WILEY PUBLICATIONS
IN STATISTICS

Walter A. Shewhart, Editor

Mathematical Statistics

DWYER—Linear Computations.

FISHER—Contributions to Mathematical Statistics.

WALD—Statistical Decision Functions.

FELLER—An Introduction to Probability Theory and Its Applications, Volume One.

WALD—Sequential Analysis.

HOEL—Introduction to Mathematical Statistics.

Applied Statistics

MUDGETT—Index Numbers.

TIPPETT—Technological Applications of Statistics.

DEMING—Some Theory of Sampling.

COCHRAN and COX—Experimental Designs.

RICE—Control Charts.

DODGE and ROMIG—Sampling Inspection Tables.

Related Books of Interest to Statisticians

HAUSER and LEONARD—Government Statistics for Business Use.

Linear Computations

PAUL S. DWYER
Professor of Mathematics
University of Michigan

New York · John Wiley & Sons, Inc.
London · Chapman & Hall, Limited

PRINTED IN THE UNITED STATES OF AMERICA

Preface

THIS BOOK IS WRITTEN FOR THE PURPOSE OF AIDING THE MANY workers in a variety of fields who have the general problem of finding numerical solutions for sets of simultaneous linear equations. Though many arrive at this mathematical problem through least squares, correlation, regression, or other statistical studies, some arrive at the problem in non-statistical ways. For this reason I have used a general mathematical presentation rather than one designed more specifically for statistical problems. The reader who is primarily interested in statistical applications should not have much difficulty in translating the mathematical results to appropriate statistical results. Chapter 18 is designed to assist him in this process of translation.

Much of this material is simplified with the use of matrices. In many cases the matrix proofs of important results are very concise, and, frequently, the matrix results describe the computational methods adequately. However, many of the workers who need these methods are not familiar with matrices; indeed, the basic computational methods can be presented to those who know the basic facts of high school algebra. It is the purpose of this book first to describe the theorems and methods in terms of elementary algebra and then to develop the subject by including introductory material on determinants (in Chapter 9) and on matrices (in Chapter 12). More powerful expositions are possible, therefore, in the later chapters. It cannot be overemphasized, however, that a real understanding of the methods involved can be obtained only with a direct application of the methods to numerical problems. I have provided many illustrative problems, throughout the book, to assist the reader in translating the mathematical results to concise calculational methods. Many of the illustrations chosen have been selected from the illustrations of previous writers so as to make possible a direct comparison between the old and the new techniques.

Though I have organized the material in such a way that the book may properly serve as a textbook for a course on linear computations, or as a textbook for individual study, it is also arranged to serve as a useful reference for the many workers in applied fields who seek to apply improved techniques to specific problems. A rather extensive

the earlier methods. Indicate the size of the error resulting from the application of (1.4.5) and show that the error is approximately $\frac{\sqrt{501}}{8(501)^2}$ indicated by (1.4.5). (This calls for a ten-place machine. Translate the problem to a six-decimal-place problem if you have an eight-place machine.)

8. The value of $\sqrt{5.01}$ is $\frac{1}{10}\sqrt{501}$ and so can be obtained from the result of the last problem by moving a decimal position. The value of $\sqrt{50.1}$ is $\frac{1}{\sqrt{10}}\sqrt{501}$. It can be obtained by dividing the results of exercise 7 by $\sqrt{10} = 3.162277660$. It can also be obtained by using the divisor $2\sqrt{50.0} = 14.14213562$ and (1.4.5). Compare the results.

9. Find $\sqrt[3]{28}$ by successive approximations. Use 3 as the first approximation and form $28 \div 3^2 = 3\frac{1}{9}$. Use the average of 3, 3, and $3\frac{1}{9}$ as the next approximation. Continue until agreement is reached to five decimal places.

10. Derive formulas corresponding to (1.4.3) and (1.4.5) when the cube root of N is desired. In the first case divide by $(\sqrt[3]{N} + \epsilon)^2$ and in the second case divide by $3(N + \epsilon)^{2/3}$.

11. Find $\sqrt[3]{28}$ to four decimal places by the method indicated in exercise 10.

12. Derive formulas corresponding to (1.4.3) and (1.4.5) when the rth root of N is desired. In the first case divide by N by $(\sqrt[r]{N} + \epsilon)^{r-1}$ and in the second case divide $N + (r-1)(N + \epsilon)$ by $r(N + \epsilon)^{(r-1)/r}$. Obtain (1.4.3) and (1.4.5) when $r = 2$ and the results of exercise 10 when $r = 3$.

CHAPTER 2

Computation with Approximate Numbers

2.1. Introduction. The effective use of any digital system or device requires that each number to be used in calculation shall be expressible as a digital number. The very nature of measurement also necessitates the use of approximate numbers. Although counting by integers results in exact numbers, most measurements, whether direct or indirect, result eventually in comparisons with some sort of scale. Numbers resulting from these comparisons are, in general, approximate rather than exact. Although the length of a line can be determined to the nearest inch, it cannot be determined exactly. Even if the line were exactly ten inches in length, there is no way in which we could ascertain that fact.

2.2 Approximate numbers. The limitations of digital systems of calculation and the very origin of the quantities to be used as bases of calculation, then, force us to make use of approximate numbers. An *approximate number*, or more precisely the *approximate value of a number*, is some number that differs from the true value by some amount, presumably small. If x represents the true number and x' the approximate number, then the error ϵ is given by

$$\epsilon = \epsilon(x) = x - x' = \Delta x = dx, \tag{1}$$

where ϵ is positive or negative according as $x > x'$ or $x < x'$. The error, ϵ, is not usually known exactly, but is specified to be less in absolute value than some quantity η. If the error and the approximation were known, it would be possible to solve for x in (1). This would give the exact value of x, and the use of approximate numbers would not be necessary. Approximate numbers are especially useful when the condition

$$|\epsilon| \leq \eta \tag{2}$$

is satisfied. This is the general situation resulting when measurements are made to the nearest unit. Thus η is 0.5 inch when measurements

are made to the nearest inch, η is 0.0005 inch when measurements are made to the nearest 0.001 inch, etc.

Approximate numbers with unspecified but limited errors may be indicated in different ways. The number x', if accompanied by the absolute value of the maximum possible error, enables us to find the range within which the true number lies. Thus if $x' = 112$ and $\eta = 4$, the true number satisfies the relation

$$108 \leq x \leq 116.$$

The approximate number may be indicated either by the range 108 to 116 or by x' with the greatest possible error, 112 ± 4. A dual number such as $\begin{bmatrix} 116 \\ 108 \end{bmatrix}$, where the upper entry is the highest possible value of the number and the lower entry is the lowest possible value of the number, may be used. The term *range number* is used here to indicate an approximate number when expressed in this form. The algebraic representation of an approximate number in range form is then $\begin{bmatrix} x_H \\ x_L \end{bmatrix}$, where x_H is the highest possible value of x and x_L is the lowest. The two recorded values are the *components of the range number*.

The form $x' \pm \eta$, where η is the absolute value of the largest possible error, may also be used to represent an approximate number with unspecified but limited error. Since this form features an approximation to the number accompanied by a statement of the largest possible error, we may refer to numbers of this form as *approximation-error numbers*. Also a condensed notation may be used in which the maximum possible error is inserted in parentheses after the x'. The decimal point may be disregarded in writing the error if it is understood that the error term is expressed in the unit of the last figure of the x'. Thus 1.12 ± 0.04 appears as 1.12(4) and 0.00132 ± 0.00017 may be written compactly as 0.00132(17). This convention also tends to avoid confusion with customary probable error and standard error notations.

No matter whether we use range numbers or approximation-error numbers, it is important to note that an approximate number represents a range within which the true value of the number is located.

The reader who understands these two forms of approximate numbers will be able to change at once from approximation-error numbers to range numbers and vice versa. Thus

$$(3) \qquad x_H = x' + \eta \quad \text{and} \quad x_L = x' - \eta$$

and

$$(4) \qquad x' = \tfrac{1}{2}(x_H + x_L) \quad \text{and} \quad \eta = \tfrac{1}{2}(x_H - x_L).$$

For example,

$$1643(17) = \begin{bmatrix} 1660 \\ 1626 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1.43 \\ 1.16 \end{bmatrix} = 1.295(135).$$

2.3 Significant figures and significant numbers. Both approximation-error numbers and range numbers are dual numbers. The recording of an approximation-error number can be accomplished by a single number if an agreement is made as to the maximum size of the error permissible. It is conventional to record the result of a measurement (or the result of an approximation to an irrational number) by a digital number, so that the recorded number is correct to the last recorded digit, that is, the error is at most one-half unit in the last recorded place. In this case it is not necessary to record the η in the error number since it is by agreement equal to one-half unit. Thus the recorded number 6.738 implies the approximation-error number 6.738(½) and might be written as the approximation-error number 6.7380(5). In the notation of range numbers this number would be represented by $\begin{bmatrix} 6.7385 \\ 6.7375 \end{bmatrix}$.

The digits used in this method of recording approximate numbers, neglecting the zeros necessary to indicate positive or negative powers of ten, are known as *significant figures* or *significant digits*. Thus the number 6.738 mentioned above has four significant digits.

Approximate numbers that are expressed in terms of significant figures might be called *significant numbers*. A significant number may be viewed as an approximation-error number in which the maximum error is one-half unit in the last decimal position.

There is some ambiguity about the number of significant digits in a significant number such as 30720. Is the last cipher a significant digit or does it merely indicate a power of ten? This ambiguity should be removed by the person who introduces the number or, preferably (see section 2.5), a notation should be adopted that resolves the ambiguity.

The process of replacing a number, exact or approximate, by a significant number with a smaller number of significant figures is known as rounding off.* Thus 3.1416, the five-figure approximation to π, rounds off successively to 3.142 and 3.14. It is conventional to round off to the

* It is conventional to use the symbol for equals, rather than the symbol for approximation, in rounding off. Thus it is accepted practice to write $\pi = 3.1416$ and not necessarily $\pi \cong 3.1416$ or $\pi \doteq 3.1416$.

even digit when the number to be rounded off is exactly half way between two successive digits.*

2.4 Limitations of significant numbers. In view of the simplicity of significant numbers, it is not surprising that these, rather than range numbers or approximation-error numbers, have been used extensively in computational work. However, significant numbers are far from ideal as a means of expressing the results of fundamental operations with approximate numbers.

The limitations of significant numbers begin to be apparent when we attempt to transform range numbers and approximation-error numbers to significant numbers. The transformations by which range numbers are written as equivalent approximation-error numbers, and by which approximation-error numbers are written as equivalent range numbers, are shown in (2.2.3) and (2.2.4). It is impossible to transform these numbers to equivalent significant numbers.

A significant number is a special case of an approximation-error number with the range restricted to a unit of the last digital position so that significant numbers constitute a rather restricted subclass of all approximate digital numbers, any of which may be stated in range or approximation-error form. It is possible to transform significant numbers to equivalent approximation-error numbers and range numbers, but it is impossible, in general, to transform approximation-error numbers and range numbers to the subclass of significant numbers. Thus

$$1.196 = 1.196(\tfrac{1}{2}) = 1.1960(5) = \begin{bmatrix} 1.1965 \\ 1.1955 \end{bmatrix} \tag{1}$$

but

$$\begin{bmatrix} 1.24 \\ 1.14 \end{bmatrix} = 1.19(5)$$

cannot be expressed as an equivalent significant number. It certainly cannot be expressed as the significant number 1.19 because it represents the range number $\begin{bmatrix} 1.195 \\ 1.185 \end{bmatrix}$, nor as the significant number 1.2 because it represents the number $\begin{bmatrix} 1.25 \\ 1.15 \end{bmatrix}$. These two numbers have a range of

* It is helpful to place a dash above a final 5 that results from rounding off a number whose digit in this position is less than 5. Thus 2.76147 should appear as $2.761\bar{5}$ when rounded to 5 significant figures. This number when rounded to four figures appears as 2.761, which is in error by less than one-half unit in the last place, while the rounding off of 2.7615 yields 2.762, and this is in error by more than one-half unit in the last place.

the same length, and nine-tenths of the range is common to the two numbers, but they are not the same numbers. The number $\begin{bmatrix} 1.24 \\ 1.14 \end{bmatrix}$ must be represented by the significant number $1 = 1.0(5) = \begin{bmatrix} 1.5 \\ 0.5 \end{bmatrix}$. It is true that the range of the significant number 1 does cover the range of 1.19(5), but the numbers are certainly not equivalent.

This illustration shows that there may be considerable loss in information in using significant numbers as a means of expressing results of computations, for we deliberately take a larger error than is necessary. The loss in the number of significant figures in products and quotients (stated in section 2.11), for example, is due not so much to the accumulation of errors as to the simplicity that has been gained at the expense of precision.

We need, then, to carry out our calculations with the use of range or approximation-error numbers if we wish precise results. Rules for manipulation with range and approximation-error numbers, together with an outline of some of the classical material on calculation with significant numbers, are presented in the later sections of this chapter.

2.5 Scientific and significant integer notation. Multiplication by a power of ten can be used to make the number of digits in a significant number the same as the number of significant figures. Application of this device results in so-called *scientific notation*. Any significant number can be written as a significant number between 1 and 10 multiplied by some power of 10. Thus $3720 = 3.720 \times 10^3$, $93{,}000{,}000 = 9.3 \times 10^7$, $0.0000153 = 1.53 \times 10^{-5}$.

This form of scientific notation is closely related to the laws of common logarithms since any significant number can be written (approximately) as a power of 10 if a table is available giving the (approximate) powers of 10 of the numbers between 1 and 10.

Another form of notation, in which each significant number consisting of n significant figures is multiplied by a power of 10 to make the result a significant n-place integer, might be called a *significant integer notation*. If I represents the significant integer and x the significant number, then

$$I = x \cdot 10^c, \tag{1}$$

where c may be positive or negative. Thus

$$3.9762 = 39762 \times 10^{-4} \quad \text{and} \quad 93{,}000{,}000 = 93 \times 10^6.$$

2.6 Absolute and relative error. The error of an approximate number is defined in (2.2.1). In many situations it is not so much the error

as the ratio of the error to the number that is important. The relative error of x, which is defined as

$$(1) \qquad \epsilon_r(x) = \frac{\epsilon}{x} = \frac{x - x'}{x} = \frac{\Delta x}{x} = 1 - \frac{x'}{x},$$

may be used to measure this ratio.

In many cases the true value of x is unknown and we have recorded only its approximate value x'. If ϵ is small with reference to x, an approximate value of the relative error, or an alternative definition of the relative error, is given by

$$(2) \qquad \epsilon_{r'}(x) = \frac{\epsilon}{x'} = \frac{x - x'}{x'} = \frac{\Delta x}{x'} = \frac{x}{x'} - 1.$$

The percentage error is by definition the relative error multiplied by 100.

In technical calculations the term *error* is reserved for the difference between an exact number and its approximation. Incorrect statements entering the calculation as a result of incorrect transcriptions or as a result of an improper use of the rules and laws of the computing system are due to *mistakes*. It is usually possible to eliminate mistakes from computational procedure, but, when dealing with approximate numbers, it is usually impossible to eliminate errors, although limits or bounds for these errors can often be computed.

2.7 The fundamental operations with range numbers. Range numbers may be used with the fundamental operations to secure range numbers representing sums, differences, products, and quotients.

In the operation of addition we have $x + y = \begin{bmatrix} x_H \\ x_L \end{bmatrix} + \begin{bmatrix} y_H \\ y_L \end{bmatrix}$. The sum may be as large as $x_H + y_H$ and as small as $x_L + y_L$. This follows at once no matter whether x and y are positive or negative. So

$$(1) \qquad x + y = \begin{bmatrix} (x + y)_H \\ (x + y)_L \end{bmatrix} = \begin{bmatrix} x_H + y_H \\ x_L + y_L \end{bmatrix}.$$

For example,

$$\begin{bmatrix} 2.38 \\ 2.34 \end{bmatrix} + \begin{bmatrix} 3.19 \\ 3.17 \end{bmatrix} = \begin{bmatrix} 5.57 \\ 5.51 \end{bmatrix}$$

and

$$\begin{bmatrix} 2.38 \\ 2.34 \end{bmatrix} + \begin{bmatrix} -3.17 \\ -3.19 \end{bmatrix} = \begin{bmatrix} -0.79 \\ -0.85 \end{bmatrix}.$$

This law can be applied to any number of additions simultaneously.

Thus

$$\begin{bmatrix} 1.73 \\ 1.69 \end{bmatrix} + \begin{bmatrix} 1.27 \\ 1.25 \end{bmatrix} + \begin{bmatrix} -0.63 \\ -0.67 \end{bmatrix} + \begin{bmatrix} -1.26 \\ -1.30 \end{bmatrix} = \begin{bmatrix} 1.11 \\ 0.97 \end{bmatrix}.$$

Before considering the operation of subtraction, we note that prefixing by a minus sign (multiplication by -1) changes the order of the terms in the range number in addition to changing their sign. Thus $-\begin{bmatrix} 3.19 \\ 3.17 \end{bmatrix} = \begin{bmatrix} -3.17 \\ -3.19 \end{bmatrix}$. With this adjustment subtraction is a special case of addition. We may then write

$$(2) \qquad x - y = \begin{bmatrix} (x - y)_H \\ (x - y)_L \end{bmatrix} = \begin{bmatrix} x_H \\ x_L \end{bmatrix} - \begin{bmatrix} y_H \\ y_L \end{bmatrix} = \begin{bmatrix} x_H \\ x_L \end{bmatrix} + \begin{bmatrix} -y_L \\ -y_H \end{bmatrix}$$

$$= \begin{bmatrix} x_H - y_L \\ x_L - y_H \end{bmatrix} = -\begin{bmatrix} y_H - x_L \\ y_L - x_H \end{bmatrix}.$$

Thus

$$\begin{bmatrix} 2.38 \\ 2.34 \end{bmatrix} - \begin{bmatrix} 1.19 \\ 1.17 \end{bmatrix} = \begin{bmatrix} 1.21 \\ 1.15 \end{bmatrix}$$

but

$$\begin{bmatrix} 1.19 \\ 1.17 \end{bmatrix} - \begin{bmatrix} 2.38 \\ 2.34 \end{bmatrix} = \begin{bmatrix} -1.15 \\ -1.21 \end{bmatrix} = -\begin{bmatrix} 1.21 \\ 1.15 \end{bmatrix}.$$

Addition and subtraction may be carried on simultaneously. For example,

$$\begin{bmatrix} 32.04 \\ 31.96 \end{bmatrix} - \begin{bmatrix} 2.27 \\ 2.21 \end{bmatrix} + \begin{bmatrix} 16.09 \\ 15.43 \end{bmatrix} + \begin{bmatrix} -3.08 \\ -3.16 \end{bmatrix}$$

$$= \begin{bmatrix} 32.04 - 2.21 + 16.09 - 3.08 \\ 31.96 - 2.27 + 15.43 - 3.16 \end{bmatrix} = \begin{bmatrix} 42.84 \\ 41.96 \end{bmatrix}.$$

It is usually preferable to factor the negative sign from a negative subtrahend since a subtraction of the form $a - (-b) = a + b$ is then easier to accomplish. Thus

$$\begin{bmatrix} 123 \\ 120 \end{bmatrix} - \begin{bmatrix} -110 \\ -115 \end{bmatrix} = \begin{bmatrix} 123 \\ 120 \end{bmatrix} + \begin{bmatrix} 115 \\ 110 \end{bmatrix} = \begin{bmatrix} 238 \\ 230 \end{bmatrix}.$$

Sometimes we wish to add or subtract exact numbers and approximate numbers. The above rules can be made to apply by writing the exact

(digital) numbers in range form. Thus the exact number 126 is represented in range form as $\begin{bmatrix}126\\126\end{bmatrix}$. If one approximate number is accurate to more decimal places than a second approximate number, a satisfactory result can be obtained by rounding off the more accurate number to one decimal place more than the less accurate number and treating the rounded-off number as though it were an exact number. Thus to add

$$\begin{bmatrix}245\\244\end{bmatrix} \text{ and } \begin{bmatrix}173.94397\\173.94388\end{bmatrix}$$

we form

$$\begin{bmatrix}245\\244\end{bmatrix} + \begin{bmatrix}173.9\\173.9\end{bmatrix} = \begin{bmatrix}418.9\\417.9\end{bmatrix}.$$

Similarly, if 1.37 is a significant number,

$$\pi + 1.37 = \begin{bmatrix}3.142\\3.142\end{bmatrix} + \begin{bmatrix}1.375\\1.365\end{bmatrix} = \begin{bmatrix}4.517\\4.507\end{bmatrix}.$$

Products of range numbers are handled similarly. When the numbers are positive (either exact or digital), the product is obtained by multiplying the large number by the large number and the small number by the small number. If one (or both) of the range numbers is negative, factor out the minus sign and use the above rule. Thus

$$(3) \qquad xy = \begin{bmatrix}x_H\\x_L\end{bmatrix}\begin{bmatrix}y_H\\y_L\end{bmatrix} = \begin{bmatrix}x_Hy_H\\x_Ly_L\end{bmatrix} \qquad x > 0, \qquad y > 0.$$

Also

$$\begin{bmatrix}8\\4\end{bmatrix}\begin{bmatrix}3\\2\end{bmatrix} = \begin{bmatrix}24\\8\end{bmatrix}$$

$$\begin{bmatrix}8\\4\end{bmatrix}\begin{bmatrix}-2\\-3\end{bmatrix} = -\begin{bmatrix}8\\4\end{bmatrix}\begin{bmatrix}3\\2\end{bmatrix} = -\begin{bmatrix}24\\8\end{bmatrix} = \begin{bmatrix}-8\\-24\end{bmatrix}$$

$$\begin{bmatrix}-4\\-8\end{bmatrix}\begin{bmatrix}-2\\-3\end{bmatrix} = -(-)\begin{bmatrix}8\\4\end{bmatrix}\begin{bmatrix}3\\2\end{bmatrix} = \begin{bmatrix}24\\8\end{bmatrix} \quad \text{and} \quad \begin{bmatrix}20\\20\end{bmatrix}\begin{bmatrix}3.6\\3.5\end{bmatrix} = \begin{bmatrix}72\\70\end{bmatrix}.$$

A range number may have components with different signs. This is not usual since here the error is larger than the approximation. Consider, for example, the number $\begin{bmatrix}a\\-b\end{bmatrix}$, where a and b are positive. Then

$$x' = \frac{a-b}{2} \qquad \text{and} \qquad \epsilon = \frac{a+b}{2}.$$

It follows that ϵ is larger than x', and relatively much larger if a is about the size of b. In most calculations the error term is much smaller than the approximate term, so numbers of this sort should appear infrequently.

Multiplication involving one of these numbers is easily accomplished if the sign of the number is so adjusted that the component having the largest absolute value appears as positive. Thus

$$\begin{bmatrix}1.18\\1.16\end{bmatrix}\begin{bmatrix}1.02\\-2.14\end{bmatrix} = -\begin{bmatrix}1.18\\1.16\end{bmatrix}\begin{bmatrix}2.14\\-1.02\end{bmatrix} = -\begin{bmatrix}2.5252\\-1.2036\end{bmatrix} = \begin{bmatrix}1.2036\\-2.5252\end{bmatrix}$$

and

$$\begin{bmatrix}-1.16\\-1.18\end{bmatrix}\begin{bmatrix}1.02\\-2.14\end{bmatrix} = \begin{bmatrix}1.18\\1.16\end{bmatrix}\begin{bmatrix}2.14\\-1.02\end{bmatrix} = \begin{bmatrix}2.5252\\-1.2036\end{bmatrix}.$$

Multiplication involving two of these numbers may be accomplished directly by writing the four possible products of the extreme values and selecting the highest and lowest, or by adjusting the sign of each number so that the component having the largest absolute value is positive. The product is then reduced to $\pm$ another product

$$\begin{bmatrix}a\\-b\end{bmatrix}\begin{bmatrix}c\\-d\end{bmatrix} \qquad \text{with} \qquad \begin{matrix}a \geq b \geq 0\\ c \geq d \geq 0\end{matrix}.$$

The upper component of this product is then ac, since $ac \geq bd$. The lower component is then either $(-bc)$ or $(-ad)$, whichever is smaller. Using the first method, we have

$$\begin{bmatrix}1.22\\-1.38\end{bmatrix}\begin{bmatrix}1.27\\-1.11\end{bmatrix} = \begin{bmatrix}1.5494\\-1.7526\end{bmatrix}$$

since the possible products are 1.5494, -1.3542, -1.7526, and 1.5318. Using the second method, we have

$$-\begin{bmatrix}1.38\\-1.22\end{bmatrix}\begin{bmatrix}1.27\\-1.11\end{bmatrix} = -\begin{bmatrix}1.7526\\-1.5494\end{bmatrix} = \begin{bmatrix}1.5494\\-1.7526\end{bmatrix}.$$

Significant numbers may be multiplied with the use of the corresponding range numbers. The product of the significant numbers 1.23 and 2.34 is then

$$\begin{bmatrix}1.235\\1.225\end{bmatrix}\begin{bmatrix}2.345\\2.335\end{bmatrix} = \begin{bmatrix}2.896075\\2.860375\end{bmatrix}.$$

This product may well be represented by an approximation to it such as $\begin{bmatrix}2.896\\2.860\end{bmatrix}$. A four-place range number that does cover the range of the product is $\begin{bmatrix}2.897\\2.860\end{bmatrix}$.

The calculation of products of more than two approximate numbers is carried out with repeated applications of the processes described above.

The quotients of two approximate numbers can also be computed with range numbers. The negative signs, if any are present, should first be removed from the numerator and denominator to obtain the form $\pm x/y$ with x and y positive. We then divide x_H by y_L to get the highest absolute value of the quotient and x_L by y_H to get the smallest absolute value.

$$\frac{x}{y} = \pm \begin{bmatrix}x_H\\x_L\end{bmatrix} \div \begin{bmatrix}y_H\\y_L\end{bmatrix} = \pm \begin{bmatrix}x_H/y_L\\x_L/y_H\end{bmatrix}. \tag{4}$$

So

$$\begin{bmatrix}625.7\\624.3\end{bmatrix} \div \begin{bmatrix}36.2\\35.8\end{bmatrix} = \begin{bmatrix}17.478\\17.245\end{bmatrix}^*,$$

whereas

$$\begin{bmatrix}625.7\\624.3\end{bmatrix} \div \begin{bmatrix}-35.8\\-36.2\end{bmatrix} = -\begin{bmatrix}625.7\\624.3\end{bmatrix} \div \begin{bmatrix}36.2\\35.8\end{bmatrix}$$

$$= -\begin{bmatrix}17.478\\17.245\end{bmatrix} = \begin{bmatrix}-17.245\\-17.478\end{bmatrix}$$

and

$$\begin{bmatrix}3.39\\3.15\end{bmatrix} \div \begin{bmatrix}3\\3\end{bmatrix} = \begin{bmatrix}1.13\\1.05\end{bmatrix}.$$

In rounding off the answers make sure that the range of the quotient is covered even though the error is bigger than one-half unit. (This amounts, in effect, to providing a true bound rather than an approximate limit.) Thus, though the ratio above (624.3 to 36.2) equals 17.245856, the result is recorded as 17.245 since that number represents the lower terminus of the range. If it were to represent the upper terminus, it should be recorded as 17.246.

The foregoing rule takes care of the usual situation where both components of each range number have the same sign. Additional consideration needs to be given to the case where the numerator, or the denominator, has components with different signs.

* A more precise statement would use the approximation sign, rather than the equals sign, when the results of the divisions are rounded off.

A good rule to follow, if the dividend has components with different signs, is one similar to the multiplication rule. The sign of the numerator is so adjusted that the component having the larger absolute value appears as positive. Thus

$$\begin{bmatrix}1.37\\-2.46\end{bmatrix} \div \begin{bmatrix}-1.11\\-1.16\end{bmatrix} = \begin{bmatrix}2.46\\-1.37\end{bmatrix} \div \begin{bmatrix}1.16\\1.11\end{bmatrix} = \begin{bmatrix}2.22\\-1.23\end{bmatrix}.$$

A similar rule for the case in which the divisor has components with different signs is not stated here since, in general, divisions of this type should not be performed. If the components of the prospective divisor have different signs, then the range of the divisor includes the number zero. Since division by zero is excluded, and since there is no way of knowing when the approximate number is zero, a safe procedure is to use the conservative rule, *never divide by a range number whose components have different signs.*

Range numbers are also adaptable to the operation of square root. If $x > 0$, then

$$\sqrt{x} = \sqrt{\begin{bmatrix}x_H\\x_L\end{bmatrix}} = \begin{bmatrix}\sqrt{x_H}\\\sqrt{x_L}\end{bmatrix}.$$

For example, the value of the square root of the significant number 103 is

$$\begin{bmatrix}\sqrt{103.5}\\\sqrt{102.5}\end{bmatrix} = \begin{bmatrix}10.18\\10.12\end{bmatrix}.$$

2.8 The fundamental operations with approximation-error numbers. Approximation-error numbers may be used to perform the fundamental operations and to secure approximation-error numbers representing the values of sums, differences, products, and quotients. If $x_1 = x'_1 + \epsilon_1$ and $x_2 = x'_2 + \epsilon_2$, then $x_1 \pm x_2 = x'_1 \pm x'_2 + (\epsilon_1 \pm \epsilon_2)$.

Now if ϵ_1 is an error, not greater in absolute value than η_1, and if $|\epsilon_2| \leq \eta_2$, it follows that the maximum absolute error of $x_1 \pm x_2$ is less than $\eta_1 + \eta_2$ since

$$\epsilon(x_1 \pm x_2) = (x_1 \pm x_2) - (x'_1 \pm x'_2) = \epsilon_1 \pm \epsilon_2, \tag{1}$$

and

$$|\epsilon(x_1 \pm x_2)| \leq \eta_1 + \eta_2.$$

If η_2 is small with respect to η_1, then the maximum possible error of the sum or difference is approximately that of x_1.

If x_2 is an exact number, then $\eta_2 = 0$, and the maximum possible error of the sum or difference is equal to that of x_1.

If η_1 and η_2 are each equal to or less than η, we may say

$$|\epsilon(x_1 \pm x_2)| \leq 2\eta. \tag{2}$$

This argument may be extended to include the algebraic sum of N numbers. Thus

$$|\epsilon(x_1 \pm x_2 \pm x_3 \pm \cdots \pm x_N)| \leq \eta_1 + \eta_2 + \cdots + \eta_N \tag{3}$$

and, if $\eta_i = \eta$,

$$|\epsilon(x_1 \pm x_2 \pm x_3 \pm \cdots \pm x_N)| \leq N\eta. \tag{4}$$

The rules for adding and subtracting approximation-error numbers are somewhat simpler than the rules for adding and subtracting range numbers, since one does not need to be so careful about signs. One computes the approximate sum or difference just as he does the exact sum or difference, but he adds a possible error term that is the sum of all possible errors.

Some of the addition and subtraction problems of the last section are here worked with the use of approximation-error numbers.

$$\begin{bmatrix}2.38\\2.34\end{bmatrix} + \begin{bmatrix}3.19\\3.17\end{bmatrix} = 2.36(2) + 3.18(1) = 5.54(3) = \begin{bmatrix}5.57\\5.51\end{bmatrix}$$

$$\begin{bmatrix}2.38\\2.34\end{bmatrix} + \begin{bmatrix}-3.17\\-3.19\end{bmatrix} = 2.36(2) - 3.18(1) = -0.82(3) = \begin{bmatrix}-0.79\\-0.85\end{bmatrix}$$

$$\begin{bmatrix}1.73\\1.69\end{bmatrix} + \begin{bmatrix}1.27\\1.25\end{bmatrix} + \begin{bmatrix}-0.63\\-0.67\end{bmatrix} + \begin{bmatrix}-1.26\\-1.30\end{bmatrix} = 1.71(2) + 1.26(1)$$

$$- 0.65(2) - 1.28(2) = 1.04(7) = \begin{bmatrix}1.11\\0.97\end{bmatrix}$$

$$\pi + 1.37 = 3.142(0) + 1.370(5) = 4.512(5) = \begin{bmatrix}4.517\\4.507\end{bmatrix}.$$

Products of approximation-error numbers can also be computed. We get

$$x_1x_2 = (x'_1 + \epsilon_1)(x'_2 + \epsilon_2) = x'_1x'_2 + \epsilon_2x'_1 + \epsilon_1x'_2 + \epsilon_1\epsilon_2$$

so that

$$\epsilon(x_1x_2) = x_1x_2 - x'_1x'_2 = \epsilon_2x'_1 + \epsilon_1x'_2 + \epsilon_1\epsilon_2. \tag{5}$$

The second-order error term $\epsilon_1\epsilon_2$ is usually very small and may be neglected in most problems. If we neglect it, we have the conventional formula

$$\epsilon(x_1x_2) = x'_1\epsilon_2 + x'_2\epsilon_1{}^*$$

Again if η_1 is the absolute value of the greatest possible value of ϵ_1, and η_2 of ϵ_2, we can write

$$|\epsilon(x_1x_2)| \leq |x'_1|\eta_2 + |x'_2|\eta_1. \tag{6}$$

This $ab + cd$ operation is easily performed with a computing machine. We do not even need to watch the signs. To use the earlier illustration, the product of the significant numbers 1.23 and 2.34 is

$$P = 1.230(5) \times 2.340(5).$$

The approximation term is $1.230 \times 2.340 = 2.8782$. The error term is $(1.230)(0.005) + (2.340)(0.005) = 0.01785$. The product, recorded to three decimal places, is then $2.878(18) = \begin{bmatrix}2.896\\2.860\end{bmatrix}$.

Quotients may be treated in a similar fashion since

$$\frac{x_1}{x_2} = \frac{x'_1 + \epsilon_1}{x'_2 + \epsilon_2} \tag{7}$$

$$= \frac{x'_1\left(1 + \dfrac{\epsilon_1}{x'_1}\right)}{x'_2\left(1 + \dfrac{\epsilon_2}{x'_2}\right)}$$

$$= \frac{x'_1}{x'_2}\left(1 + \frac{\epsilon_1}{x'_1}\right)\left(1 + \frac{\epsilon_2}{x'_2}\right)^{-1}$$

$$= \frac{x'_1}{x'_2}\left(1 + \frac{\epsilon}{x'_1}\right)\left(1 - \frac{\epsilon_2}{x'_2} + \frac{\epsilon_2{}^2}{x'_2{}^2} - \cdots\right)$$

$$= \frac{x'_1}{x'_2}\left(1 + \frac{\epsilon_1}{x'_1} - \frac{\epsilon_2}{x'_2} + \cdots\right).$$

* This formula can also be obtained by differential calculus since

$$d(x'_1x'_2) = x'_1\,dx'_2 + x'_2\,dx'_1.$$

An approximate value of the error of the quotient is then

$$\epsilon\left(\frac{x_1}{x_2}\right) = \frac{x'_1}{x'_2}\left(\frac{\epsilon_1}{x'_1} - \frac{\epsilon_2}{x'_2}\right) = \frac{x'_2\epsilon_1 - x'_1\epsilon_2}{{x'_2}^2}. \tag{8}$$

*

Again, if $|\epsilon_1| \leq \eta_1$ and $|\epsilon_2| \leq \eta_2$, we have

$$\left|\epsilon\left(\frac{x_1}{x_2}\right)\right| \leq \frac{|x'_2|\eta_1 + |x'_1|\eta_2}{{x'_2}^2}. \tag{9}$$

This formula also describes a single machine operation if the value of ${x'_2}^2$ is first computed. Thus

$$\frac{625.0(7)}{36.0(2)} = \frac{625.0}{36.0} \pm \frac{(36.0)(0.7) + (625.0)(0.2)}{1296}$$

$$= 17.361 \pm 0.116 = 17.361(116).$$

The formulas (6) and (9) are approximate and should not be used when the errors are relatively large. They should not be used, for example, when one of the numerators has components with different signs, for in this case the error may be larger than the approximation.

Some special quotient rules are worthy of note. If x_1 is an exact number, say A, (9) becomes

$$\left|\epsilon\left(\frac{A}{x_2}\right)\right| \leq \frac{|A|\eta_2}{{x'_2}^2}, \tag{10}$$

whereas, if x_2 is an exact number, it becomes

$$\left|\epsilon\left(\frac{x_1}{B}\right)\right| \leq \frac{\eta_1}{B}. \tag{11}$$

Thus

$$3 \div 3.27(12) = 0.9174 \pm \frac{3.00(0.12)}{(3.27)^2} = 0.9174(337)$$

and

$$3.27(12) \div 3 = 1.09(4).$$

The rule against division by zero becomes, when stated in approximation-error numbers: *never divide by an approximation-error number when the absolute value of the error term is as large or larger than the absolute value of the approximation term.*

* This formula may also be obtained with the use of the differential calculus since

$$d\left(\frac{x'_1}{x'_2}\right) = \frac{x'_2\,dx'_1 - x'_1\,dx'_2}{{x'_2}^2}.$$

The reader should note that the numerator of the right side of (9) is identical with the right side of (6), and hence that the recorded absolute value of the error of the quotient of the two numbers having relatively small errors is greater than, equal to, or less than the recorded absolute value of the error of the product of the numbers, depending on whether the absolute value of the denominator is less than, equal to, or greater than unity.

Square root may be accomplished with the use of approximation-error numbers. Thus, if $x' > 0$, $d(\sqrt{x'}) = \frac{1}{2\sqrt{x'}}\,dx'$, so that

$$|\epsilon(\sqrt{x})| \leq \frac{\eta}{2\sqrt{x'}}. \tag{12}$$

The square root of the significant number 103 is

$$10.149 \pm \frac{\frac{1}{2}}{2(10.149)} = 10.149(25) = \begin{bmatrix} 10.174 \\ 10.124 \end{bmatrix}.$$

2.9 Theorems on relative error. An alternative method of studying first-order error, particularly effective with products and quotients (and powers and roots), is by means of relative error. Formulas are obtained easily with the use of logarithmic differentiation. Thus, if

$$P = xy, \qquad Q = \frac{x}{y}, \qquad U = x^p, \qquad V = x^{1/p} \tag{1}$$

we have

$$\begin{aligned} \log_e P &= \log_e x + \log_e y \\ \log_e Q &= \log_e x - \log_e y \\ \log_e U &= p \log_e x \end{aligned} \tag{2}$$

$$\log_e V = \frac{1}{p}\log_e x.$$

It follows that

$$\begin{aligned} \frac{dP}{P} &= \frac{dx}{x} + \frac{dy}{y} \\ \frac{dQ}{Q} &= \frac{dx}{x} - \frac{dy}{y} \\ \frac{dU}{U} &= p\frac{dx}{x} \end{aligned} \tag{3}$$

$$\frac{dV}{V} = \frac{1}{p}\frac{dx}{x},$$

which gives us

$$
\begin{aligned}
&|\epsilon_r(P)| \leq |\epsilon_r(x)| + |\epsilon_r(y)| \\
&|\epsilon_r(Q)| \leq |\epsilon_r(x)| + |\epsilon_r(y)| \\
(4)\qquad &|\epsilon_r(U)| \leq p\,|\epsilon_r(x)| \\
&|\epsilon_r(V)| \leq \frac{1}{p}|\epsilon_r(x)|.
\end{aligned}
$$

These approximate inequalities may be summarized by the two theorems:

The absolute value of the relative error of a product (or quotient) is at most equal to the sum of the greatest absolute values of the relative errors of the numbers from which it is formed.

The absolute value of the relative error of a power (or root) is at most equal to the absolute value of the power (or the reciprocal of the root) times the greatest relative error of the number.

Once we have computed the relative error of a quantity we can compute the error by multiplying by the quantity or we may compute the approximate size of the error by multiplying by the approximate value of the quantity.

The errors of products and quotients (as well as powers and roots) may then be calculated by relative error. It is necessary only to compute the maximum relative error of the number in the product or quotient, to add these, and to multiply the result by the approximate product or quotient. For example, the maximum relative errors of the numbers 1.23 and 2.34 are, respectively, 0.0041 and 0.0021. The sum is 0.0062. Since the approximate product and quotient are 2.878 and 0.526, respectively, it follows that the errors are 0.018 and 0.003. Then $P = 2.878(18)$ and $Q = 0.526(3)$.

Similar treatments of $(1.23)^2$ and $\sqrt{1.23}$ give 1.513(12) and 1.109(2).

Before the introduction of computing machines, it was not practical to perform all the calculations necessary for obtaining limits or bounds for error in an extensive series of calculations. The practice was to prove certain statements that are true for large groups of approximate numbers and to use these facts in fixing an upper bound for the resultant error. These statements have usually been expressed in terms of significant numbers. A modified treatment of this general theory is presented in the following sections. Direct computation with range numbers or approximation-error numbers is recommended as a better procedure when precise statements of small maximum error are desired.

2.10 Relative errors and significant numbers. If x is known and ϵ is known, the maximum relative error may be calculated by (2.6.1). If x' is known (say positive) and η is an upper bound for the absolute value of ϵ, we may state

$$(1)\qquad \epsilon_r(x) \leq \frac{\eta}{x' - \eta}.$$

Consider the approximate number 1.295(135) of section 2.2. Application of (1) gives

$$\epsilon_r(x) \leq \frac{0.135}{1.295 - 0.135} = \frac{0.135}{1.160} = 0.1164 = 11.64\%.$$

The application of (1) to significant numbers yields

$$(2)\qquad \epsilon_r(x) \leq \frac{\frac{1}{2}\cdot 10^p}{x' - \frac{1}{2}10^p},$$

where 10^p indicates the unit in the last recorded position of the significant number. The theory relating relative errors and significant figures is conventionally developed [A] by considering the three cases $p < 0$, $p = 0$, $p > 0$. The treatment here reduces these three cases to a single case with the use of the following lemma. *If x is any significant number, there is a significant integer I having the same number of significant figures and having the same relative error.*

The first part of this lemma follows from (2.5.1). Also we know that

$$(3)\qquad \epsilon_r(I) = \frac{I - I'}{I} = \frac{x\cdot 10^c - x'\cdot 10^c}{x\cdot 10^c} = \frac{x - x'}{x} = \epsilon_r(x).$$

The use of this lemma enables us to calculate the relative errors of significant numbers without any consideration of the position of the decimal point, since all significant numbers with the same significant figures have the same relative error as the significant integer to which they may be transformed.

The application of (3) to (1) then gives

$$(4)\qquad \epsilon_r(x) = \epsilon_r(I) \leq \frac{\frac{1}{2}}{I' - \frac{1}{2}}.$$

For example, the relative error of the significant number 7.16 is indicated by

$$\epsilon_r(x) \leq \frac{\frac{1}{2}}{716 - \frac{1}{2}} = \frac{5}{7155} = 0.000699.$$

Now a bound for the relative error of a significant number may be determined quite accurately by a simple formula that depends only on the first digit of the approximate number and the number of digits in the number. Thus if k is the first non-zero digit of a number, and if the total number of significant digits is n, we may say

$$I' \geq k \cdot 10^{n-1} \qquad \text{and} \qquad I \geq k \cdot 10^{n-1} - \tfrac{1}{2}$$

so that

$$\epsilon_r(x) = \epsilon_r(I) \leq \frac{\frac{1}{2}}{k \cdot 10^{n-1} - \frac{1}{2}} = \frac{1}{2k \cdot 10^{n-1} - 1}.$$

In general, since $2k \cdot 10^{n-1} - 1 \geq k \cdot 10^{n-1}$, for $k > 0$ and $n \geq 1$, we have

$$\epsilon_r(x) = \epsilon_r(I) \leq \frac{1}{k \cdot 10^{n-1}}. \tag{5}$$

The restriction $k > 0$ implies that $n \geq 1$. If $k = 1$ and $n = 1$, we have the largest possible value of $\epsilon_r(x)$, subject to the restriction, with the relative error equal to unity. The error is as large as the number. For example, the significant number 1, when written as an approximation for the exact number 0.5, is in error by 0.5.

The statement (5) is useful in setting some sort of an upper bound for the value of the relative error without making detailed calculations with (4). A better inequality than (5) is easily obtained if an additional condition, usually satisfied, is made on the significant number. If the significant number has at least one non-zero digit besides the first digit, we may write

$$I' \geq k \cdot 10^{n-1} + l \cdot 10^{\alpha},$$

where l is a non-zero digit and α is an integer equal to or greater than zero. Application of (4) gives

$$\epsilon_r(x) = \epsilon_r(I) \leq \frac{\frac{1}{2}}{k \cdot 10^{n-1} + l \cdot 10^{\alpha} - \frac{1}{2}} = \frac{1}{2k \cdot 10^{n-1} + 2l \cdot 10^{\alpha} - 1} \tag{6}$$

$$\leq \frac{1}{2k \cdot 10^{n-1}}$$

since $2l \cdot 10^{\alpha} - 1 \geq 1$ for all permissible values of l and α.

A useful special case of (6) results when $k = 5$, for we then have

$$\epsilon_r(x) \leq \frac{1}{10^n}. \tag{7}$$

These formulas provide bounds for the relative error without the necessity of detailed calculation. Thus we may say at once that the relative errors of 1000, 1001, 5001 are not greater than 0.001, 0.0005, 0.0001, respectively. Decimal points may be inserted at any place in any of the three numbers without changing the relative errors.

It is clear that there is a close association between relative error and significant digits. The above formulas have been provided for estimating an upper bound from a knowledge of the significant numbers. The following pages are devoted to the problem of finding the number of significant digits in a significant number when the maximum relative error is known.

The number of significant digits of the significant number x is the same as the number of significant digits of the significant integer I. It is then only necessary to get a bound for the absolute error of I in order to indicate the number of proved significant places in x since

$$\epsilon(I) = I\epsilon_r(I). \tag{8}$$

We first prove the theorem: *If the relative error of a significant number* $\epsilon_r(x) \leq \frac{1}{(k+1)10^{n-1}}$, *where* k *is the first significant digit of* x, *then the error of* x *is not more than one unit in the* n*-th figure of* x. This follows since multiplication of I and $\epsilon_r(I)$ in

$$I < (k+1)10^{n-1}$$

$$\epsilon_r(I) \leq \frac{1}{(k+1)10^{n-1}}$$

results in

$$\epsilon(I) < 1.$$

In this case we are not permitted to say that x is significant to n places, since the error may be larger than one-half unit in the last position. We can say only that

$$\epsilon(I) < 1$$

$$\epsilon(x) < 1 \text{ unit in the } n\text{th digital position.} \tag{9}$$

We next prove the theorem: *If the relative error of a significant number* $x \leq \frac{1}{2(k+1)10^{n-1}}$, *where* k *is the first digit of* x, *then* x *has* n *significant*

digits. This follows at once, since now

$$I < (k+1)10^{n-1}$$

$$\epsilon_r(I) \leq \frac{1}{2(k+1)10^{n-1}}$$

so that

$$\epsilon(I) < \tfrac{1}{2}$$

and

(10) $\epsilon(x) < \frac{1}{2}$ unit in the nth position of x.

In this case we may say that x has n significant figures.

A third theorem is sometimes used in determining the number of significant figures when the relative error is known: *If the relative error of a significant number, $\epsilon_r(x) \leq 1/(2\cdot 10^n)$, then x is significant to n figures.* This follows since $I < (k+1)10^{n-1}$, with

$$\epsilon(I) \leq \frac{k+1}{20} \leq \tfrac{1}{2} \text{ (for } k = 1, 2, \cdots, 9)$$

so

(11) $\epsilon(x) \leq \frac{1}{2}$ in the nth position of x.

A fourth theorem is: *If the relative error of a significant number $\epsilon_r(x) \leq 1/10^n$, then x has at least $n-1$ significant places.* This is really a special case of Theorem 1, since $1/10^n = 1/(10\cdot 10^{n-1}) \leq 1/(k+1)\cdot 10^{n-1}$. It follows that x has an error no larger than unity in the nth position, so that the $n-1$ values are guaranteed.

2.11 The fundamental operations with significant numbers. We are now in a position to discuss calculation with significant numbers. In a general way a significant number is a special case of an approximation-error number, so it would seem that the general methods of computation with approximation-error numbers outlined in section 2.8 might be applicable to significant numbers. This would be true were it not for the fact that the limited ranges of significant numbers force considerable rounding off so that the results may be recorded as significant numbers. This rounding-off process deliberately discards essential information for the sake of ease of recording and computation, and it is not to be recommended if precise results are desired and if computing machines are available. However, it is the method usually presented in books dealing with approximate numbers.

The rules for addition and subtraction of significant numbers follow those of section 2.8. The only difference is that it is necessary to round off the result sufficiently to obtain some significant number whose range

includes the true range. Thus $1.68 + 7.43 = 9.11$, with a possible error of 0.01. Although the answer is indicated accurately with the error number 9.11(1) or the range number $\begin{bmatrix}9.12\\9.10\end{bmatrix}$, we are forced to use the significant number 9.1, which is identical with the range number $\begin{bmatrix}9.15\\9.05\end{bmatrix}$ if the answer is to be expressed as a significant number. Similarly, the value $\pi + 1.37 = 3.142 + 1.37 = 4.512(5) = \begin{bmatrix}4.517\\4.507\end{bmatrix}$ cannot be represented by the significant number $4.51 = \begin{bmatrix}4.515\\4.505\end{bmatrix}$. It is necessary to use the significant number $4.5 = \begin{bmatrix}4.55\\4.45\end{bmatrix}$, which has a much larger range.

The case with which the sums and differences of approximation-error numbers can be computed, when compared with the arbitrariness of significant numbers, indicates the use of approximation-error numbers rather than significant numbers in pure addition and subtraction.

The situation is somewhat the same in the case of multiplication and division. There is unnecessary restriction in expressing the results in the form of significant numbers. However, the number of significant figures may be determined, without the extensive computation demanded by approximation-error numbers, from a rule that is developed from the theorems of the last section. This rule is: *The product (or quotient) of two numbers, each containing n significant figures (at least two of which are not zero), is a significant number of at least $n-2$ figures. If the leading digits of these numbers are both equal to or greater than 2, then the product (or quotient) has at least $n-1$ significant figures.* Let I_1 and I_2 be the significant integers corresponding to x_1 and x_2. Then

$$I_1 \leq k_1\cdot 10^{n-1} + l\cdot 10^{\alpha_1}, \qquad I_2 \leq k_2\cdot 10^{n-1} + l\cdot 10^{\alpha_2}$$

so that by (2.10.6)

$$\epsilon_r(I_1) \leq \frac{1}{2k_1\cdot 10^{n-1}} \quad \text{and} \quad \epsilon_r(I_2) \leq \frac{1}{2k_2\cdot 10^{n-1}}.$$

It follows from application of (2.10.3) and (2.9.4) that

(1) $$\epsilon_r(x_1x_2) = \epsilon_r(I_1I_2) = \frac{1}{2k_1\cdot 10^{n-1}} + \frac{1}{2k_2\cdot 10^{n-1}}$$

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

for all values of k_1 and k_2.

Now by Theorem 4 of the last section, the value of x_1x_2 is guaranteed to only $n-2$ places. If, however, $k_1 \geq 2$ and $k_2 \geq 2$, (1) becomes

$$\epsilon_r(x_1x_2) \leq \frac{1}{2\cdot 10^{n-1}} \tag{2}$$

and x_1x_2 is guaranteed to $n-1$ figures by Theorem 3.

An almost identical argument holds for the quotient.

Application of this rule does not lead to precise results. Thus the product of 1.23 and 2.34 is given by the significant number $3 = \begin{bmatrix} 3.5 \\ 2.5 \end{bmatrix}$. The use of approximation-error numbers in 2.8 shows that a much better answer is $2.878(18) = \begin{bmatrix} 2.896 \\ 2.860 \end{bmatrix}$.

If the factors of the product have different numbers of significant places, the number of significant places in the product is controlled by the factor having the smallest number of significant places. This is shown by applying (2.10.6) to (2.9.4). A similar rule holds for quotients.

2.12 Roots and powers with significant numbers. The conventional rules for the number of significant places of powers and roots follow a similar pattern. If the number has at least two non-zero digits, then by (2.9.4) and (2.10.6) we have

$$\epsilon_r(x_i^p) \leq \frac{p}{2k\cdot 10^{n-1}} \tag{1}$$

If $p = k$, $\epsilon_r(x_i^p) \leq 1/(2\cdot 10^{n-1})$ and x_i^p has at least $n-1$ significant digits by Theorem 3 of section 2.10, whereas, if $p \leq 10k$, $\epsilon_r(x_i^p) \leq 1/(2\cdot 10^{n-2})$ and x_i^p has $n-2$ significant places.

Similarly

$$\epsilon_r(x_i^{1/p}) \leq \frac{1}{2pk\cdot 10^{n-1}}. \tag{2}$$

If $pk \geq 10$, the right-hand value is equal to or less than $1/(2\cdot 10^n)$ and the root has n significant figures. If $pk < 10$, the right-hand side is equal to or less than $1/(2\cdot 10^{n-1})$ and the root has $n-1$ significant figures.

The formula for square root is a special case with $p = 2$. Then (2.9.4) gives

$$\epsilon_r(\sqrt{x}) \leq \frac{1}{4k\cdot 10^{n-1}}. \tag{3}$$

It follows that the square root of a n-place number is significant to $n-1$ places if the first digit of the number is 5 or less, and to n places otherwise.

The limitations of this conventional method of handling computations with approximate numbers are apparent when we apply it to the problem of finding the square root of the approximate number 103. Application of the rule leads to a two-place number, the significant number $10 = \begin{bmatrix} 10.5 \\ 9.5 \end{bmatrix}$, whereas calculation with approximation-error numbers shows that the precise result is 10.149(25), with an error of less than 3 in the fourth digit.

The reader is referred to Scarborough and to Walker and Sanford [A] for further discussion of significant numbers.

2.13 Recommendations for computation with approximate numbers. The selection of a suitable type of approximate number depends upon the purpose of the computation. Operations with significant numbers, particularly when supplemented with the use of the theory of the last two sections, are easier and simpler than the corresponding operations with range or approximation-error numbers. They are quite satisfactory when additions, subtractions, or a single multiplication or division are involved. They are also satisfactory when we are not concerned with the loss of significant figures in each operation. In most computational work we cannot afford this luxury.

Range numbers or approximation-error numbers are preferred to significant numbers for precise calculation with approximate numbers. The methods used in obtaining range numbers are more accurate than those used in getting approximation-error numbers, though the difference is trivial in the usual case in which the relative errors of the numbers are very small.

For most operations, approximation-error numbers are preferable to range numbers for ease of calculation. Computations with range numbers demand dual calculations at each step and constant attention to signs. Approximation-error numbers demand a single computation for the approximation, with an auxiliary computation for the error, which is usually accomplished easily with the machine. The use of approximation-error numbers, in general, requires the recording of fewer digits than the use of range numbers.

Both range numbers and approximation-error numbers have this undesirable property: different orders of computation in complex calculations may lead to different results. For example, the evaluation of a determinant of approximate numbers by conventional methods and the use of either range or approximation-error numbers may lead to dif-

ferent bounds, depending on the choice of the terms in the elimination process. Some general rules can be provided for situations of this sort, as von Neumann and Goldstine have provided rules of algebra for pseudo-operations [B]. For the basic linear problems the method of the next paragraph is to be preferred.

An alternative method is the use of incomplete numbers. An *incomplete number* is an approximation-error number in which the error term is omitted. These numbers look very much like significant numbers, but, unlike significant numbers, the results may be recorded to any desired number of places. This method makes for ease with a machine, since all numbers to be placed on the machine may be rounded off to the same number of places. It must be remembered that any recorded number is not necessarily a significant number in the technical sense, that is, we do not know what the bound for the error may be.

Calculation with incomplete numbers, then, amounts to calculation with a desirable form of the approximation term of an approximation-error number. The omission of the calculation of the error term would be very unsatisfactory were it not for the fact that, frequently, independent calculations of the error are available. These may be computed separately and then be attached to the result obtained with the use of incomplete numbers. Incomplete numbers are used in handling the linear computations of this book, since separate estimates may be made of the maximum errors of determinants, solutions of simultaneous equations, and the elements of the inverse matrix.

REFERENCES

A. 1. J. B. Scarborough, *Numerical Mathematical Analysis*, Second Edition, Johns Hopkins Press, Baltimore, 1950.

2. Helen Walker and Vera Sanford, "The accuracy of computation with approximate numbers," *The Annals of Mathematical Statistics*, **5**, 1–12 (1934).

B. J. von Neumann and H. H. Goldstine, "Numerical inverting of matrices of high order," *Bulletin of the American Mathematical Society*, **53**, 1021–1099 (1947).

This article contains much interesting material, including a discussion of the sources of errors in a computation and the rounding off of errors and their cumulation. Application is made to the calculation of the inverse matrix with the method of elimination (the method of single division of Chapter 6).

EXERCISES

1. Consider the number $\begin{bmatrix}2.43\\2.16\end{bmatrix}$. Write it in approximation-error form, and calculate its relative error.

2. Write the number 1.8923(46) in range form. Calculate its relative error.

3. Express in scientific notation and in significant integer notation.

(*a*) 0.00639
(*b*) $63.9\cdot10^{-8}$
(*c*) 92,500,000
(*d*) $62.5\cdot10^{20}$

4. Perform the indicated operations.

(*a*) $\begin{bmatrix}2.97\\2.83\end{bmatrix}+\begin{bmatrix}0.88\\0.86\end{bmatrix}-\begin{bmatrix}1.92\\1.87\end{bmatrix}+\begin{bmatrix}-1.13\\-1.23\end{bmatrix}$

(*b*) $\begin{bmatrix}2.98\\2.83\end{bmatrix}\cdot\begin{bmatrix}0.88\\0.86\end{bmatrix}$

(*c*) $\begin{bmatrix}2.97\\2.83\end{bmatrix}\div\begin{bmatrix}0.88\\0.66\end{bmatrix}$

(*d*) $\pi+\begin{bmatrix}2.24\\2.13\end{bmatrix}$

(*e*) $\pi-\sqrt{2}$ (to three decimal places)

(*f*) $\begin{bmatrix}-1.24\\0.96\end{bmatrix}\div 4$

5. Evaluate and express the results in approximation-error form.

(*a*) 8.321(15) + 6.297(2) − 7.777(77)
(*b*) 2.345(2) − 3.456(3)
(*c*) $\dfrac{2.345(2)}{3.456(3)}$
(*d*) $\sqrt{2.345(2)}$
(*e*) 2.345(2) ÷ 5
(*f*) $\dfrac{5}{2.345(2)}$

6. Work exercise 5(*c*) and exercise 5(*d*), using relative error formulas.

7. Using conventional rules, write the values of ab, a/b, and $\sqrt{a}$ in significant numbers if a is the significant number 2.345 and b is the significant number 3.456.

8. Find the perimeter and the area of a rectangle with sides $a = 163.0(2)$ feet and $b = 276.3(4)$ feet. Find the maximum relative error of the perimeter and of the area.