NUMERICAL MATHEMATICAL ANALYSIS

CHAPTER I

THE ACCURACY OF APPROXIMATE CALCULATIONS

1. Introduction. Since applied mathematics comes down ultimately to numerical results, the worker in applied mathematics will encounter all kinds of numbers and all kinds of formulas. He must be able to use the numbers and evaluate the formulas so as to get the best possible result in any situation. What he learned about numerical calculation in his earlier study of arithmetic is inadequate for handling the numerical side of applied mathematics. For example, the numerical data used in solving the problems of everyday life are usually not exact, and the numbers expressing such data are therefore not exact. They are merely approximations, true to two, three, or more figures.

Not only are the data of practical problems usually approximate, but sometimes the methods and processes by which the desired result is to be found are also approximate. An approximate calculation is one which involves approximate data, approximate methods, or both.

It is therefore evident that the error in a computed result may be due to one or both of two sources: errors in the data and errors of calculation. Errors of the first type cannot be remedied, but those of the second type can usually be made as small as we please. Thus, when such a number as π is replaced by its approximate value in a computation, we can decrease the error due to the approximation by taking π to as many figures as desired, and similarly in most other cases. We shall therefore assume in this chapter that the calculations are always carried out in such a manner as to make the errors of calculation negligible.

Nearly all numerical calculations are in some way approximate, and the aim of the computer should be to obtain results consistent with the data with a minimum of labor. The object of the present chapter is to set forth some basic ideas and methods relating to approximate calculations and to give methods for estimating the accuracy of the results obtained.

2. Approximate Numbers and Significant Figures.

(a) *Approximate Numbers.* In the discussion of approximate computation, it is convenient to make a distinction between numbers which are absolutely exact and those which express approximate values. Such numbers as 2, 1/3, 100, etc. are exact numbers because there is no approximation or uncertainty associated with them. Although such numbers as π, $\sqrt{2}$, e, etc. are exact numbers, they cannot be expressed exactly by a finite number of digits. When expressed in digital form, they must be written as 3.1416, 1.4142, 2.7183, etc. Such numbers are therefore only approximations to the true values and in such cases are called approximate numbers. An approximate number is therefore defined as a number which is used as an approximation to an exact number and differs only slightly from the exact number for which it stands.*

(b) *Significant Figures.* A *significant figure* is any one of the digits $1, 2, 3, \cdots 9$; and 0 is a significant figure except when it is used to fix the decimal point or to fill the places of unknown or discarded digits. Thus, in the number 0.00263 the significant figures are 2, 6, 3; the zeros are used merely to fix the decimal point and are therefore not significant. In the number 3809, however, all the digits, including the zero, are significant figures. In a number like 46300 there is nothing in the number as written to show whether or not the zeros are significant figures. The ambiguity can be removed by writing the number in the powers-of-ten notation as 4.63×10^4, 4.630×10^4, or 4.6300×10^4, the number of significant figures being indicated by the factor at the left.

3. Rounding of Numbers.

If we attempt to divide 27 by 13.1, we get

$$27/13.1 = 2.061068702 \cdots,$$

a quotient which never terminates. In order to use such a number in a practical computation, we must cut it down to a manageable form, such as 2.06, or 2.061, or 2.06107, etc. This process of cutting off superfluous digits and retaining as many as desired is called rounding off.

To *round off* or simply *round* a number is to retain a certain number of digits, *counted from the left,* and drop the others. Thus, to round off π to three, four, five, and six figures, respectively, we have 3.14, 3.142, 3.1416, 3.14159. Numbers are rounded off so as to cause the *least possible error.* This is attained by rounding according to the following rule:

* Some readers may object to the term "approximate number" and insist that one should always say "approximate value" of a number. The shorter term, however, is less cumbrous, is perfectly definite as defined above, and reminds us by its very name that it stands for the approximate value of a number. It has been used in this sense by no less an authority than Jules Tannery in his *Leçons d'Arithmétique.*

To round off a number to n significant figures, discard all digits to the right of the nth place. If the discarded number is less than half a unit in the nth place, leave the nth digit unchanged; if the discarded number is greater than half a unit in the nth place, add 1 to the nth digit. If the discarded number is *exactly* half a unit in the nth place, leave the nth digit unaltered if it is an even number, but increase it by 1 if it is an odd number; in other words, round off so as to leave the nth digit an *even* number in such cases.

When a number has been rounded off according to the rule just stated, it is said to be *correct to* n *significant figures.*

The following numbers are rounded off correctly to four significant figures:

29.63243	becomes	29.63
81.9773	"	81.98
4.4995001	"	4.500
11.64489	"	11.64
48.365	"	48.36
67.495	"	67.50

When the above rule is followed consistently, the errors due to rounding are largely cancelled by one another.

Such is not the case, however, if the computer follows an old rule which is sometimes advocated. The old rule says that when a 5 is dropped the preceding digit should always be increased by 1. This is bad advice and is conducive to an accumulation of rounding errors and therefore to inaccuracy in computation. It should be obvious to any thinking person that when a 5 is cut off, the preceding digit should be increased by 1 in only *half* the cases and should be left unchanged in the other half. Since even and odd digits occur with equal frequency, on the average, the rule that the odd digits be increased by 1 when a 5 is dropped is logically sound.

The case where the number to be discarded is exactly half a unit in the nth place deserves further comment. From purely logical considerations the digit preceding the discarded $5000\cdots$ might just as well be left odd, but there is a practical aspect to the matter. Rounded numbers must often be divided by other numbers, and it is highly desirable from the standpoint of accuracy that the division be exact as often as possible. An even number is always divisible by 2, it may be divisible by other even numbers, and it may also be divisible by several odd numbers; whereas an odd number is not divisible by any even number and it may not be divisible by any odd number. Hence, in general, even numbers are exactly divisible

by many more numbers than are odd numbers, and therefore there will be fewer left-over errors in a computation when the rounded numbers are left even. The rule that the last digit be left even rather than odd is thus conducive to accuracy in computation.

In certain rare instances the rule for cutting off $50000\cdots$ should be modified. For example, if a 5 is to be cut off from two or more numbers in a column that is to be added, the preceding digit should be increased by 1 in *half* the cases and left unchanged in the other half, regardless of whether the preceding digit is even or odd. Other cases might arise where common sense should be the guide in making the errors neutralize one another.

4. Absolute, Relative, and Percentage Errors. The *absolute error* of a number, measurement, or calculation is the numerical difference between the true value of the quantity and its approximate value as given, or obtained by measurement or calculation. The *relative error* is the absolute error divided by the true value of the quantity. The *percentage error* is 100 times the relative error. For example, let Q represent the true value of some quantity. If ΔQ is the absolute error of an approximate value of Q, then

$$\Delta Q/Q = \text{relative error of the approximate quantity.}$$
$$100\Delta Q/Q = \text{percentage error of the approximate quantity.}$$

If a number is correct to n significant figures, it is evident that its absolute error can not be greater than half a unit in the nth place. For example, if the number 4.629 is correct to four figures, its absolute error is not greater than $0.001 \times \frac{1}{2} = 0.0005$.

Remark. It is to be noted that relative and percentage errors are independent of the unit of measurement, whereas absolute errors are expressed in terms of the unit used.

5. Relation between Relative Error and the Number of Significant Figures. The belief is widespread, even in scientific circles, that the accuracy of a measurement or of a computed result is indicated by the number of decimals required to express it. This belief is erroneous, for the accuracy of a result is indicated by the number of *significant figures* required to express it. The true index of the accuracy of a measurement or of a calculation is the relative error. For example, if the diameter of a 2-inch steel shaft is measured to the nearest thousandth of an inch, the result is less accurate than the measurement of a mile of railroad track to the nearest foot. For although the absolute errors in the two measure-

ments are 0.0005 inch and 6 inches, respectively, the relative errors are $0.0005/2 = 1/4000$ and $1/10{,}560$. Hence in the measurement of the shaft we make an error of one part in 4000, whereas in the case of the railroad we make an error of one part in 10,560. The latter measurement is clearly the more accurate, even though its absolute error is 12,000 times as great.

The relation between the relative error and the number of correct figures is given by the following fundamental theorem:

Theorem I. If the first significant figure of a number is k, *and the number is correct to* n *significant figures, then the relative error is less than* $1/(k \times 10^{n-1})$.

Before giving a literal proof of this theorem we shall first show that it holds for several numbers picked at random. Henceforth we shall denote absolute and relative errors of numbers by the symbols E_a and E_r, respectively.

Example 1. Let us suppose that the number 864.32 is correct to five significant figures. Then $k = 8$, $n = 5$, and $E_a \leqq 0.01 \times \frac{1}{2} = 0.005$. For the relative error we have

$$E_r \leqq \frac{0.005}{864.32 - 0.005} = \frac{5}{864320 - 5} = \frac{1}{2 \times 86432 - 1}$$
$$= \frac{1}{2(86432 - \frac{1}{2})} < \frac{1}{2 \times 8 \times 10^4} < \frac{1}{8 \times 10^4}.$$

Hence the theorem holds here.

Example 2. Next, let us consider the number 369,230. Assuming that the last digit (the zero) is written merely to fill the place of a discarded digit and is therefore not a significant figure, we have $k = 3$, $n = 5$, and $E_a \leqq 10 \times \frac{1}{2} = 5$. Then

$$E_r \leqq \frac{5}{369230 - 5} = \frac{1}{2 \times .36923 - 1} = \frac{1}{2(36923 - \frac{1}{2})}$$
$$< \frac{1}{2 \times 3 \times 10^4} < \frac{1}{3 \times 10^4}.$$

Example 3. Finally, suppose the number 0.0800 is correct to three significant figures. Then $k = 8$, $n = 3$, $E_a \leqq 0.0001 \times \frac{1}{2} = 0.00005$, and

$$E_r \leqq \frac{0.00005}{0.0800 - 0.00005} = \frac{5}{8000 - 5} = \frac{1}{1600 - 1}$$
$$= \frac{1}{2(800 - \frac{1}{2})} < \frac{1}{8 \times 10^2}.$$

It is to be noted that in this example the relative error is not certainly less than $1/(2k \times 10^{n-1})$, as was the case in Examples 1 and 2 above.

To prove the theorem generally, let

N = any number (exact value),
n = number of correct significant figures,
m = number of correct decimal places.

Three cases must be distinguished, namely $m < n$, $m = n$, and $m > n$.

Case 1. $m < n$. Here the number of digits in the integral part of N is $n - m$. Denoting the first significant figure of N by k, as before, we have

$$E_a \leqq 1/10^m \times \tfrac{1}{2}, \qquad N \geqq k \times 10^{n-m-1} - 1/10^m \times \tfrac{1}{2}.$$

Hence

$$E_r \leqq \frac{1/10^m \times \frac{1}{2}}{k \times 10^{n-m-1} - 1/10^m \times \frac{1}{2}} = \frac{10^{-m}}{2k \times 10^{n-1} \times 10^{-m} - 10^{-m}}$$

$$= \frac{1}{2k \times 10^{n-1} - 1} = \frac{1}{2(k \times 10^{n-1} - \frac{1}{2})}.$$

Remembering now that n is a *positive integer* and that k stands for any one of the digits from 1 to 9 inclusive, we readily see that $2k \times 10^{n-1} - 1 > k \times 10^{n-1}$ in all cases except $k = 1$ and $n = 1$. But this is the trivial case where $N = 1$, 0.01, etc.; that is, where N contains only one digit different from zero and this digit is 1—a case which would never occur in practice. Hence for all other cases we have $2k \times 10^{n-1} - 1 > k \times 10^{n-1}$, and therefore

$$E_r < \frac{1}{k \times 10^{n-1}}$$

Case 2. $m = n$. Here N is a decimal and k is the first decimal figure. We then have

$$E_a \leqq 1/10^m \times \tfrac{1}{2}, \qquad N \geqq k \times 10^{-1} - 1/10^m \times \tfrac{1}{2}.$$

$$\therefore E_r \leqq \frac{10^{-m} \times \frac{1}{2}}{k \times 10^{-1} - 10^{-m} \times \frac{1}{2}} = \frac{10^{-m}}{2k \times 10^{-1} - 10^{-m}} = \frac{1}{2k \times 10^{m-1} - 1}$$

$$= \frac{1}{2k \times 10^{n-1} - 1} < \frac{1}{k \times 10^{n-1}}.$$

Case 3. $m > n$. In this case k occupies the $(m - n + 1)$th decimal place and therefore

$$N \geqq k \times 10^{-(m-n+1)} - 1/10^m \times \tfrac{1}{2}, \qquad E_a \leqq 1/10^m \times \tfrac{1}{2}.$$

$$\therefore E_r \leqq \frac{10^{-m} \times \frac{1}{2}}{k \times 10^{-m} \times 10^{n-1} - 10^{-m} \times \frac{1}{2}} = \frac{10^{-m}}{2k \times 10^{-m} \times 10^{n-1} - 10^{-m}}$$

$$= \frac{1}{2k \times 10^{n-1} - 1} < \frac{1}{k \times 10^{n-1}}.$$

The theorem is therefore true in all cases.

Corollary 1. Except in the case of approximate numbers of the form $k(1.000 \cdots) \times 10^p$, in which k is the only digit different from zero, the relative error is less than $1/(2k \times 10^{n-1})$.

Corollary 2. If $k \geqq 5$ and the given approximate number is not of the form $k(1.000 \cdots) \times 10^p$, then $E_r < 1/10^n$; for in this case $2k \geqq 10$ and therefore $2k \times 10^{n-1} \geqq 10^n$.

To find the number of correct figures corresponding to a given relative error we can not take the converse of the theorem stated at the beginning of this article, for the converse theorem is not true. In proving the formula for the relative error we took the lower limit for N in order to obtain the upper limit for E_r. Thus, for the lower limit of N we took its first significant figure multiplied by a power of 10. In the converse problem of finding the number of correct figures corresponding to a given relative error we must find the upper limit of the absolute error E_a; and since $E_a = NE_r$, we should use the upper limit for N. This upper limit will be $k + 1$ times a power of 10, where k is the first significant figure in N. For example, if the approximate value of N is 6895, the lower limit to be used in finding the relative error is 6×10^3, whereas the upper limit to be used in finding the absolute error is 7×10^3.

To solve the converse problem we utilize Theorem II:

Theorem II. If the relative error in an approximate number is less than $1/[(k+1) \times 10^{n-1}]$, *the number is correct to* n *significant figures, or at least is in error by less than a unit in the* n*th significant figure.*

To prove this theorem let

N = the given number (exact value),
n = number of correct significant figures in N,
k = first significant figure in N,
p = number of digits in the integral part of N.

Then

$n - p$ = number of decimals in N,

and $N \leqq (k+1) \times 10^{p-1}$.

Let

$$E_r < \frac{1}{(k+1)\times 10^{n-1}}.$$

Then

$$E_a < (k+1)\times 10^{p-1} \times \frac{1}{(k+1)\times 10^{n-1}} = \frac{1}{10^{n-p}}.$$

Now $1/10^{n-p}$ is one unit in the $(n-p)$th decimal place, or in the nth significant figure. Hence the absolute error E_a is less than a unit in the nth significant figure.

If the given number is a pure decimal, let

p = number of zeros between the decimal point and first significant figure. Then $n+p$ = number of decimals in N, and

$$N \leqq \frac{(k+1)}{10^{p+1}}.$$

Hence if

$$E_r < \frac{1}{(k+1)\times 10^{n-1}},$$

we have

$$E_a < \frac{(k+1)}{10^{p+1}} \times \frac{1}{(k+1)\times 10^{n-1}} = \frac{1}{10^{n+p}}.$$

But $1/10^{n+p}$ is one unit in the $(n+p)$th decimal place, or in the nth significant figure. Hence the absolute error E_a is less than a unit in the nth significant figure.

Corollary 3. If $E_r < 1/[2(k+1)\times 10^{n-1}]$, then E_a is less than half a unit in the nth significant figure and the given number is correct to n significant figures in all cases.

Corollary 4. Since k may have any value from 1 to 9 inclusive, it is evident that $k+1$ may have any value from 2 to 10. Hence the upper and lower limits of the fraction $1/[2(k+1)\times 10^{n-1}]$ are $1/(4\times 10^{n-1})$ and $1/(2\times 10^n)$, respectively. We can therefore assert that

If the relative error of any number is not greater than $1/(2\times 10^n)$ *the number is certainly correct to* n *significant figures.*

Remark. The reader can readily see from the preceding discussion that *the absolute error* is connected with the number of *decimal places*, whereas the *relative error* is connected with the number of *significant figures.*

6. The General Formula for Errors. Let

$$N = f(u_1, u_2, u_3, \cdots u_n) \tag{1}$$

denote any function of several independent quantities $u_1, u_2, \cdots u_n$, which are subject to the errors $\Delta u_1, \Delta u_2, \cdots \Delta u_n$, respectively. These errors in the u's will cause an error ΔN in the function N, according to the relation

$$N + \Delta N = f(u_1 + \Delta u_1, u_2 + \Delta u_2, \cdots u_n + \Delta u_n). \tag{2}$$

To find an expression for ΔN we must expand the right-hand member of (2) by Taylor's theorem for a function of several variables. Hence we have

$$f(u_1 + \Delta u_1, u_2 + \Delta u_2, \cdots u_n + \Delta u_n) = f(u_1, u_2, \cdots u_n) + \Delta u_1 \frac{\partial f}{\partial u_1} + \Delta u_2 \frac{\partial f}{\partial u_2} + \cdots + \Delta u_n \frac{\partial f}{\partial u_n} + \tfrac{1}{2}\left[(\Delta u_1)^2 \frac{\partial^2 f}{\partial u_1^2} + \cdots + (\Delta u_n)^2 \frac{\partial^2 f}{\partial u_n^2} + 2\Delta u_1 \Delta u_2 \frac{\partial^2 f}{\partial u_1 \partial u_2} + \cdots\right] + \cdots.$$

Now since the errors $\Delta u_1, \Delta u_2, \cdots \Delta u_n$ are always relatively small,* we may neglect their squares, products, and higher powers and write

$$N + \Delta N = f(u_1, u_2, u_3, \cdots u_n) + \Delta u_1 \frac{\partial f}{\partial u_1} + \Delta u_2 \frac{\partial f}{\partial u_2} + \cdots + \Delta u_n \frac{\partial f}{\partial u_n} \tag{3}$$

Subtracting (1) from (3), we get

$$\Delta N = \frac{\partial f}{\partial u_1} \Delta u_1 + \frac{\partial f}{\partial u_2} \Delta u_2 + \cdots + \frac{\partial f}{\partial u_n} \Delta u_n,$$

or

$$\Delta N = \frac{\partial N}{\partial u_1} \Delta u_1 + \frac{\partial N}{\partial u_2} \Delta u_2 + \frac{\partial N}{\partial u_3} \Delta u_3 + \cdots + \frac{\partial N}{\partial u_n} \Delta u_n. \tag{6.1}$$

This is the general formula for computing the error of a function, and it includes all possible cases. It will be observed that the right-hand member of (6.1) is merely the total differential of the function N.

For the relative error of the function N we have

$$E_r = \frac{\Delta N}{N} = \frac{\partial N}{\partial u_1} \frac{\Delta u_1}{N} + \frac{\partial N}{\partial u_2} \frac{\Delta u_2}{N} + \cdots + \frac{\partial N}{\partial u_n} \frac{\Delta u_n}{N}. \tag{6.2}$$

When N is a function of the form

$$N = \frac{K a^m b^n c^p}{d^q e^r}, \tag{6.3}$$

* A quantity P is said to be relatively small in comparison with a second quantity Q when the ratio P/Q is small in comparison with unity. The squares and products of such small ratios ars negligible in most calculations.

then by (6. 2) the relative error is

$$E_r = \Delta N/N = m(\Delta a/a) + n(\Delta b/b) + p(\Delta c/c) - q(\Delta d/d) - r(\Delta e/e).$$

But since the errors $\Delta a, \cdots \Delta e$, etc. are just as likely to be negative as positive, we must take all the terms with the positive sign in order to be sure of the maximum error in the function N. Hence we write

(6. 4) $$E_r \leqq m\,|\,\Delta a/a\,| + n\,|\,\Delta b/b\,| + p\,|\,\Delta c/c\,| + q\,|\,\Delta d/d\,| + r\,|\,\Delta e/e\,|.$$

7. Application of the Error Formulas to the Fundamental Operations of Arithmetic and to Logarithms. We shall now apply the preceding results to the fundamental operations of arithmetic.

7*a*). *Addition.* Let

$$N = u_1 + u_2 + \cdots + u_n.$$

Then

(7. 1) $$\Delta N = E_a = \Delta u_1 + \Delta u_2 + \cdots + \Delta u_n.$$

The absolute error of a sum of approximate numbers is therefore equal to the algebraic sum of their absolute errors.

The proper way to add approximate numbers of different accuracies is shown in the two examples below.

Example 1. Find the sum of the approximate numbers 561.32, 491.6, 86.954, and 3.9462, each being correct to its last figure but no farther.

Solution. Since the second number is known only to the first decimal place, it would be useless and absurd to retain more than two decimals in any of the other numbers. Hence we round them off to two decimals, add the four numbers, and give the result to one decimal place, as shown below:

$$\begin{array}{r} 491.6 \\ 561.32 \\ 86.95 \\ 3.95 \\ \hline 1143.8 \end{array}$$

By retaining two decimals in the more accurate numbers we eliminate the errors inherent in these numbers and thus reduce the error of the sum to that of the least accurate number. The final result, however, is uncertain by one unit in its last figure.

Example 2. Find the sum of 36490, 994, 557.32, 29500, and 86939, assuming that the number 29500 is known to only three significant figures.

Solution. Since one of the numbers is known only to the nearest hundred, we round off the others to the nearest ten, add, and give the sum to hundreds, as shown below:

$$\begin{array}{r} 29500 \\ 86940 \\ 36490 \\ 990 \\ 560 \\ \hline 154500 \text{ or } 1.545 \times 10^5. \end{array}$$

The result is uncertain by one unit in the last significant figure.

In general, if we find the sum of m numbers each of which has been rounded off correctly to the same place, the error in the sum may be as great as $m/2$ units in the last significant figure.

7*b*). *Averages.* An important case in the addition of numbers must here be considered. Suppose we are to find the mean of several approximate numbers. Is this mean reliable to any more figures than are the numbers from which it was obtained? The answer is yes, but in order to see why let us consider the following concrete case.

The first column below contains the mantissas of ten consecutive logarithms taken from a six-place table. The second column contains these same mantissas rounded off to five decimals. The third column gives the errors due to rounding, expressed in units of the sixth decimal place.

N	N'	E
0.961421	0.96142	1
0.961469	0.96147	−1
0.961516	0.96152	−4
0.961563	0.96156	3
0.961611	0.96161	1
0.961658	0.96166	−2
0.961706	0.96171	−4
0.961753	0.96175	3
0.961801	0.96180	1
0.961848	0.96185	−2
Average, 0.9616346	Av., 0.961635	Sum, −4
= 0.961635		Av., −0.4

Here we have the relation

$$N = N' + E$$

for each of the numbers and therefore the further relations

$$\Sigma N = \Sigma N' + \Sigma E$$

and

$$\Sigma N/n = \Sigma N'/n + \Sigma E/n.$$

It will be noticed that the average of the rounded numbers is in error by only 0.4 of a unit in the sixth decimal place. We may therefore call it correct to six decimals, or to one more place than the rounded numbers.

The entries in all numerical tables and the results of all measurements are rounded numbers in which the error is not greater than half a unit in the last significant figure. These errors (due to rounding) are in general as likely to be positive as negative and hence their algebraic sum is never large. Usually it is less than a unit in the last figure.

The foregoing considerations justify the computer in retaining one more figure in the mean of a set of numbers than are given in the numbers themselves. But rarely should he retain the mean to more than one additional figure.

7c). *Subtraction.* Here

$$N = u_1 - u_2$$

and

$$\Delta N = E_a = \Delta u_1 - \Delta u_2. \tag{7.2}$$

Since the errors Δu_1 and Δu_2 may be either positive or negative, however, we must take the sum of the absolute values of the errors in order to get the maximum error. We then have the result that the absolute error of the difference of two approximate numbers may equal the *sum* of their absolute errors.

When one approximate number is to be subtracted from another, they must both be rounded off to the same place before subtracting. Thus, to subtract 46.365 from 779.8, assuming that each number is approximate and correct only to its last figure, we have

$$779.8 - 46.4 = 733.4.$$

It would be absurd to write $779.800 - 46.365 = 733.435$, because the last two figures in the larger number as here written are not zeros.

7d). *Loss of Significant Figures by Subtraction.*

The most serious error connected with the subtraction of approximate numbers arises from the subtraction of numbers which are nearly equal. Suppose, for example, that the numbers 64.395 and 63.994 are each correct

to five figures, but no more. Their difference, $64.395 - 63.994 = 0.401$, is correct to only *three* figures. Again, if the numbers 16950 and 16870 are each correct to only four significant figures, their difference $16950 - 16870 = 80$ is correct to only *one* significant figure, and even this figure may be in error by one unit.

Errors arising from the disappearance of the most important figures on the left, as in the two examples of the preceding paragraph, are of frequent occurrence and sometimes render the result of a computation worthless. They must be carefully guarded against and eliminated wherever possible.

The inaccuracy resulting from the loss of the most important significant figures in the subtraction of two nearly equal numbers can be lessened, and sometimes entirely avoided, in one of two ways:

1. By approximating each of the numbers with sufficient accuracy *before* subtraction, when this is possible. Thus, to find the difference $\sqrt{2.03} - \sqrt{2}$ correct to five significant figures, we take $\sqrt{2.03} = 1.424781$ and $\sqrt{2} = 1.414214$. Then $1.424781 - 1.414214 = 0.010567$. Note that a slide-rule computation is worthless in such a case as this.

This method is limited when the two given numbers are approximate and true to only a few digits.

2. By transforming the expression whose value is desired. Thus, to find the value of $1 - \cos x$ when x is small and no extended table is at hand, write $1 - \cos x = 2 \sin^2 (x/2)$ in some cases, and in other cases replace $\cos x$ by its Taylor expansion. Then

$$1 - \cos x = 1 - \left(1 - \frac{x^2}{2} + \frac{x^4}{4!} - \cdots\right) = \frac{x^2}{2} - \frac{x^4}{24} + \cdots$$

In finding the area of a circular segment having a small central angle, replace $\sin \theta$ by its Taylor expansion. Thus

$$\text{Area} = \frac{R^2}{2}(\theta - \sin\theta) = \frac{R^2}{2}\left[\theta - \left(\theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \cdots\right)\right]$$
$$= \frac{R^2}{2}\left(\frac{\theta^3}{6} - \frac{\theta^5}{120} + \cdots\right),$$

otherwise the area of a plainly visible segment might turn out to be zero when 4- or 5-place tables are used.

Sometimes in the evaluation of such an expression as $\sqrt{a} - \sqrt{b}$, where b is only slightly less than a, one or more significant figures can be saved by rationalizing the expression as the first step in the calculation. Thus,

$$\sqrt{a} - \sqrt{b} = \frac{a - b}{\sqrt{a} + \sqrt{b}}.$$

This method is of value only when fewer digits are lost by taking $a - b$ than by taking $\sqrt{a} - \sqrt{b}$.

The general solution of a certain type of ladder problem in elementary mechanics is

$$P = \frac{\frac{W}{2}\cot\theta - W\mu}{\mu - \left(\frac{l-c}{l}\right)\cot\theta}.$$

Here the terms in the numerator may be nearly equal for particular values of W, θ, and μ; and the terms of the denominator may also be nearly equal for certain values of μ, l, c, and θ. In such cases a slide-rule computation may be worthless.

In making a transformation to prevent loss of significant figures by subtraction, each problem must be treated individually. There is no known method or procedure that will fit all cases.

The loss of the leading significant figures in the subtraction of two nearly equal numbers is the greatest source of inaccuracy in most computations, and it forms the weakest links in a chain computation where it occurs. The computer must be on his guard against it at all times.

In general, if we desire the difference of two approximate numbers to n significant figures, and if it is known beforehand that the first m figures at the left will disappear by subtraction, we must start with $m + n$ significant figures in *each* of the given numbers.

7*e*). *Multiplication.* In this case

$$N = u_1 u_2 u_3 \cdots u_n.$$

Since this is of the form (6. 3), in which $m = n = \cdots = r = 1$, we have by (6. 4)

$$(7.3) \qquad E_r = \Delta N/N = \Delta u_1/u_1 + \Delta u_2/u_2 + \cdots + \Delta u_n/u_n.$$

The relative error of a product of n approximate numbers is therefore equal to the algebraic sum of the relative errors of the separate numbers.

The accuracy of a product should always be investigated by means of the relative error. The absolute error, if desired, can be found from the relation $E_a = E_r N$.

When it is desired to find the product of two or more approximate numbers of different accuracies, the more accurate numbers should be rounded off so as to contain *one more* significant figure than the least accurate factor, for by so doing we eliminate the error due to the more accurate factors and thus make the error of the product due solely to the

errors of the less accurate numbers. The final result should be given to as many significant figures as are contained in the least accurate factor, and no more. The proper method of procedure in such cases will be illustrated by examples later on.

7f). *Division.* Here we have

$$N = u_1/u_2 .$$

This is also of the form (6. 3), where the exponents are all unity. Hence by (6. 4)

$$E_r = \Delta u_1/u_1 + \Delta u_2/u_2 . \tag{7. 4}$$

The relative error of a quotient is therefore equal to the algebraic sum of the relative errors of divisor and dividend, but in order to get the maximum error one should take the arithmetical sum of the errors.

A simple formula for the absolute error of a quotient can be found directly, as follows:

Let ΔQ = absolute error of the quotient u_1/u_2. Then

$$\Delta Q = \frac{u_1 + \Delta u_1}{u_2 + \Delta u_2} - \frac{u_1}{u_2} = \frac{u_2\Delta u_1 - u_1\Delta u_2}{u_2(u_2 + \Delta u_2)} = \frac{u_1\left(\frac{\Delta u_1}{u_1} - \frac{\Delta u_2}{u_2}\right)}{u_2 + \Delta u_2}$$

Now let ω denote the greatest absolute value of either $\Delta u_1/u_1$ or $\Delta u_2/u_2$, and take the signs of Δu_1 and Δu_2 so as to get the greatest value of ΔQ. Then since $\Delta u_2/u_2 \leqq \omega$, we have $\Delta u_2 \leqq \omega u_2$; and therefore if u_1 and u_2 are *both* subject to errors of the *same order of magnitude* we have

$$\Delta Q \leqq \frac{u_1(\omega + \omega)}{u_2 - \omega u_2} = \frac{2u_1\omega}{u_2(1 - \omega)} .$$

If only u_1 *or* u_2 is subject to error and the other is free from error in comparison with it, then

$$\Delta Q \leqq \frac{u_1(\omega)}{u_2 - \omega u_2} = \frac{u_1\omega}{u_2(1 - \omega)} .$$

Finally, if ω is negligible in comparison with 1, we get

$$\Delta Q \leqq 2(u_1/u_2)\omega \tag{7. 5}$$

if u_1 and u_2 are both subject to errors of the same order of magnitude; and

$$\Delta Q \leqq (u_1/u_2)\omega \tag{7. 6}$$

if only u_1 *or* u_2 is subject to error.

As in the case of products, the accuracy of a quotient should always be investigated by means of the relative error, and all the statements made above in regard to products hold for quotients. In particular, if one of the numbers (divisor or dividend) is more accurate than the other, the more accurate number should be rounded off so as to contain *one more* significant figure than the less accurate one. The result should be given to as many significant figures as the less accurate number, and no more. The following examples will illustrate the proper methods of investigating the accuracy of products and quotients.

Example 1. Find the product of 349.1×863.4 and state how many figures of the result are trustworthy.

Solution. Assuming that each number is correct to four figures but no more, we have $\Delta u_1 \leqq 0.05$, $\Delta u_2 \leqq 0.05$. Hence

$$E_r \leqq \frac{0.05}{349.1} + \frac{0.05}{863.4} = 0.000143 + 0.000057 = 0.00020.$$

The product of the given numbers is 301413 to six figures. The absolute error of this product is

$$E_a = 301413 \times 0.00020 = 60, \text{ possibly.}$$

The true result therefore lies between 301473 and 301353, and the best we can do is to take the mean of these numbers to four significant figures, or

$$349.1 \times 863.4 = 301400 = 3.014 \times 10^5.$$

Even then there is some uncertainty about the last figure.

Theorem II of Art. 5 also tells us that the above result is uncertain in the fourth figure, but that the error in that figure is less than a unit.

Example 2. Find the number of correct figures in the quotient $56.3/\sqrt{5}$, assuming that the numerator is correct to its last figure but no farther.

Solution. Here we take $\sqrt{5} = 2.236$ so as to make the divisor free from error in comparison with the dividend. Then

$$E_r \leqq \frac{0.05}{56.3} < 0.0009;$$

and since $56.3/2.236 = 25.2$ we have

$$E_a < 25.2 \times 0.0009 < 0.023.$$

Since this error does not affect the third figure of the quotient, we take 25.2 as the correct result.

Note that formula (7. 6) also gives this result.

We could have seen at a glance, without any investigation, that the error of the quotient in this example would be less than 0.025; for the denominator is free from error and the possible error of 0.05 in the numerator is to be divided by 2.236, thereby making the error of the quotient less than half that amount.

Example 3. Find how many figures of the quotient $4.89\pi/6.7$ are trustworthy, assuming that the denominator is true to only two figures.

Solution. The only appreciable error to be considered here is the possible 0.05 in the denominator. The corresponding relative error is

$$E_r \leqq \frac{0.05}{6.7} < 0.0075.$$

The quotient to three figures is

$$\frac{4.89 \times 3.14}{6.7} = 2.29.$$

Hence the possible absolute error is $E_a \leqq 2.29 \times 0.0075 < 0.02$. Since the third figure of the quotient may be in error by nearly two units, we are not justified in calling the result anything but 2. 3, or

$$\frac{4.89\pi}{6.7} = 2.3.$$

Formula (7. 6) also gives this same result.

Example 4. Find the number of trustworthy figures in the quotient of 876.3/494.2, assuming that both numbers are approximate and true only to the number of digits given.

Solution. Here the largest relative error is

$$\omega = \frac{0.05}{494.2} = 0.000101,$$

and the quotient is

$$\frac{876.3}{494.2} = 1.7732.$$

Hence by (7. 5)

$$\Delta Q = 2(1.7732)(0.000101) = 0.000358.$$

Since this error affects the fourth decimal place but not the third, we take the quotient to be 1.773.

Note. The greatest and least values of the above quotient are

$$\frac{876.35}{494.15} = 1.7734 \quad \text{and} \quad \frac{876.25}{494.25} = 1.7729.$$

These values agree to four significant figures and both give 1.773.

7g). *Powers and Roots.* Here N has the form

$$N = u^m.$$

Hence by (6. 4)

$$E_r \leqq m(\Delta u/u).$$

For the pth power of a number we put $m = p$ and have

$$E_r \leqq p(\Delta u/u).$$

The relative error of the pth power of a number is thus p times the relative error of the given number.

For the rth root of a number we put $m = 1/r$ and get

$$E_r \leqq \frac{1}{r} \frac{\Delta u}{u}.$$

Hence the relative error of the rth root of an approximate number is only $1/r$th of the relative error of the given number.

Example. Find the number of trustworthy figures in $(0.3862)^4$, assuming that the number in parentheses is correct to its last figure but no farther.

Solution. Here the relative error of the given number is

$$E_r = \frac{0.00005}{0.3862} < 0.00013.$$

The relative error of the result is therefore less than 4×0.00013, or 0.00052.

The required number to five figures is $(0.3862)^4 = 0.022246$. Hence the absolute error of the result is $0.022246 \times 0.00052 = 0.000012$. Since this error affects the fourth significant figure of the result, the best we can do is to write

$$(0.3862)^4 = 0.02225$$

and say that the last figure is uncertain by one unit.

The relative error of the fourth root of 0.3862 is less than $\frac{1}{4}(0.00013) = 0.000032$, and since this fourth root is 0.78832 the absolute error of the result is about $0.78832 \times 0.000032 = 0.000026$. Hence the fourth root is 0.7883 correct to four figures.

7h). *Logarithms.* Here we have

$$N = \log_{10} u = 0.43429 \log_e u.$$

Hence

$$\Delta N = 0.43429(\Delta u/u),$$

or

$$\Delta N < \frac{1}{2} \frac{\Delta u}{u}$$

The absolute error in the common logarithm of a number is thus less than half the relative error of the given number.

An error in a logarithm may cause a disastrous error in the antilogarithm or corresponding number, for from the first formula for ΔN above we have

$$\Delta u = \frac{u \Delta N}{0.43429} = 2.3026 u \Delta N.$$

The error in the antilog may thus be many times the error in the logarithm. For this reason it is of the utmost importance that the logarithm of a result be as free from error as possible.

Example 1. Suppose $N = \log_{10} u = 3.49853$ and $\Delta N < 0.000005$, so that the given logarithm is correct to its last figure. Then $u = 3151.6$ and therefore

$$\Delta u = 2.3 \times 3151.6 \times 0.000005 = 0.036.$$

Since this error does not affect the fifth figure in u, the antilog is correct to five figures.

Example 2. Suppose $N = \log_{10} u = 2.96384$ and $\Delta N = 0.00001$. Then $u = 920.11$ and

$$\Delta u = 2.3 \times 920.11 \times 0.00001 = 0.021.$$

This error affects the fifth figure in u and makes it uncertain by two units.

Inasmuch as the logarithm of most results is obtained by the addition of other logarithms, it is evident that such a logarithm is likely to be in error by a unit in the last figure, due to the addition of rounded numbers. Hence the corresponding number may frequently be in error by one or two units in its last significant figure when the number of significant figures in the antilog is the same as the number of decimals in the logarithm.

Remarks. The reader should bear in mind the fact that the number of correct figures in the antilog corresponds to the number of correct

decimals in the logarithm. The integral part, or characteristic, of the logarithm plays no part in determining the accuracy of the antilog. This fact is at once evident from a consideration of the equation

$$\Delta u/u = 2.3\Delta N.$$

For inasmuch as the number of correct figures in the antilog u is measured by its relative error, and since this latter quantity depends only on the absolute error ΔN and not at all on the characteristic, it is plain that the accuracy of the antilog depends only on the number of correct decimals in the mantissa.

It is an easy matter to determine the number of correct figures in any antilog when the number of correct decimals in the mantissa is given. Suppose, for example, that we are using m-place log tables and that the possible error in the logarithm of a result is one unit in the last decimal place, as is usually the case. Then $\Delta N = 1/10^m$ and we have

$$\Delta u/u = \frac{2.3}{10^m} = \frac{2.3}{10 \times 10^{m-1}} = \frac{1}{4.34 \times 10^{m-1}} < \frac{1}{2 \times 10^{m-1}}.$$

Hence by Corollary 4, Art. 5, the antilog u is certainly correct to $m - 1$ significant figures.

The equation $\Delta u/u = 1/(4.34 \times 10^{m-1})$ shows that if the mantissa is in error by two units in its last figure the antilog is still correct to $m - 1$ significant figures, for in this case the relative error of the antilog is

$$\Delta u/u = \frac{1}{2.17 \times 10^{m-1}},$$

which is less than $1/(2 \times 10^{m-1})$. We are therefore justified in asserting that if the mantissa of a logarithm is not in error by more than two units in the last decimal place the antilog is certainly correct to $m - 1$ significant figures.

8. The Impossibility, in General, of Obtaining a Result More Accurate than the Data Used. The reader will have observed that in all the examples worked in the preceding pages no result has been more accurate than the numbers used in obtaining it. This, of course, is what we should have expected, but sometimes computers seem to try to get more figures in the result than are used in the data. When we apply Corollaries 1 and 4 of Art. 5 to the errors of products, quotients, powers, roots, logarithms, and antilogarithms, we find that in no case is the result true to more figures than are the numbers used in computing it. The results for these operations are as follows:

(a) *Products and Quotients.* If k_1 and k_2 are the first significant figures of two numbers which are each correct to n significant figures, and if neither number is of the form $k(1.000\cdots)\times 10^p$, then their product or quotient is correct to

$n-1$ significant figures if $k_1 \geqq 2$ and $k_2 \geqq 2$,

$n-2$ significant figures if either $k_1 = 1$ or $k_2 = 1$.

(b) *Powers and Roots.* If k is the first significant figure of a number which is correct to n significant figures, and if this number contains more than one digit different from zero, then its pth power is correct to

$n-1$ significant figures if $p \leqq k$,

$n-2$ significant figures if $p \leqq 10k$;

and its rth root is correct to

n significant figures if $rk \geqq 10$,

$n-1$ significant figures if $rk < 10$.

(c) *Logs and Antilogs.* If k is the first significant figure of a number which is correct to n significant figures, and if this number contains more than one digit different from zero, then for the absolute error in its common logarithm we have

$$E_a < \frac{1}{4k \times 10^{n-1}}.$$

If a logarithm (to the base 10) is not in error by more than two units in the mth decimal place, the antilog is certainly correct to $m-1$ significant figures.

To prove the foregoing results for the accuracy of products and quotients, let k_1 and k_2 represent the first significant figures of the given numbers. Then by Corollary 1 of Art. 5 the relative errors of the numbers are less than $1/(2k_1 \times 10^{n-1})$ and $1/(2k_2 \times 10^{n-1})$, respectively; and since the relative error of the product or quotient of two numbers may equal the sum of their relative errors, we have

Relative error of result

$$< \frac{1}{2k_1 \times 10^{n-1}} + \frac{1}{2k_2 \times 10^{n-1}} = \left(\frac{1}{k_1} + \frac{1}{k_2}\right)\frac{1}{2 \times 10^{n-1}}.$$

Now if $(1/k_1 + 1/k_2) \leqq 1$ we have $E_r < 1/(2 \times 10^{n-1})$, and the product or quotient is certainly correct to $n-1$ significant figures. But this quantity is not greater than 1 if $k_1 \geqq 2$ and $k_2 \geqq 2$. Hence in this case the result is correct to $n-1$ significant figures. If, however, either $k_1 = 1$

or $k_2 = 1$, the quantity $(1/k_1 + 1/k_2) > 1$ and therefore the relative error of the result may be greater than $1/(2 \times 10^{n-1})$. Hence the result may not be correct to $n - 1$ significant figures, but it is certainly correct to $n - 2$ figures.

To prove the above results for the accuracy of powers and roots let k represent the first significant figure of the given number. Then the relative error of this number is less than $1/(2k \times 10^{n-1})$. Hence the relative error of its pth power is less than

$$\frac{p}{2k \times 10^{n-1}} = \frac{p}{k} \quad \frac{1}{2 \times 10^{n-1}}$$

The result will therefore be correct to $n - 1$ significant figures if $(p/k) \leqq 1$, or $p \leqq k$, and to $n - 2$ significant figures if $p \leqq 10k$.

The error of the rth root is less than

$$\frac{1}{r} \quad \frac{1}{2k \times 10^{n-1}} = \frac{1}{rk} \quad \frac{1}{2 \times 10^{n-1}} = \frac{10}{rk} \quad \frac{1}{2 \times 10^{n}}$$

Hence the result will be correct to n significant figures if $rk \geqq 10$ and to $n - 1$ significant figures if $rk < 10$.

To prove the result for the error of the common logarithm we recall that $\Delta N < \frac{1}{2}(\Delta u/u)$, and since $\Delta u/u < 1/(2k \times 10^{n-1})$ we have

$$\Delta N < \frac{1}{4k \times 10^{n-1}} .$$

The proof for the accuracy of the antilog has already been given at the end of Art. 7.

Since the separate processes of multiplication, division, raising to powers, and extraction of roots can not give a result more accurate than the data used in obtaining it, no combination of these processes could be expected to give a more accurate result except by accident. Hence when only these processes are involved in a computation, the result should never be given to more significant figures than are contained in the least accurate of the factors used. Even then the last significant figure will usually be uncertain. In a computation involving several distinct steps, retain at the end of each step one more significant figure than is required in the final result.

While it is true in general that a computed result is not more accurate than the numbers used in obtaining it, an exception must be made in the cases of addition and subtraction. When only these processes are involved, the result may be much more accurate than one of the quantities added or subtracted. For example, the sum $3463 + \sqrt{3} = 3463 + 1.7 = 3464.7$ is correct to five significant figures (assuming 3463 to be an exact number)

even though one of the numbers used in obtaining it is correct to only two figures. A similar result would evidently follow in the case of subtraction.

9. Further Considerations on the Accuracy of a Computed Result. In commenting on formulas (7.1) and (7.3), it was stated that the absolute error of a sum is equal to the *algebraic* sum of the errors of the numbers added, and that the relative error of a product is equal to the *algebraic* sum of the relative errors of the factors. The word "algebraic" deserves emphasis in these cases because errors of measuremnt and errors due to rounding are compensating to a very great extent, so that in most cases the error in a computed result is not equal to the arithmetical sum of the errors of the numbers from which the result was obtained.

We saw in (7b) that the error in a sum was only a small fraction of the arithmetic sum of the separate errors. That the errors of the factors in a product are also compensating may be seen by considering the product of two *exact* numbers:

$$649.3 \times 675.8 = 438{,}796.94.$$

Now suppose we round off these numbers to 649 and 676. Their product is then $649 \times 676 = 438{,}724$. The actual error of this product is 72.94, and the relative error is

$$\frac{72.94}{438{,}796.94} = 0.000166.$$

The relative errors of the factors are $0.3/649.3 = 0.000462$ and $-0.2/675.8 = -0.000296$. The relative error of the product is thus *less* than the relative error of either factor and is actually equal to their *algebraic* sum. The product in this case is more accurate than either factor.

When a long computation is carried out in several steps and the intermediate results are properly rounded at the end of each step, there is no accumulation of rounding errors. If there were, long astronomical computations, such as those of eclipses and the orbits of comets, would be worthless. Time and experience have proved the correctness of such astronomical computations. In a chain computation the loss of significant figures by subtraction is the chief source of error.

Bad advice is sometimes given in regard to computation. In the addition of numbers of unequal accuracy, some writers advise that all the numbers first be rounded off to the number of decimal places given in the least accurate number. When this is done, the computer throws away definite information and replaces it with uncertainty. In adding a column of several numbers, the uncertainties might largely cancel one another, but this would not be the case with only a few numbers. The proper

method is to add the more accurate numbers separately and then round off their sum to the same decimal place as the least accurate number or numbers. In this way, the sum is as accurate as the least accurate of the numbers added.

Similar bad advice is given in the case of multiplication and division. When multiplying or dividing numbers of unequal accuracy, some writers advise that all numbers first be rounded off to the same number of significant figures as contained in the least accurate factor. To make all factors as rough as the roughest one is folly. There is no sense in throwing away perfectly definite information and replacing it with a question mark. The more accurate factors should be kept with *one more* significant figure than the least accurate factor. Then the result will usually be as accurate as the least accurate factor. The correct procedure in all ordinary computations can be stated in

A Sound and Safe Rule: When computing with rounded or approximate numbers of unequal accuracy, retain from the beginning *one more* significant figure in the more accurate numbers than are contained in the least accurate number. Then round off the result to the same number of significant figures as the least accurate number.

In the case of addition, retain in the more accurate numbers one more *decimal* digit than is contained in the least accurate number.

This rule follows from equations (7. 1), (7. 3), and (7. 4). By retaining one more digit in the more accurate numbers, we reduce to zero the errors of those terms and thus reduce the error of the final result.

In the case of subtraction or of addition of only *two* numbers, round of the more accurate number to the same number of decimal places as the less accurate one *before* subtracting or adding.

10. Accuracy in the Evaluation of a Formula or Complex Expression. The two fundamental problems under this head are the following:

(a) Given the errors of several independent quantities or approximate numbers, to find the error of any function of these quantities.

(b) To find the allowable errors in several independent quantities in order to obtain a prescribed degree of accuracy in any function of these quantities.

10a). *The Direct Problem.* The first of these problems is solved by replacing the given *approximate* numbers by the letters $a, b, c, \cdots$ or u_1, u_2, u_3, taking the partial derivatives of the function with respect to each of these letters, and then substituting in formula (6. 1) or (6. 2). An *exact* number, such as 2, 3, 10, etc., is *not* replaced by a letter before

taking the derivatives.* We shall now work some examples to show the method of procedure.

Example 1. Find the error in the evaluation of the fraction $\cos 7^\circ 10'/\log_{10} 242.7$, assuming that the angle may be in error by $1'$ and that the number 242.7 may be in error by a unit in its last figure.

Solution. Since this is a quotient of two functions, it is better to compute the relative error from the formula $E_r \leqq \Delta u_1/u_1 + \Delta u_2/u_2$ and then find the absolute error from the relation $E_a = NE_r$. Hence if we write

$$N = \frac{\cos 7^\circ 10'}{\log_{10} 242.7} = \frac{\cos x}{\log_{10} y} = u_1/u_2$$

we have

$$\Delta u_1 = \Delta \cos x = -\sin x \Delta x,$$

$$\Delta u_2 = \Delta \log_{10} y = 0.43429\,(\Delta y/y).$$

$$\therefore E_r \leqq \frac{\sin x}{\cos x}\,\Delta x + \frac{0.43429}{y \log y}\,\Delta y,$$

or

$$E_r \leqq \tan x \Delta x + \frac{0.435}{y \log y}\,\Delta y.$$

Now taking $x = 7^\circ 10'$, $\Delta x = 1' = 0.000291$ radian, $y = 242$, $\Delta y = 0.1$, and using a slide rule for the computation, we have

$$E_r < 0.126 \times 0.000291 + \frac{0.435 \times 0.1}{242 \times 2.38} = 0.00011.$$

Since $N = \cos 7^\circ 10'/\log 242.7 = 0.41599$, we have

$$E_a = 0.00011 \times 0.416 = 0.000046,$$

or $E_a < 0.00005$.

The value of the fraction is therefore between 0.41604 and 0.41594, and we take the mean of these numbers to four figures as the best value of the fraction, or

$$\underline{N = 0.4160.}$$

Example 2. The hypotenuse and a side of a right triangle are found by measurement to be 75 and 32, respectively. If the possible error in

* Adopted or accepted values of physical, chemical, and astronomical constants are to be treated as exact numbers, but results obtained by using these numbers as multipliers or divisors are not to be relied upon to more significant figures than are used in the constants themselves.

the hypotenuse is 0.2 and that in the side is 0.1, find the possible error in the computed angle A.

Solution. Lettering the triangle in the usual manner, we have

$$\sin A = 32/75 = a/c.$$

$$\therefore A = \sin^{-1}(a/c),$$

and

$$\Delta A = (\partial A/\partial a)\Delta a + (\partial A/\partial c)\Delta c.$$

Now

$$\partial A/\partial a = 1/\sqrt{c^2 - a^2},$$

$$\partial A/\partial c = -a/(c\sqrt{c^2 - a^2}).$$

Taking the numerical values of c and a in such a manner as to give the upper limits for $\partial A/\partial a$ and $\partial A/\partial c$, and remembering that $\Delta a = 0.1$, $\Delta c = 0.2$, we have

$$\Delta A < \frac{1}{\sqrt{(74.8)^2 - (32.1)^2}} \times 0.1 + \frac{32.1}{74.8\sqrt{(74.8)^2 - (32.1)^2}} \times 0.2 = 0.00275,$$

or

$$\Delta A < 0.0028 \text{ radian} = 9' 38''.$$

The possible error in A is therefore less than $9' 38''$.

10b). *The Inverse Problem.* We now turn our attention to the second fundamental problem mentioned at the beginning of this article: that of finding the allowable errors in $u_1, u_2, \cdots u_n$ when the function N is desired to a given degree of accuracy. This problem is mathematically indeterminate, since it would be possible to choose the errors Δu_1, Δu_2, etc. in a variety of ways so as to make ΔN less than any prescribed quantity. The problem is solved with the least labor by using what is known as the *principle of equal effects.** This principle assumes that all the partial differentials $(\partial N/\partial u_1)\Delta u_1$, $(\partial N/\partial u_2)\Delta u_2$, etc., contribute an equal amount in making up the total error ΔN. Under these conditions all the terms in the right-hand member of equation (6. 1) are equal to one another, so that

$$\Delta N = n\frac{\partial N}{\partial u}\Delta u_1 = n\frac{\partial N}{\partial u_2}\Delta u_2 = \cdots = n\frac{\partial N}{\partial u_n}\Delta u_n.$$

Hence

$$\Delta u_1 = \frac{\Delta N}{n\dfrac{\partial N}{\partial u_1}}, \quad \Delta u_2 = \frac{\Delta N}{n\dfrac{\partial N}{\partial u_2}}, \cdots \quad \Delta u_n = \frac{\Delta N}{n\dfrac{\partial N}{\partial u_n}}.$$

* See Palmer's *Theory of Measurements*, pp. 147-148.

Example 3. Two sides and the included angle of a trianglar city lot are approximately 96 ft., 87 ft., and 36°, respectively. Find the allowable errors in these quantities in order that the area of the lot may be determined to the nearest square foot.

Solution. Writing $b = 96$, $c = 87$, $A = 36°$, and denoting the area by u, we have

$$u = \tfrac{1}{2}bc \sin A = \tfrac{1}{2}(96 \times 87 \sin 36°) = 2455 \text{ sq. ft.}$$

Hence

$$\partial u/\partial b = \tfrac{1}{2}c \sin A, \quad \partial u/\partial c = \tfrac{1}{2}b \sin A, \quad \partial u/\partial A = \tfrac{1}{2}bc \cos A.$$

Substituting these quantities in (6. 2), we find

$$\Delta u/u = \Delta b/b + \Delta c/c + \Delta A/\tan A.$$

Now since the area is to be determined to the nearest square foot we must have $\Delta u < 0.5$; and by the principle of equal effects we must have

$$\frac{\Delta b}{b} = \frac{1}{3}\,\frac{\Delta u}{u} < \frac{0.5}{3 \times 2455} = \frac{1}{14730} < 0.000068.$$

Hence $\Delta b < 96 \times 0.000068 = 0.0065$ ft.

In like manner ,

$$\frac{\Delta c}{c} = \frac{1}{3}\,\frac{\Delta u}{u}, \text{ or } \Delta c < 87 \times 0.000068 = 0.0059 \text{ ft.};$$

and

$$\frac{1}{3}\,\frac{\Delta u}{u} = \frac{\Delta A}{\tan A}, \text{ or } \Delta A < \tan 36° \times 0.000068 = 0.000049 \text{ radian.}$$

Hence from a table for converting radians to degrees we find $\Delta A = 10''$.

It thus appears that in order to attain the desired accuracy in the area the sides must be measured to the nearest hundredth of a foot and the included angle to the nearest 20″ of arc.

This problem could also be solved by assuming that the possible errors in the measured sides might be 0.005 ft. and then computing the permissible error in the measured angle.

Example 4. The value of the function $6x^2(\log_{10} x - \sin 2y)$ is required correct to two decimal places. If the approximate values of x and y are 15.2 and 57°, respectively, find the permissible errors in these quantities.

Solution. Putting

$$u = 6x^2(\log_{10} x - \sin 2y) = 6(15.2)^2(\log_{10} 15.2 - \sin 114°)$$
$$= 371.9,$$

we have

$$\partial u/\partial x = 12x(\log_{10} x - \sin 2y) + 6x \times 0.43429 = 88.54,$$

$$\partial u/\partial y = -12x^2 \cos 2y = 1127.7.$$

Hence

$$\Delta u = (\partial u/\partial x)\Delta x + (\partial u/\partial y)\Delta y = 88.54\Delta x + 1127.7\Delta y.$$

In order that the required result be correct to two decimal places we must have $\Delta u < 0.005$. Then by the principle of equal effects we have

$$\Delta x = \frac{\Delta u}{2\dfrac{\partial u}{\partial x}} < \frac{0.005}{2 \times 88.54} = 0.000028,$$

$$\Delta y = \frac{\Delta u}{2\dfrac{\partial u}{\partial y}} < \frac{0.005}{2 \times 1127.7} = 0.0000022 \text{ rad.}$$

$$= 0''.45.$$

Since the permissible error in x is only 0.00003, it will be necessary to take x to seven significant figures in order to attain the required degree of accuracy in the result. The value of y can then be taken to the nearest second.

The reason why the permissible errors in x and y are so small in this example is that the factor $\log_{10} x - \sin 2y$ causes the loss of one significant figure by subtraction.

Remark. It is neither necessary nor desirable to investigate the accuracy of all proposed computations. But when we are in doubt about the possibility of attaining a certain degree of accuracy in the final result, we should make the necessary investigation. It usually suffices to carry all computations to one more figure than is desired in the final result and then round off the result to the desired number of figures, if the accuracy of the given independent quantities is such as to permit this.

11. Accuracy in the Determination of Arguments from a Tabulated Function. In many problems it is necessary to compute some function of an unknown quantity and then determine the quantity from tabulated values of the function. Examples of this kind are the determination of numbers from a table of logarithms, and angles from trigonometric tables. If the computed function happens to be affected with an error, the argument determined from this function is necessarily incorrect in some degree. The purpose of this article is to investigate the accuracy of the argument whose value is required.

In tables of single entry are tabulated functions of a single argument. Calling x the argument and y the tabulated function, we have

$$y = f(x).$$

From this we get the relation

$$\Delta y = f'(x)\Delta x, \text{ approximately,}$$

from which we have

$$\Delta x = \Delta y / f'(x). \tag{11.1}$$

This is the *fundamental equation* for computing the error in arguments taken from a table. Here Δy represents the error in the computed function whose values are tabulated, and Δx is the corresponding error in the argument. It will be noted that the magnitude of Δx depends upon three things: the error in the function, the nature of the function, and the magnitude of the argument itself. We shall now apply (11.1) to several functions whose values are tabulated.

1. Logarithms.

(*a*)
$$f(x) = \log_e x.$$
$$f'(x) = 1/x.$$

(1)
$$\therefore \Delta x = x\Delta y, \text{ from (11.1).}$$

(*b*)
$$f(x) = \log_{10} x.$$
$$f'(x) = M/x, \text{ where } M = 0.43429.$$
$$\therefore \Delta x = x\Delta y/M = 2.3026x\Delta y.$$

Hence

(2)
$$\Delta x < 2.31x\Delta y.$$

2. Trigonometric Functions.

(*a*)
$$f(x) = \sin x.$$
$$f'(x) = \cos x.$$

(3)
$$\therefore \Delta x = \Delta y/\cos x = \sec x\Delta y \text{ radians,}$$

or

(4)
$$(\Delta x)'' = 206264.8 \sec x\Delta y \text{ seconds.}$$

(*b*)
$$f(x) = \tan x.$$
$$f'(x) = \sec^2 x.$$

(5) $$\therefore \Delta x = \cos^2 x \Delta y \text{ radians,}$$

or

(6) $$(\Delta x)'' = 206264.8 \cos^2 x \Delta y \text{ seconds.}$$

(c) $$f(x) = \log_{10} \sin x.$$

$$f'(x) = M \frac{\cos x}{\sin x} = M \cot x.$$

(7) $$\therefore \Delta x = \frac{\Delta y}{M \cot x} = 2.3026 \tan x \Delta y \text{ radians,}$$

or

(8) $$(\Delta x)'' < 475000 \tan x \Delta y \text{ seconds.}$$

(d) $$f(x) = \log_{10} \tan x.$$

$$f'(x) = M \frac{\sec^2 x}{\tan x} = \frac{M}{\sin x \cos x} = \frac{2M}{\sin 2x}.$$

$$\therefore \Delta x = \frac{\sin 2x \Delta y}{2M} = 1.1513 \sin 2x \Delta y,$$

or

(9) $$\Delta x < 1.16 \sin 2x \Delta y \text{ radians;}$$

and

(10) $$(\Delta x)'' < 238000 \sin 2x \Delta y \text{ seconds.}$$

3. *Exponential Functions.*

$$f(x) = e^x.$$

$$f'(x) = e^x.$$

(11) $$\therefore \Delta x = \Delta y / e^x.$$

4. *Other Tabulated Functions.* By means of the fundamental equation (11. 1) we can compute the error in any argument when the derivative of the given function is given or can be easily found. In Jahnke and Emde's *Funktionentafeln*, for instance, are tabulated the derivatives of $\log \Gamma(x+1)$, the error function $\int_0^x e^{-x^2} dx$, the Weierstrass p-function, $p(u)$, and Legendre's polynomials $P_n(x)$. Hence by means of these tables we can determine the arugment and also its error.

Elliptic integrals are functions of two arguments. The error in each of these arguments can not be determined uniquely, but by using formula (6. 1) and assuming the principle of equal effects we can find definite formulas for the errors in the arguments. Thus, denoting an elliptic integral by I and the function of the arguments by $F(\theta, \phi)$, we have

$$I = F(\theta, \phi).$$

Hence

$$\Delta I = (\partial F/\partial\theta)\Delta\theta + (\partial F/\partial\phi)\Delta\phi.$$

By assuming that the two terms on the right-hand side are equal, we get

$$\Delta\theta = \frac{\Delta I}{2\dfrac{\partial F}{\partial\theta}}, \qquad \Delta\phi = \frac{\Delta I}{2\dfrac{\partial F}{\partial\phi}}.$$

Knowing the error ΔI of the integral, we can find from these formulas the corresponding errors in θ and ϕ.

Remarks. Comparison of formulas (3) and (5) shows that the error made in finding an angle from its tangent is always less than when finding it from its sine, because $\cos^2 x$ is less than $\sec x$. The latter may have any value from 1 to ∞, whereas the value of the former never exceeds 1.

Formulas (7) and (9) show still more clearly the advantage of determining an angle from its tangent. It is evident from (9) that the error in x can rarely exceed the error in y, since $\sin 2x$ can not exceed 1, but (7) shows that when the angle is determined from its log sine the error in x may be many times that in y.

Let us consider a numerical case. Suppose we are to find x from a 5-place table of log sines. Since all the tabular values are rounded numbers, the value of Δy may be as large as 0.000005, due to the inherent errors of the table itself. Taking $x = 60°$ and substituting in (7), we get

$$\begin{aligned}\Delta x &= 2.3026\sqrt{3} \times 0.000005 \\ &= 0.00002 \text{ radian, about,} \\ &= 4''.1.\end{aligned}$$

The unavoidable error may therefore be as great as 4 seconds if we find x from its log sine.

If, on the other hand, we find x from a table of log tangents we have from (9)

$$\Delta x < 1.16 \times \tfrac{1}{2}\sqrt{3} \times 0.000005 = 0.000005 \text{ rad.}$$
$$= 1''.$$

The error is thus only one-fourth as great as in the preceding case.

The foregoing formulas simply substantiate what has long been known by computers: that an angle can be determined more accurately from its tangent or cotangent than from its sine or cosine.

Note. The problem of determining the maximum possible error in a

result found by means of tables is rather involved. The reader will find a masterly treatment of this matter in J. Lüroth's *Vorlesungen über numerisches Rechen,* Leipzig, 1900.

However, the problem is of little practical importance, because the errors in such a computation rarely if ever combine so as to produce their maximum aggregate effect. They neutralize one another as the calculation proceeds.

12. The Accuracy of Series Approximations. It is frequently easier to find the numerical value of a function by expanding it into a power series and evaluating the first few terms than by any other method. In fact, this is sometimes the only possible method of computing it. The general method for expanding functions into power series is by means of Taylor's formula. The two standard forms of this formula are the following:

$$(1)\quad f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + \cdots + \frac{(x-a)^{n-1}}{(n-1)!}f^{(n-1)} + \frac{(x-a)^n}{n!}f^{(n)}[a+\theta(x-a)], \quad 0<\theta<1.$$

$$(2)\quad f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!}f''(x) + \cdots + \frac{h^{n-1}}{(n-1)!}f^{(n-1)}(x) + \frac{h^n}{n!}f^{(n)}(x+\theta h), \quad 0<\theta<1.$$

On putting $a = 0$ in (1) we get Maclaurin's formula:

$$(3)\quad f(x) = f(0) + xf'(0) + \frac{x^2}{2!}f''(0) + \cdots + \frac{x^{n-1}}{(n-1)!}f^{(n-1)}(0) + \frac{x^n}{n!}f^{(n)}(\theta x), \quad 0<\theta<1.$$

The last term in each of these three formulas is the *remainder after n terms.* This remainder term is the quantity in which we shall be interested in this article. The forms of the remainder given above are not the only ones, however. Another useful form will be given below.

12a). The Remainder Terms in Taylor's and Maclaurin's Series. Denoting by $R_n(x)$ the remainder after n terms in the Taylor and Maclaurin expansions, we have the following useful forms:

1. For Taylor's formula (1):

$$(a)\qquad R_n(x) = \frac{(x-a)^n}{n!}f^{(n)}[a+\theta(x-a)], \quad 0<\theta<1.$$

$$(b) \qquad R_n(x) = \frac{1}{(n-1)!}\int_0^{x-a} f^{(n)}(x-t)\,t^{n-1}dt.$$

2. For Taylor's formula (2):

$$(a) \qquad R_n(x) = \frac{h^n}{n!}f^{(n)}(x+\theta h), \quad 0<\theta<1.$$

$$(b) \qquad R_n(x) = \frac{1}{(n-1)!}\int_0^{h} f^{(n)}(x+h-t)\,t^{n-1}dt.$$

3. For Maclaurin's formula:

$$(a) \qquad R_n(x) = \frac{x^n}{n!}f^{(n)}(\theta x), \quad 0<\theta<1.$$

$$(b) \qquad R_n(x) = \frac{1}{(n-1)!}\int_0^{x} f^{(n)}(x-t)\,t^{n-1}dt.$$

It will be observed that the second form (the integral form) is perfectly definite and contains no uncertain factor θ. In using either form, however, it is necessary first to find the nth derivative of $f(x)$.

Since the integral form of $R_n(x)$ is not usually given in the textbooks on calculus, we shall show how to apply it to an example.

Example. Find the remainder after n terms in the expansion of $\log_e(x+h)$.

Solution. Here

$$f(x) = \log_e x,$$
$$\therefore\ f'(x) = 1/x,$$
$$f''(x) = -(1/x^2),$$
$$f'''(x) = 2/x^3,$$
$$f^{iv}(x) = -(6/x^4),$$
$$\cdots\cdots\cdots\cdots$$
$$f^{(n)}(x) = \frac{(-1)^{n-1}(n-1)!}{x^n}.$$

$$\therefore\ R_n(x) = (-1)^{n-1}\frac{(n-1)!}{(n-1)!}\int_0^{h}\frac{1}{(x+h-t)^n}\,t^{n-1}dt.$$

Now since t varies from 0 to h, the greatest value of $R_n(x)$ is obtained by putting $t=h$ in the integrand. We then have, omitting the factor $(-1)^{n-1}$, which is never greater than 1,

$$R_n(x) < \int_0^h \frac{t^{n-1}dt}{x^n} = \frac{1}{x^n}\int_0^h t^{n-1}dt = \frac{1}{x^n}\,\frac{h^n}{n} = \frac{1}{n}\left(\frac{h}{x}\right)^n.$$

Suppose $x = 1$, $h = 0.01$. Then $h/x = 0.01$. If, therefore, we wish to know how many terms in the expansion of $\log_e 1.01$ are necessary in order to get a result correct to seven decimal places we take $R_n \leqq 0.00000005$.

$$\therefore\ (1/n)\,(0.01)^n = 0.00000005.$$

It is evident by inspection that $n = 4$ will give a remainder much smaller than the allowable error. Hence we take four terms of the expansion of $\log_e (x + h)$.

The reader can easily verify that the first form of remainder gives the same result as that just found.

12b). *Alternating Series.* An alternating series is an infinite series in which the terms are alternately positive and negative. Such a series is convergent if (a) each term is numerically less than the preceding and (b) the limit of the nth term is zero when n becomes infinite.

Alternating series are of frequent occurrence in applied mathematics and are the most satisfactory for purposes of computation, because it is always an easy matter to determine the error of a computed result. The rule for determining the error is simply this:

In a convergent alternating series the error committed in stopping with any term is always less than the first term neglected.

Thus, since

$$\log_e (1 + x) = x - x^2/2 + x^3/3 - x^4/4 + x^5/5 - \cdots,$$

we have

$$\log_e (1.01) = 0.01 - \frac{(0.01)^2}{2} + \frac{(0.01)^3}{3} + R,$$

where $R < |(0.01)^4/4| = 0.0000000025$.

We therefore get a result true to eight decimal places by taking only three terms of the expansion.

12c). *Some Important Series and Their Remainder Terms.* Below are given some of the most useful series and their remainder terms, alternating series not being included because their remainder terms can be computed by the rule given above.

1. The Binomial Series.

$$(1 + x)^m = 1 + mx + \frac{m(m-1)}{2!}x^2 + \frac{m(m-1)(m-2)}{3!}x^3 + \cdots + \frac{m(m-1)(m-2)\cdots(m-n+2)}{(n-1)!}x^{n-1} + R_n,$$

where

$$(a) \quad R_n = \frac{m(m-1)(m-2)\cdots(m-n+1)}{n!} x^n (1+\theta x)^{m-n}, \quad 0 < \theta <$$

in all cases.

$$(b) \quad R_n < \left| \frac{m(m-1)(m-2)\cdots(m-n+1)}{n!} x^n \right| \text{ if } x > 0.$$

$$(c) \quad R_n < \left| \frac{m(m-1)(m-2)\cdots(m-n+1)}{n!} \frac{x^n}{(1+x)^{n-m}} \right|$$

if $x < 0$ and $n > m$.

$$(d) \qquad R_n < |x^n| (1+x)^m \text{ if } -1 < m < 0.$$

If m is a fraction, positive or negative, or a negative integer, the binomial expansion is valid only when $|x| < 1$. Also, except when m is a *positive integer*, a binomial such as $(a+b)^m$ must be written in the form

$$a^m\left(1+\frac{b}{a}\right)^m \text{ if } a > b, \text{ or } b^m\left(1+\frac{a}{b}\right)^m \text{ if } b > a,$$

before expanding it.

2. Exponential Series.

$$(a) \quad e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^{n-1}}{(n-1)!} + \frac{x^n}{n!} e^{\theta x}.$$

$$(b) \quad a^x = 1 + x\log a + \frac{(x\log a)^2}{2!} + \cdots + \frac{(x\log a)^{n-1}}{(n-1)!} + \frac{(x\log a)^n}{n!} a^{\theta x}.$$

If in (a) we put $x = 1$ we get the following series for computing e:

$$(c) \qquad e = 1 + 1 + \frac{1}{2} + \frac{1}{3!} + \frac{1}{4!} + \cdots + \frac{1}{(n-1)!} + \frac{e^\theta}{n!}.$$

Here

$$R_n = \frac{e^\theta}{n!}.$$

But since $e < 3$ and $\theta \leqq 1$, it is plain that

$$(d) \qquad R_n < \frac{3}{n!}.$$

A more definite formula for R_n can be found as follows:

Writing more than n terms of the series (c), we have

$$e = \left[1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{(n-1)!}\right] + \frac{1}{n!} + \frac{1}{(n+1)!} + \frac{1}{(n+2)!} +$$

where the remainder after n terms is

$$R_n = \frac{1}{n!} + \frac{1}{(n+1)!} + \frac{1}{(n+2)!} + \cdots$$
$$= \frac{1}{n!}\left(1 + \frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \cdots\right)$$

The quantity in parenthesis on the right is clearly less than the sum of the geometric series

$$1 + 1/n + 1/n^2 + 1/n^3 + \cdots,$$

the sum of which is

$$\frac{1}{1-1/n} = \frac{n}{n-1}.$$

Hence

$$(e) \qquad R_n < \frac{1}{n!}\,\frac{n}{n-1}, \text{ or } R_n < \frac{1}{(n-1)(n-1)!}.$$

By means of this formula (e) we can find the requisite number of terms in the expansion (c) to give the value of e correct to any desired number of decimal places. Thus, if we wished to find e correct to ten decimal places by means of the series (c) we would find n from the equation $1/(n-1)(n-1)! = 0.00000000005$. With the aid of a table of the reciprocals of the factorials we find that $n-1 = 13$, or $n = 14$. We should therefore take 14 terms of the series (c). We find in like manner that in order to compute e correct to 100 decimal places we should take 71 terms of the series (c).

3. Logarithmic Series.

$$(a) \quad \log_e(m+1) = \log_e m + 2\left[\frac{1}{2m+1} + \frac{1}{3(2m+1)^3} + \frac{1}{5(2m+1)^5}\right.$$
$$\left. + \cdots \quad \frac{1}{(2n-1)(2m+1)^{2n-1}}\right] + R_n,$$

To find an upper limit for R_n we have

$$R_n = 2\left[\frac{1}{(2n+1)(2m+1)^{2n+1}} + \frac{1}{(2n+3)(2m+1)^{2n+3}}\right.$$
$$\left. + \frac{1}{(2n+5)(2m+1)^{2n+5}} + \cdots\right]$$

Each term of the series in brackets, after the first, is less than the corresponding term of the series

$$\frac{1}{(2n+1)(2m+1)^{2n+1}} + \frac{1}{(2n+1)(2m+1)^{2n+3}} + \frac{1}{(2n+1)(2m+1)^{2n+5}} + \cdots,$$

or

$$\frac{1}{(2n+1)(2m+1)^{2n+1}}\left[1 + \frac{1}{(2m+1)^2} + \frac{1}{(2m+1)^4} + \cdots\right],$$

which is a geometric series with ratio $\frac{1}{(2m+1)^2}$ and sum $\frac{1}{1-\frac{1}{(2m+1)^2}}$, or $\frac{(2m+1)^2}{4m(m+1)}$. Hence

$$R_n < 2\left(\frac{1}{(2n+1)(2m+1)^{2n+1}}\right)\frac{(2m+1)^2}{4m(m+1)} = \frac{1}{2}\,\frac{1}{m(m+1)(2n+1)(2m+1)^{2n-1}}.$$

Therefore

$$(b) \qquad R_n < \frac{1}{2}\,\frac{1}{m(m+1)(2n+1)(2m+1)^{2n-1}}.$$

Example 1. To compute $\ln 2$ * by taking three terms of (a) we have, since $m = 1$, $n = 3$,

$$\ln 2 = 2\left[\frac{1}{3} + \frac{1}{3(3)^3} + \frac{1}{5(3)^5}\right] = 0.693004;$$

and by (b),

$$R_n < \frac{1}{2}\,\frac{1}{2(7)(3)^5} = 0.000147,$$

which affects the fourth decimal place. Since the true value of $\ln 2$ to eight decimal places is 0.69314718, the error in the value found above is 0.000143, which is less than 0.000147.

Example 2. To find $\ln 5$ correct to ten decimal places we have $m = 4$, $R_n = (1/2) \cdot (1/10^{10})$. Hence, by (b),

$$\frac{1}{2}\,\frac{1}{4 \times 5(2n+1)(9)^{2n-1}} = \frac{1}{2}\,\frac{1}{10^{10}},$$

or

$$(2n+1)(9)^{2n-1} = 5 \times 10^8 = 500{,}000{,}000.$$

We find by trial that n is about 4. 1, and that for $n = 5$ the logarithm will be correct to 11 decimal places.

* Frequently in this book we shall write ln for $\log_e$.

12d). Some nth Derivatives. In computing the remainder term in a series it is necessary to have the nth derivative of the given function. To facilitate the calculation of R_n we therefore give below a list of nth derivatives of some simple functions. The symbol D denotes differentiation with respect to x, or $D = d/dx$.

$$(a) \qquad D^n a^x = a^x (\log_e a)^n.$$

$$(b) \qquad D^n \sin x = \sin[x + n(\pi/2)].$$

$$(c) \qquad D^n \cos x = \cos[x + n(\pi/2)].$$

$$(d) \qquad D_n\left(\frac{1}{a+bx}\right) = \frac{(-1)^n n! b^n}{(a+bx)^{n+1}}.$$

$$(e) \qquad D^n\left(\frac{1}{\sqrt{a+bx}}\right) = \frac{(-1)^n 1\cdot 3\cdot 5\cdot\cdots(2n-1)}{2^n (a+bx)^{(2n+1)/2}} b^n.$$

$$(f) \qquad D^n \log_e(a+bx) = \frac{(-1)^n (n-1)! b^n}{(a+bx)^n}.$$

$$(g) \qquad D^n\left(\frac{\log_e x}{x}\right) = \frac{(-1)^n n!}{x^{n+1}}[\log_e x - (1 + \tfrac{1}{2} + \tfrac{1}{3} + \cdots + 1/n)].$$

$$(h) \qquad D^n \log_e (1+x^2) = \frac{(-1)^{n-1} 2(n-1)! \cos\left[n \sin^{-1}\left(\frac{1}{\sqrt{1+x^2}}\right)\right]}{(1+x^2)^{n/2}}.$$

$$(i) \qquad D^n \tan^{-1} x = \frac{(-1)^{n-1}(n-1)!}{(1+x^2)^{n/2}} \sin\left[n \sin^{-1}\left(\frac{1}{\sqrt{1+x^2}}\right)\right].$$

$$(j) \qquad D^n \frac{1}{1+x^2} = \frac{(-1)^n n!}{(1+x^2)^{(n+1)/2}} \sin\left[(n+1)\sin^{-1}\left(\frac{1}{\sqrt{1+x^2}}\right)\right].$$

$$(k) \qquad D^n\left(\frac{\alpha+\beta x}{(x-a)^2+b^2}\right) = \frac{(-1)^n n!}{b\rho^{n+1}}[\beta b \cos(n+1)\theta + (\alpha+\beta a)\sin(n+1)\theta],$$

where

$$\rho = \sqrt{(x-a)^2 + b^2}$$

$$\theta = \tan^{-1}\frac{b}{x-a}$$

For an extensive investigation of nth derivatives the reader is referred to Steffensen's *Interpolation*, pp. 231-241.

13. Errors in Determinants. When the elements in a determinant are inexact numbers, due to rounding or otherwise, the value of the determinant may be seriously affected by the loss of the most important significant figures in the expansion or evaluation process. The amount of such losses cannot be determined in advance. We can, however, determine the upper limit of the error in a determinant whose elements are subject to given possible errors. For purposes of illustration we consider a determinant of the third order.

Let

$$D = \begin{vmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{vmatrix}. \tag{1}$$

Now if the elements are subject to possible errors of unknown signs but of magnitudes Δx_1, Δy_1, etc., which are small in comparison with x_1, y_1, etc., then the value of D will be subject to the possible error ΔD such that

$$D + \Delta D = \begin{vmatrix} x_1 + \Delta x_1 & x_2 + \Delta x_2 & x_3 + \Delta x_3 \\ y_1 + \Delta y_1 & y_2 + \Delta y_2 & y_3 + \Delta y_3 \\ z_1 + \Delta z_1 & z_2 + \Delta z_2 & z_3 + \Delta z_3 \end{vmatrix}. \tag{2}$$

By the addition theorem of determinants the right member of (2) can be expressed as the sum of eight determinants, the first of which is the original determinant D. Each of three of the remaining determinants contains one column of error elements, each of three of the others contains two columns of error elements, and the remaining determinant has three columns of error elements. All determinants containing more than one column of error elements will be neglected, because, when expanded, the resulting terms will all contain second and third powers of the errors and will therefore be negligible in comparison with terms containing only the first powers of the errors. The value of ΔD is thus the sum of three determinants each containing a single column of error elements.

But those determinants are only the *differential* of D, and we therefore have

$$dD = \begin{vmatrix} dx_1 & x_2 & x_3 \\ dy_1 & y_2 & y_3 \\ dz_1 & z_2 & z_3 \end{vmatrix} + \begin{vmatrix} x_1 & dx_2 & x_3 \\ y_1 & dy_2 & y_3 \\ z_1 & dz_2 & z_3 \end{vmatrix} + \begin{vmatrix} x_1 & x_2 & dx_3 \\ y_1 & y_2 & dy_3 \\ z_1 & z_2 & dz_3 \end{vmatrix}, \tag{3}$$

or

$$(4)\qquad dD = (y_2z_3 - y_3z_2)\,dx_1 - (x_2z_3 - x_3z_2)\,dy_1 + (x_2y_3 - x_3y_2)\,dz_1$$
$$- (y_1z_3 - y_3z_1)\,dx_2 + (x_1z_3 - x_3z_1)\,dy_2 - (x_1y_3 - x_3y_1)\,dz_2$$
$$+ (y_1z_2 - y_2z_1)\,dx_3 - (x_1z_2 - x_2z_1)\,dy_3 + (x_1y_2 - x_2y_1)\,dz_3.$$

The maximum possible error would occur when the signs of the elements and the signs of the errors were such that all the eighteen terms in the right member of (4) were of the same sign—a very remote possibility.

Equation (4) shows that the error in a determinant composed of inexact elements may be anything from zero up to a number of considerable magnitude. It must be borne in mind, however, that the terms in (4) will largely cancel one another so that, in general, dD will not be large.

14. A Final Remark. The present chapter may appropriately close with the following lines from Alexander Pope:

> A little learning is a dangerous thing;
> Drink deep, or taste not the Pierian spring:
> There shallow draughts intoxicate the brain,
> And drinking largely sobers us again.

Pope was probably not thinking of approximate calculation when he wrote those lines, but no better advice could be given with respect to that subject. A smatter of knowledge of approximate calculation is worse than no knowledge at all. Fragmentary knowledge may lead to rough results that cannot be trusted. The author has seen students and teachers obtain far worse results from applying hazy ideas of the subject than if they had never heard of it. Their faulty work was due mostly to drastic rounding of numbers (at the beginning of a computation or at intermediate steps) or to dropping non-negligible terms in a series.

The essence of this chapter cannot be given in one or two recitations, nor in two or three. If the teacher has only two or three recitations to devote to it, he had better leave it out entirely.

EXERCISES I

1. Round off the following numbers correctly to four significant figures:

63.8543, 93487, 0.0063945, 83615, 363042, 0.090038, 53908.

2. A carpenter measures a 10-foot beam to the nearest eighth of an inch, and a machinist measures a $\frac{1}{2}$-inch bolt to the nearest thousandth of an inch.* Which measurement is the more accurate?

3. The following numbers are all approximate and are correct as far as their last digits only. Find their sum.

$$136.421,\ 28.3,\ 321,\ 68.243,\ 17.482.$$

4. Find the sum of the following approximate numbers, each being correct only to the number of significant figures given:

$$0.15625,\ 86.43,\ 191.6,\ 432.0 \times 10,\ 930.42.$$

5. The numbers 48.392 and 6852.4 are both approximate and true only to their last digits. Find their difference and state how many figures in the result are trustworthy.

6. Find the value of $\sqrt{10} - \pi$ correct to five significant figures.

7. The theoretical horsepower available in a stream is given by the formula

$$H.P. = \frac{whQ}{550},$$

where h = head in feet, Q = discharge in cubic feet per second, and w = weight of a cubic foot of water. The weight of fresh water varies from 62.3 to 62.5 lbs. per cubic foot, depending upon its temperature and purity.

If the measured values of Q and h are $Q = 463$ cu. ft./sec. and $h = 16.42$ ft., find the H. P. of the stream and indicate how many figures of the result are reliable.

8. The velocity of water flowing in long pipes is given by the formula

$$v = \sqrt{\frac{2ghd}{fl}} \text{ ft./sec.},$$

where
g = acceleration of gravity = 32.2 ft./sec.^{2}
h = head in feet,
d = diameter of pipe in feet,
l = length of pipe in feet,
f = coefficient of pipe friction.

* When a measurement is recorded to the nearest unit, the absolute error of the measurement is not more than *half* a unit.

In this problem the factor f is the most uncertain. It varies from 0.01 to 0.05 and is usually somewhere between 0.02 and 0.03. Assuming that f is within the limits 0.02 and 0.03 and taking

$$g = 32.2,$$
$$h = 112 \text{ feet},$$
$$d = \tfrac{1}{2} \text{ foot, (assumed exact)}$$
$$l = 1865 \text{ feet},$$

find v and indicate its reliability.

9. The velocity of water in a short pipe is given by the formula

$$v = \sqrt{\frac{2gh}{1.5 + fl/d}}$$

where g, h, f, l, and d have the same meanings as in the preceding example. Taking $l = 75$ feet and the other data the same as in Ex. 8, find v and indicate its reliability.

10. The acceleration of gravity at any point on the earth's surface is given by the formula

$$g = 32.1721 - 0.08211 \cos 2L - 0.000003H,$$

where H = altitude in feet above sea level, and L = latitude of the place. It thus appears that the value of g is not 32, nor 32.2, nor even 32.17.

Compute the kinetic energy of a 100-pound projectile moving with a velocity of 2000 feet per second by taking g equal to 32, 32.2, and 32.17 in succession and note the extent to which the results disagree after the first two or three figures.

11. How accurately should the length and time of vibration of a pendulum be measured in order that the computed value of g be correct to 0.05 per cent?

12. If in the formula

$$R = \frac{r^2}{2h} + \frac{h}{2}$$

the percentage error in R is not to exceed 0.3 per cent, find the allowable percentage errors in r and h when $r = 48$ mm. and $h = 56$ mm.

13. When the index of refraction of a liquid is determined by means of a refractometer, the index n is given by the formula

$$n = \sqrt{N^2 - \sin^2 \theta}.$$

If $N = 1.62200$ with an uncertainty of 0.00004 and $\theta = 38°$ approximately, find $\Delta\theta$ in order that n may be reliable to 0.02 per cent.

14. The area of the cross section of a rod is desired to 0.2 per cent. How accurately should the diameter be measured?

15. The approximate latitude of a place can be easily found by measuring the altitude h of Polaris at a known time t and using the formula

$$L = h - p \cos t,$$

where p = polar distance = 90° — declination.

Treating p as a constant and equal to 1°07′30″, and taking $h = 41°25'$, $t = 0^h38^m42^s$, find the error in L due to errors of 1 in h and 5^s in t.

16. In the preceding example find the allowable errors in h and t in order that the error in L shall not exceed 1′, using the same values of p, t, and h as before.

17. The distance between any two points P_1 and P_2 on the earth's surface is given by the formula

$$\cos D = \sin L_1 \sin L_2 + \cos L_1 \cos L_2 \cos(\lambda_1 - \lambda_2),$$

where L_1, L_2 and λ_1, λ_2 denote the respective latitudes and longitudes of the two places. Find the allowable errors in L_1, L_2, λ_1, λ_2 in order that the error in D shall not exceed 1′ (a geographical mile), taking

$$L_1 = 36°10' N, \quad L_2 = 58°43' N, \quad \lambda_1 = 82°15' W, \quad \lambda_2 = 125°42' W.$$

18. The fundamental equations of practical astronomy are:

$$\sin h = \sin \delta \sin L + \cos \delta \cos L \cos t, \tag{1}$$

$$\cos h \cos A = -\sin \delta \cos L + \cos \delta \sin L \cos t, \tag{2}$$

$$\cos h \sin A = \cos \delta \sin t, \tag{3}$$

where δ denotes declination, t hour angle, h altitude, and A azimuth of a celestial body and L denotes the latitude of a place on the earth. The declination δ is always accurately known and may therefore be considered free from error.

Differentiating (1) by considering δ constant and h, L, t as variables, we have

$$\cos h \, dh = \sin \delta \cos L \, dL - \cos \delta \sin L \cos t \, dL - \cos \delta \cos L \sin t \, dt.$$

Replacing $\cos \delta \sin L \cos t$ and $\cos \delta \sin t$ on the right by their values from (2) and (3), respectively, we get

$$dh = -(\cos A \, dL + \sin A \cos L \, dt).$$

Solving for dL,

$$dL = -(\sec A\, dh + \tan A \cos L\, dt). \tag{4}$$

This equation shows that the numerical value of dL is least when A is near 0° or 180°, that is, when the body is near the *meridian*. If A should be near 90°, that is, if the body should be near the prime vertical, the error in L might be enormous. Hence when determining latitude the observed body should be as near the meridian as possible.

Using equation (4), compute dL when $dh = 1'$, $dt = 10^s$, $L = 40°$, $A = 10°$, and $A = 80°$.

19. Using the formula $dL = -(\sec A dh + \tan A \cos L dt)$, find the allowable errors in t and h in order that the error in L may not exceed $1'$ when $L = 40°$ and (a) $A = 10°$ and (b) $A = 75°$.

20. From the relation

$$\cos h\, dh = (\sin \delta \cos L - \cos \delta \sin L \cos t)\, dL - \cos \delta \cos L \sin t\, dt$$

we find by means of (2) and (3) of Ex. 18

$$dt = -\frac{dh + \cos A\, dL}{\sin A \cos L}$$

This equation shows that dt is least numerically when A is near 90°, that is, when the observed body is near the prime vertical; it also shows that when the body is on or near the prime vertical an error in the assumed latitude has practically no effect on the error in t.

Compute dt when $dh = 1'$, $dL = 5'$, $L = 40°$, $A = 10°$, and $A = 80°$.

21. Using the formula for dt in the preceding example, find the allowable errors in L and h in order that dt may not exceed 3^s, taking $L = 40°$, $A = 10°$, and $A = 80°$.

22. Using the formula of Ex. 20, take $dt = 3^s$, $dh = 1'$, and find dL for $A = 10°$ and $A = 80°$.

23. In the equation

$$x = a \sin (kt + \alpha)$$

suppose a, k and α are subject to the errors Δa, Δk, $\Delta \alpha$, respectively. Compute Δx and see which of the errors Δa, Δk, $\Delta \alpha$ is the most potent in causing an error in x.

24. Find the value of

$$I = \int_0^{0.8} (\sin x/x)\, dx$$

correct to five decimal places.

25. Compute the value of the integral

$$I = \int_0^{\pi/2} \sqrt{1 - 0.162 \sin^2 \phi}\, d\phi$$

correct to five significant figures by first expanding the integrand by the binomial theorem and then integrating the result term by term.

26. In the formula

$$\cos k = \frac{\sin \frac{1}{2}(h_1 - h_3)\cos \frac{1}{2}(h_1 + h_3)}{\sin \frac{1}{2}(t_1 - t_3)\cos h_2}$$

k denotes an angle; h_1, h_2, h_3, t_1, t_3 are all positive; $(h_1 - h_3)$ and $(t_1 - t_3)$ are small quantities; and $(h_1 - h_3)$ is small in comparison with h_2. Find the maximum error in k due to errors in t and h, assuming that

$$|dh_1| = |dh_2| = |dh_3| \text{ and } |dt_1| = |dt_3|.$$

27. Using the result found in Ex. 26, find the maximum value of dk when $dt = 0^s.05$, $dh = 0'.05$, $h_1 = 40°$, $h_2 = 40° 15'$, $h_3 = 40° 30'$, $t_1 - t_3 = -4^m$.