed in an attempt to prove that it does not hold generally. If ever the suggestion fails, the computer may change the sequence; but in any case he is obliged, as stated above, to calculate the maximum possible computational error that can enter into his calculations. Dr. Deming's suggestion is this: "To evaluate some but not all of the unknowns to the highest possible computational accuracy, retaining as few significant figures as possible in the intermediate steps, solve the equations by successive elimination, eliminating first and evaluating last the unknowns of greatest inherent relative accuracy."

15. **Summary.** Expressions are given for the maximum observational error in the unknowns of a system of simultaneous linear equations, in terms of the relative errors of the coefficients and absolute terms therein. In order to extract all the information possible from a system of linear equations representing observational results, it is not sufficient in general to assume that the relative errors in the unknowns are as large as the relative error in the determinant of the system. In many problems the computation of some of the unknowns must therefore be carried to more significant figures than are determinate in the determinant of the system. Methods are outlined for evaluating computational error in the solution of linear equations to ensure that the computations are adequate.

In conclusion I wish to express my thanks to Dr. W. Edwards Deming who has given much of his time to assist me in the preparation of this paper. He has made valuable suggestions on the material to be included and the general manner of presentation. In addition he has criticized the manuscript in detail and assisted in the final revision.

## REFERENCES

[1] Edward B. Roessler, *Science*, Vol. 84, (1936), pp. 289–290.
[2] Joseph Berkson, ibid., p. 437.
[3] P. J. Rulon, ibid., pp. 483–484.
[4] F. R. Moulton, ibid., pp. 574–575.
[5] W. Edwards Deming, *Science*, Vol. 85 (1937), pp. 451–454. *Least Squares* (Department of Agriculture Graduate School, 1938), pp. 105, 111, 121, 135.
[6] I. M. H. Etherington, *Edinb. Math. Soc. Proc.*, Vol. 3 (1932), Part 2, pp. 107–117.