ELEMENTS OF
NUMERICAL ANALYSIS

Academic Press
Textbooks in Mathematics

Consulting Editor: **Ralph P. Boas, Jr.,**
Northwestern University

HOWARD G. TUCKER. An Introduction to Probability and Mathematical Statistics

EDUARD L. STIEFEL. An Introduction to Numerical Mathematics

WILLIAM PERVIN. Foundations of General Topology

JAMES SINGER. Elements of Numerical Analysis

PESI MASANI, R. C. PATEL and D. J. PATIL. Elementary Calculus

ELEMENTS OF
NUMERICAL ANALYSIS

James Singer

Department of Mathematics
Brooklyn College
Brooklyn, New York

NEW YORK **ACADEMIC PRESS** LONDON

ACADEMIC PRESS INC.
11 Fifth Avenue, New York, New York 10003

United Kingdom Edition published by
ACADEMIC PRESS INC. (LONDON) LTD.
Berkeley Square House, London W.1

LIBRARY OF CONGRESS CATALOG CARD NUMBER: 64-18216

PRINTED IN THE UNITED STATES OF AMERICA

To H and R and J

Preface

This book is written with two sets of readers in mind, the practicing scientific worker and the "pure" mathematician. The practicing scientific worker—the chemist, the physicist, the engineer, the economist, anyone who is concerned with the quantitative aspects of the physical, biological, social and applied sciences—knows only too well that much of his effort is directly or indirectly devoted to the determination of numerical results and to the derivation of natural laws, which are nothing but relations between numbers endowed with "dimensions." This book aims to tell him how to obtain a numerical result and how to judge the reliability or trustworthiness of his answer. The scientific worker will find many of the necessary formulas and many special tables to help him in his computations, he will find detailed descriptions of the methods and procedures, he will be aided by many illustrative examples worked out in the text, he will be guided by many remarks, observations, and words of caution.

The "pure" mathematician is usually interested, if at all concerned, with the art rather than the practice of computation. This book attempts to give him a coherent, systematic and, I trust, lucid treatment of the classical or traditional theory of mathematical computation. He will find careful and honest proofs where proofs are given; and he will learn that there is frequently an amazing amount of real mathematics behind a prosaic numerical answer, correct to five decimal places.

It is my earnest hope, however, that as far as possible the two sets of readers merge into one. It has always been my contention that the scientific worker interested in a numerical answer would do well to delve into the foundations of his methods, to learn "why" as well as "how"; an understanding of the underlying concepts is a powerful tool when he must cope with new problems or with old problems in new dress. On the other hand, it is my hope that those not now intrigued with computation will nevertheless plunge in to help discover new and better methods and more sound results if for no other reason than the fun of it.

For these reasons, the text not only includes set algorithms and tables, but attempts to give the reader some feeling for and insight into the subject so that he will be more than ready to strike out on his own.

This book is intended as a first course in numerical computation. It is not geared to electronic computers although it will serve as an introduction for those interested in high speed calculators. The methods and procedures that

are described can readily be modified, if modifications are needed, for use on electronic computors; but fundamentally, the procedures were intended to be carried out on desk calculators or even longhand.

For an understanding of most of the text, the reader will need a good introductory course in calculus; for some portions, some advanced calculus and differential equations will be necessary; for some of the material, not even the calculus is necessary. The references listed at the end of the book are few in number; they have been listed either because they can be used for supplementary reading or because they themselves contain extensive bibliographies. Various tables, not readily found elsewhere, are included in the text, but the serious reader should supply himself with a set of ordinary tables including the usual trigonometric, logarithmic and exponential tables.

The reader will find two chapters not usually covered in present day texts, one on geometric methods and nomography and one on curve fitting; he will also find many illustrative examples throughout the text. It is suggested that these be more than read; the reader should also work them out and compare his results with those in the text. In some cases, the examples worked out are merely illustrations of theory or algorithms previously discussed in the text; in some cases, the examples worked out serve as the vehicle for the explanations of new theory or modes of operation.

The text can be covered thoroughly in two semesters. Those who desire a faster pace can cover a good portion of it in one semester and finish it in a second semester with further topics such as matrix solutions or partial differential equations that are omitted from this book.

A final word addressed to the teacher. The examples, by and large, were intended to be worked out with the aid of desk calculators but if these are not available, the number of required significant figures or decimal places should be cut to prevent prohibitively long calculations.

JAMES SINGER

Brooklyn, New York

Contents

ELEMENTS OF
NUMERICAL ANALYSIS

Numbers and Errors

1.1. Significant Figures. In this chapter we develop some of the basic properties of numbers that are peculiar to the science (or art) of computation. The reader will please bear with us if we begin with some very elementary considerations.

Numbers used by the scientific worker are usually written in the decimal notation. Let us recall that in this notation the successive places to the left of the decimal point are the *unit's*, *ten's*, *hundred's*, *thousand's*, *ten-thousand's*, etc., places and the successive places to the right of the decimal point are the *tenth's*, *hundredth's*, *thousandth's*, *ten-thousandth's*, etc., places. We use the convention of enumerating the digits of a number written in decimal form from left to right to simplify some of the later definitions; the first digit is then the one on the extreme left and the last digit is the one on the extreme right.

The decimal representations of 22/5, 22/7, and π are different in character. The first decimal expression *terminates* or is *finite*, the second is *nonterminating* but *periodic*, the third is *nonterminating* and *nonperiodic*. Since the scientific worker rarely if ever uses any but the first kind of decimal expression, we too, unless otherwise indicated, shall use only finite or terminating decimals. This implies that frequently a written number is only an approximation to some other number. (We remark that any number, be it 22/5, 22/7, $\sqrt{2}$, or π, is exact; it becomes "inexact" or "approximate" only when it is considered as an evaluation or representation of some other number.) We now pave the way to a better understanding of these approximations.

DEFINITION 1. The *numerical unit* of a number written in the decimal notation is the name of the place occupied by the last digit, except in the case of a whole number which terminates in one or more zeros (all to the left of the decimal point). The numerical unit in the exceptional case, if not implied by the context, must be specifically stated and may be either the name of the place occupied by the last nonzero digit or the name of the place occupied by any one of the zeros to the right of the last nonzero digit.

For example, the numerical units of the numbers 3.04, 0.0700, 67, are hundredth, ten-thousandth, and unit, respectively. The numerical unit of 67,000 may be a thousand, hundred, ten, or unit and if not implied by the text must be explicitly stated.

The last illustration indicates that two numbers may be numerically equal but can have different numerical units. We wish to emphasize this point. Consider the numbers 3.04 and 3.040. They are numerically equal but differ in form; the numerical unit of the first is a hundredth; that of the second is a thousandth.

It is convenient to extend the concept of a numerical unit. We shall regard it not only as the name of a place in the decimal representation of a number but also as a number which is an appropriate power of 10. Thus, the numerical units a thousandth and a hundred will be represented by the powers 10^{-3} and 10^{2}, respectively. If the numbers above are written in the forms

$$3.04 = 304 \times 10^{-2},$$

$$0.0700 = 700 \times 10^{-4},$$

$$67 = 67 \times 10^{0},$$

$$67{,}000 = 67 \times 10^{3} = 670 \times 10^{2} = 6700 \times 10 = 67{,}000 \times 10^{0},$$

the power of 10 in each case indicates the numerical unit. In general, any number n can be written in the *numerical unit form*

$$n = n' \times 10^{u}, \tag{1.1:1}$$

where n' is a whole number and 10^{u} is the numerical unit of n. (We use the notation 1.2:3 to signify that the corresponding formula, equation, or statement is in Chapter 1, Section 2, and is numbered third in that section.) It follows, of course, that u, too, is an integer, positive, negative, or zero. If all the digits of a number are zero, as in 0.00, we put n' equal to zero.

DEFINITION 2. The *significant digits* or *figures* of a number n are the digits in n' when n is written in the numerical unit form.

Thus, 3.04, 0.0700, 67, and 0.00 have 3, 3, 2, and 1 significant digits, respectively. The number 67,000 may have 2, 3, 4, or 5 significant digits depending on the numerical unit. Omitting this exceptional case of an integer that terminates in one or more zeros, the number of significant figures of a number written in the decimal notation is the number of its digits excluding all digits that precede the first nonzero digit.

The significant figures of a number are so named because they are the ones that specify the number of numerical units.

We call the attention of the reader to another notation often used similar to the numerical unit form. It is frequently used in the printing of tables and in the tabulation of data and is called the *scientific* or *standard notation* or *form*. A number is written in the standard notation as

$$n = n'' \times 10^{v}, \tag{1.1:2}$$

where n'' has the same digits as n' in the numerical unit form but has just one nonzero digit left of the decimal point. Thus, 3.04, 0.0700, and 67 are

$$3.04 \times 10^{0},$$
$$7.00 \times 10^{-2},$$
$$6.7 \times 10,$$

respectively, in standard notation. The number zero shall be written as 0.00×10^{u} in the standard notation. The standard notation is particularly useful for numbers like 0.0000720 or 95,000,000 (where the numerical unit is a million, say) which are written as 7.20×10^{-5} and 9.5×10^{7}, respectively.

Generally speaking, a number used by a scientific worker arises in one of three ways. It may, first of all, be a "pure" number, that is, one which is the result of a count, or one which is the result of a mathematical or other definition. As examples of pure numbers we have the number (three) of sides of a triangle, the ratio of the circumference to the diameter of a circle, the value of $\sin 23°$, or $\int_1^2 e^{-t^2}\,dt$, the number of feet in a mile, the number of days in a week, the number of pounds in the maximum load of an elevator. Secondly, there are numbers that arise as values of direct measurements. (By a direct measurement we mean one in which the result is read off some measuring instrument such as the measurement of a distance by a ruler or the measurement of a temperature with a thermometer.) Thirdly, there are numbers that arise as results of computations performed on numbers of the first two types.

But, as we know, relatively very few numbers can be written exactly as finite decimals, measurements are at best approximate, and calculations are subject at the very least to all the inaccuracies of the numbers involved. Hence a number used by a scientific worker is usually an approximation to some "true" value. It is therefore important that he should indicate in some fashion the goodness of the approximation, the reliability, or the margin of error of a stated number. This can be done

in a variety of ways. He may write 6.040 ± 0.003 to indicate that the correct value is in the range from 6.037 to 6.043, inclusive. Note that if one wants to indicate a margin of error of 0.0003, say, one should not write 6.04 ± 0.0003 but 6.0400 ± 0.0003. The scientific worker will also use 6.04^- to indicate that the true value of a number is less than 6.04 but closer to it than to 6.03. Likewise, 6.04^+ indicates a true value greater than 6.04 but closer to it than to 6.05. These methods of writing approximate numbers clearly indicate that the numbers are approximate and give the margins of their errors but as matters of notation they are just a bit clumsy. The scientific worker will most frequently write 6.04 with the intent and understanding that this does not represent the number 6.04 exactly but a number which is closer to 6.04 than it is to 6.03 or 6.05. Likewise, 6.040 indicates a number which is closer to 6.040 than it is to 6.039 or 6.041.

The last notation determines a number with a margin of error equal to one-half the numerical unit; the preceding notation also determines a number with the same margin of error but also indicates whether the error is one of excess or default. The first notation like the last does not indicate the direction of the error but usually indicates a more precise margin of error.

Let us note in passing that the margin of error is closely linked with the numerical unit of the stated number and is, in the last notation, just one-half of that unit. Thus the margin of error in 6.040 is one-tenth the margin of error of 6.04. Since the number of significant figures in a number and the numerical unit of the number are themselves closely related, one must beware of using more significant figures than are warranted in writing a number. Just how many one should use will appear shortly.

The following definition will be useful.

DEFINITION 3. If a number a with k significant figures is an approximation to a number n and is the best approximation to n of all numbers with k significant figures, then a is said to be *correct to k significant* figures as an approximation to n.

Thus, 3.1, 3.14, 3.142, and 3.1418 are correct to 2, 3, 4, and 5 significant figures, respectively, when considered as approximations to $28/9$, $\sqrt[3]{31}$, π, and $\log_{10} 1386$, respectively.

It is desirable for some purposes to "round off" a number which is written in the usual decimal notation with $k + m$ significant figures to one that has only k significant figures. We do this by deleting those of the last m digits that are to the right of the decimal point and substituting zeros for those that are to the left of the decimal point. No further change

is necessary if the m deleted or replaced digits represent less than one-half unit in the kth place; but if the deleted or replaced digits represent more than one-half unit in the kth place, the kth significant figure is increased by unity. (If the kth significant figure is 9, it changes to 0 and the preceding digit is increased by unity. Note the last illustration in the table below.) If the deleted or replaced digits represent exactly one-half unit in the kth place, usage varies. Some people treat this case like the preceding one and increase the kth digit by unity; others increase the kth digit by unity if it is odd and leave it alone if it is even. The reasoning behind this latter rule is specious; in actual practice, it matters little which system is used.

ILLUSTRATIONS

Number	Rounded off to: 5 significant figures	4 significant figures	3 significant figures	2 significant figures
32.0769	32.077	32.08	32.1	32
0.856025	0.85603	0.8560	0.856	0.86
123456	123460	123500	123000	120000
1234.56	1234.6	1235	1230	1200
1.34996	1.3500	1.350	1.35	1.3
0.999777	0.99978	0.9998	1.00	1.0

In particular, note that 1.34996 becomes 1.3 when rounded off to two significant figures and 1.35 when rounded off to three significant figures. If, however, we were given 1.35 and told to round it off to two significant figures, the correct answer is 1.4. Many authors write $1.3\bar{5}$ to indicate 1.35^{-}; rounded off to two significant figures, this number is 1.3. In brief, to round off a number with $k + m$ significant figures to one with k significant figures is to rewrite it correct to k significant figures as an approximation to its original form.

The numbers 3.14209 and 3.14285 are approximations to $\pi = 3.14159\cdots$. Neither one is correct to six significant figures. If they are rounded off to five significant digits to 3.1421 and 3.1428 (or 3.1429), respectively, they remain incorrect to five significant digits. But when they are rounded off to four significant digits to 3.142 and 3.143, respectively, the first becomes correct to four significant digits as an approximation to π. The latter becomes correct when rounded off to three significant digits. We are thus led to the following extension of Definition 3.

DEFINITION 4. If a number a with $k + m$ significant digits when rounded off to $k + 1$ significant digits is not correct to $k + 1$ significant digits as an approximation to a number n but when rounded off to k significant digits is correct to k significant digits, then a is said to be correct to k significant digits as an approximation to n.

Thus, 1.33530 is correct to four significant digits when considered as an approximation to $\sec 41°30' = 1.3352$ and is correct to two significant figures when considered as an approximation to $\frac{4}{3}$. Similarly, $\frac{1}{3}$ expressed as a decimal would be correct to two significant figures as an approximation to $\sin 19°30' = 0.33381$ and to three significant figures as an approximation to $\sqrt{0.\overline{111}} = 0.33317$.

EXERCISE 1.1

1. State the numerical unit of each of the following numbers and write each numerical unit in the form 10^n.

a. 436 **b.** 750.2 **c.** 2.006 **d.** 0.05 **e.** 0.000050
f. 400.0 **g.** 0.00000 **h.** 1.976530 **i.** 1.000001 **j.** 883.09000.

2. Do the same for each of the following numbers; give all the possibilities if there are several.

a. 956000 **b.** 906000 **c.** 1000000 **d.** 1000001 **e.** 999999 **f.** 3020010.

3. How many significant digits are there in each of the following numbers?

a. 4029 **b.** 40.29 **c.** 53.670 **d.** 0.0002 **e.** 190
f. 2.000000 **g.** 2.000006 **h.** 3.0002 **i.** 83.10400 **j.** 0.08040.

4. Write each number in examples 1, 2, and 3 in standard notation.

5. Round off each of the following numbers to four significant digits.

a. 4.32974 **b.** 682.548 **c.** 28.9956 **d.** 102843.1 **e.** 0.0765402
f. 8976.49 **g.** 0.999996 **h.** 1.35000 **i.** 407.391 **j.** 32.1089.

6. Write each of the following numbers correct to four significant digits.

a. 22/7 **b** π **c.** 100000/3 **d.** $\cos 0°$ **e.** $\cos 25'$ **f.** $\sqrt{0.00809}$
g. $\sqrt[3]{0.00000685}$ **h.** 10! **i.** π^3 **j.** the number of inches in a mile.

7. Write each of the numbers of example 6 correct to the nearest tenth.

8. The first number in each of the following pairs is an approximation to the second number. Write each approximation as a decimal if not already so written and state the number of correct significant figures in the approximations.

a. 563.201, 563.257 **b.** 0.00632, 0.00636 **c.** 52, 000, 000, 52, 475, 913
d. 4.732093, 4.732102 **e.** 3800, 3826.4 **f.** $\sqrt{3}/10$, $\sin 10°$
g. 3/4, $\log 5.624$ **h.** 1, $\cos 30'$ **i.** 19/6, $\sqrt{10}$ **j.** $\sqrt[3]{3.87}$, $\pi/2$.

1.2. Errors. It was pointed out in the last section that for a variety of reasons a number used by a scientific worker is usually an approximation to some true value. We propose to examine these errors a little further in this section.

The difference e between a number n and an approximation a to it is defined as the *actual error* in a; in symbols,

$$e = n - a, \tag{1.2:1}$$

whence

$$n = a + e. \tag{1.2:2}$$

The *relative actual error* is defined by the statement

$$r = \left| \frac{e}{n} \right|, \tag{1.2:3}$$

and the *per cent relative actual error* is defined as

$$100r\%. \tag{1.2:4}$$

It is to be noted that for a and n real, e may be positive, negative, or zero, whereas the relative errors are zero or positive only.

Thus, the actual error committed in approximating π by 22/7 is

$$\begin{aligned} e &= \pi - 22/7 \\ &= 3.14159265^{+} - 3.14285714^{+} \\ &= -0.0012645^{-}; \end{aligned}$$

the relative actual error is

$$r = \frac{0.0012645^{-}}{3.14159265^{+}} = 0.00040^{+};$$

and the per cent relative actual error is

$$0.040^{+}\%.$$

The actual error in approximating π by 3.14 is

$$\begin{aligned} e &= \pi - 3.14 \\ &= 3.14159^{+} - 3.14 \\ &= 0.00159^{+}, \end{aligned}$$

and the relative actual error is

$$r = \frac{0.00159^{+}}{3.14159^{+}} = 0.00050^{+}.$$

Note that in these two illustrations the actual and relative actual errors can be calculated to as many significant figures as we wish provided that π is given with a sufficiently great number of correct significant figures.

Let us now imagine that the members of a class read, one by one, a barometer furnished with a vernier scale. Their readings will not be all alike and range, say, from 761.5 to 762.5 mm; let us suppose that it is decided to record the atmospheric pressure as 762 mm. This value, 762 mm, is, of course, an approximation to the true value of the atmospheric pressure and is the a of formula 1.2:1. However, the true value n is not known and therefore the value of e is not known. The best we can say is that n is between 761.5 and 762.5 and that the actual value of e is at most 0.5.

In general, if the true value of a number t is not known but it is known that it differs from an approximation a by an amount which is less than a positive number h, we have

$$a - h \leqslant t \leqslant a + h. \tag{1.2:5}$$

We call h the *margin of error* or the *maximum error* of a; the ratio

$$m = \left| \frac{h}{a} \right| \tag{1.2:6}$$

is called the *maximum relative error* of a; and

$$100 \left| \frac{h}{a} \right| \% \tag{1.2:7}$$

is called the *per cent maximum relative error*. Note that the maximum relative error has the approximate number in the denominator whereas the relative actual error has the exact value in the denominator. The approximate number must be used here because the exact value is not known. Some authors use the approximate value in all cases, but it seems more natural to use the exact value when it is known.

To illustrate these definitions, suppose that the height of a mountain is given as 6703 ft but is in error by 6 in. or less, that is, the margin of error or the maximum error is 6 in. The true height of the mountain is between 6702.5 and 6703.5 ft; the maximum relative error is approximately 0.0000746 or 0.00746%. Again, suppose the width of a paper is measured as 10.0 in. with the true value somewheres between 9.95 and 10.05 in. The maximum error is 0.05 inches and the maximum relative error is 0.005 or 0.5%. Thus, the maximum error in the first